

Brace Yourself

Prospects and Problems for Improving the Mental Environment

Master's Thesis submitted for the
Research Master's Philosophy
Utrecht University
June 14th 2019

Charlie T. Blunden
Student ID: 6083714
Word count: 39987 (excluding Abstract, Contents page, and
Acknowledgements)

Supervisor: Dr. Hanno Sauer
Second Reader: Dr. Joel Anderson
Third Reader: Prof. Dr. Daniel Cohnitz



Universiteit Utrecht

For Margi, Martin, Ruth, & Nick,
without whom this would not have been possible.

Abstract

Psychology and cognitive science have given us important insights into human irrationality: notably into the sources of our cognitive biases and the features of our environments that are likely to either ameliorate or exacerbate them. Combined with theories from political science, these developments enable us to explain individual and political irrationality. In this thesis I will explore a new type of public policy tool, which has been discussed in existing literature but not yet given a systematic treatment. I will call these policies braces. They are intended to structure our environments in ways that combat irrationality. I will explore the relation of braces to existing public policy tools such as nudges and boosts; argue that braces can be given a coherent normative justification; argue that concrete applications for braces exist; and, finally, explore some of the practical issues which could prevent braces from being successfully implemented in our current political environments.

Contents

Introduction.....	1
1. Politics, Psychology, and the Mental Environment.....	4
1.1 Dual-Process Theory.....	4
1.2 Cognitive Biases and the Mental Environment.....	8
1.3 Political implications of Dual-Process theory: Nudges and Boosts.....	14
2. Beyond Nudges and Boosts: Braces.....	18
2.1 Heath's Proposals.....	19
2.2 Brennan's Proposals.....	28
2.3 Other Brace Proposals in the Literature.....	32
2.4 Justifying Braces.....	33
2.5 Braces: Prospects and Problems.....	38
3. Prospects: Bracing Ourselves Against Epistemic Deceptors.....	40
3.1 Deceptors: Conditions and Types.....	41
3.2 Ideal Public Knowledge.....	43
3.3 Fake News (and other Falsehoods): Problems and Brace Solutions.....	47
3.4 Politically Motivated Reasoning: Problems and Brace Solutions.....	56
4. Problems: Braces and Bootstrapping.....	64
4.1 Ethical vs. Practical Problems for Braces.....	66
4.2 The Implementation Problem.....	67
4.3 The Democratic Problem.....	74
Conclusion.....	84
Acknowledgements.....	87
References.....	88

Introduction

If one is a citizen of a modern democracy, the last couple of years have made it easy to feel as though one is living through an age of political irrationality. The election and policies of the Trump administration have created this feeling in the US (Millies et al. 2016); the Brexit vote and the subsequent fractious discussions about how to implement the results of that vote have created this feeling in the UK (Pfeiffer 2016); the electoral strategy and subsequent policies of the Harper government created this feeling in Canada (Heath 2014b, 5–7, 245–46, 255–57). Meanwhile, in mainland Europe, right-wing populist parties like the Forum voor Democratie in the Netherlands and Alternative für Deutschland in Germany have succeeded in the polls while embracing climate change denial (The Economist 2019; Connolly 2019).

Is democratic politics really becoming more irrational? This is hard to answer, and would require a detailed historical analysis to answer definitively (for an argument that democratic politics is *not* becoming more irrational, see Tyler Cowen's entry in Read and Wallace-Wells 2019). What we can say with certainty is that we currently understand more about the sources of both individual and political irrationality than we ever have before. On the individual level, work in cognitive science, particularly in the heuristics and biases literature, has given us detailed insights into the idiosyncratic nature of human cognition: allowing us to identify widespread and pervasive biases which affect our ability to think rationally, as well as highlighting the importance of our environment for stifling or structuring our rational capacities. On the political level, work in political science has comprehensively uncovered high levels of ignorance and misinformation among voters, and suggested a mechanism which explains this phenomenon: the phenomenon of *rational irrationality*, whereby the extremely limited value of each individual voter's vote makes it individually rational for them to be ignorant or to indulge beliefs which flatter their biases. I will discuss both of these scientific developments in Chapters 1 and 2 of this thesis.

If we have identified the factors which cause irrationality, have we also identified measures which could ameliorate these problems? There are two research programs which have suggested such measures: the nudge program, developed by Cass Sunstein and Richard Thaler, and the boost program, developed by Ralph Hertwig and Till

Grüne-Yanoff. The nudge program recommends using nudges: subtle non-coercive interventions in our environments, designed to guide our behaviour in ways that protect us from the harmful (by our own lights) effects of many of our cognitive biases. The program has been very fruitful, resulting in many policy recommendations, many of which have been implemented in democratic societies. The more recent boost program recommends using boosts: non-coercive interventions which aim to increase people's cognitive competences, enabling them to avoid falling prey to cognitive biases (I will discuss both programs more fully in section 1.3). However, while both programs should undoubtedly play a role in combatting irrationality, they have limitations: notably because nudges and boosts are non-coercive there are limits to what they can achieve.

In this thesis, I will argue that there is a new research program emerging in the literature. This new trend shares some of the characteristics of the nudge program, with a focus on changing our environments in order to protect us from our biases: but it also recommends stronger, sometimes coercive, policy interventions. Furthermore, this new trend is best exemplified by the work of two philosophers, Joseph Heath and Jason Brennan, who have a distinct focus on using environmental interventions to improve the quality of democratic decision-making: a focus which I will maintain. I call this new trend the *brace program*, and the policies which it recommends *braces*. Much as there is currently a lively academic debate about nudges and boosts, the aim of this thesis is to make the first move in starting an academic conversation about braces. This aim can be broken down into four specific aims, which I will fulfil in this thesis. I will argue that:

- (1) There is a new trend in the literature, the brace program, which has been endorsed in various forms by a number of authors: I will argue for this point in Chapter 2.
- (2) The brace program can be given a coherent and powerful normative justification: I will argue for this point in section 2.4.
- (3) The brace program can issue concrete policy proposals: it provides theoretical resources which enable problematic elements of our political environment to be fruitfully analysed, and it can suggest specific braces which could be used to combat these elements. I will argue for this point in Chapter 3.
- (4) The brace program is faced with two significant practical problems, both of which suggest that the problems that braces are meant to address have the potential to undermine the possibility of braces being successfully implemented. I will elucidate these two problems, and suggest the kinds of responses that could be given in favour of braces: however, these problems must, I will argue, be addressed for each particular brace that is suggested. I will argue for these points in Chapter 4.

By fulfilling each of these aims, I hope to give the reader a sense of what braces are, how they are justified, the problems that specific brace policies could be used to address, what specific brace policies look like, and what I believe the main problems facing this research program are. My hope is that, by fulfilling these aims, an informed discussion of this new type of policy tool will be possible.

1. Politics, Psychology, and the Mental Environment

Over the past four decades there has been an increasing body of research in cognitive science, philosophy, and multiple fields of psychology suggesting that human cognition can be fruitfully modelled with a dual-process theory of cognition (Evans 2008; Frankish 2010; Frankish and Evans 2009, 10–22). Dual-process theory, and the associated literature on cognitive biases, paints a potentially unnerving picture of human cognition and the ways in which it systematically errs from rationality: furthermore, this is a picture in which the external environment ends up playing a key role in determining whether cognition is rational or not. Some have argued that this picture of human cognition has political implications. In this chapter I will firstly describe dual-process theory. Secondly, I will outline the biases that can and do result from human cognition, and the importance of the environment for our ability to think rationally. Thirdly, I will discuss two of the main political developments to have resulted from dual process theory and the related work on cognitive biases: the nudge program promoted by Richard Thaler and Cass Sunstein and the promotion of boosts by Ralph Hertwig and Till Grüne-Yanoff. Outlining these topics will set the stage for my introduction of the brace program in Chapter 2.

1.1 Dual-Process Theory

Dual-process theories have been proposed in cognitive and social psychology to describe higher cognition, including “thinking, reasoning, decision-making, and social judgment” (Evans 2008, 256, see also 257, Table 1, for an overview of authors who have used dual-process distinctions; see also Stanovich 2004, 35–36, Table 2.1 for further authors who have used dual-process distinctions). These theories divide the human mind into two types of cognition, non-descriptively called Type 1 and Type 2 cognition (Evans and Stanovich 2013, 223–24).¹ There has been remarkable convergence on descriptions

¹ The terms System 1 and System 2 have also been used to describe the two types of cognition that have been observed (Kahneman 2011, esp. 20-21; Frankish 2010, 919–21), however I follow Keith Stanovich & Jonathan Evans’ terminology of Type 1/Type 2 due to the fact that Type 1 is in fact made up of many diverse subsystems (what Stanovich calls the autonomous set of systems (TASS) (2011, 32)), and is thus not accurately described as being one system. This also seems to be a point that Daniel Kahneman appreciates, as he warns the reader not to take his use of System 1/System 2 language literally: “I must make it absolutely clear that [System 1 and System

of Type 1 and 2 as having certain attributes (Evans 2006a, 202; Frankish 2010, 922, Table 1): Type 1 is generally described as unconscious, automatic, fast, and intuitive, whereas Type 2 is conscious, deliberative, slow, and effortful (Evans 2008, 256; Evans and Frankish 2009, v; Kahneman 2011, 20–21; Evans and Stanovich 2013, 223).

In section 1.2 I will look more closely at particular Type 1 processes, but for now a more complete list of the attributes associated with Type 1 and 2 processing can be found in Table 1 (below), distinguished into four main “clusters”.

Table 1 *Common features of Type 1 and 2 cognition, sorted into 4 clusters* (adapted from Evans 2008, 257).

Type 1	Type 2
Cluster 1 (Consciousness)	
Unconscious Implicit Automatic Low effort Rapid High capacity Default process Holistic, perceptual	Conscious Explicit Controlled High effort Slow Low capacity Inhibitory Analytic, reflective
Cluster 2 (Evolution)	
Evolutionarily old Evolutionarily rationality Shared with other animals Nonverbal Modular cognition	Evolutionarily recent Individual rationality Uniquely human Linked to language Fluid intelligence
Cluster 3 (Functional characteristics)	
Associative Domain specific Contextualized Pragmatic Parallel Stereotypical	Rule based Domain general Abstract Logical Sequential Egalitarian
Cluster 4 (Individual differences)	
Universal Independent of general intelligence Independent of working memory	Heritable Linked to general intelligence Limited by working memory capacity

2] are fictitious characters. Systems 1 and 2 are not systems in the standard sense of entities with interacting aspects or parts. And there is no one part of the brain that either of the systems would call home” (2011, 29; see also Kahneman and Frederick 2005, 267).

Type 1 and 2 processing tend to have the attributes outlined in Table 1, however not every Type 1 attribute is always correlated with every instance of Type 1 processing, and similarly for Type 2. For example, both Type 1 and 2 processing can have conscious and nonconscious aspects (Evans and Stanovich 2013, 227); Type 1 processing is not uniformly evolutionarily ancient, which is to be expected given the diversity of systems deemed Type 1 (Evans 2006a, 202–3); and Type 1 cognition is not accurately described as not being rule-based, given that any cognitive process that can be computationally modelled can be described as following rules (*idem*, 204).

The distinction between Type 1 and Type 2 cognition has been attacked due to the fact that the attributes in Table 1 do not always correlate with instances of Type 1 and 2 processing: critics object that there is no feature which is necessary and sufficient for a process to be Type 1 or to be Type 2, and so the distinction has no basis (Melnikoff and Bargh 2018). Defenders of the distinction have responded that the list of features commonly associated with Type 1 and 2 cognition should be thought of as “typical correlates”, which do not define a process as being either Type 1 or Type 2 but rather are commonly associated with these types of cognition. The “defining features” of Type 1 and 2 cognition are, they argue, as follows: Type 2 processes rely on working memory, and Type 1 processes are autonomous, meaning that they do not require any controlled attention and thus place minimal demands on working memory (Evans and Stanovich 2013, 235–36; Pennycook et al. 2018).

One important point is that the reliance Type 2 processing on working memory gives it a strong connection to hypothetical thinking, defined as “the imagination of possibilities that go beyond the representation of factual knowledge about the world”, to the extent that Type 2 processing “is involved whenever hypothetical thought is required” (Evans 2006b, 379). The simulation of alternative possibilities achieved during hypothetical thinking (so called “secondary representations”, which can then be manipulated in order to test hypotheses) must remain separate from representations of the real world (primary representations), as must simulations about possible actions: otherwise, one risks becoming confused between one’s simulated representations and reality (Stanovich 2011, 48–49). In order to achieve this separation, one must carry out the cognitive operation of *decoupling*, which is very difficult to achieve and which must be maintained throughout episodes of hypothetical thinking. One’s capacity for cognitive decoupling is predicted by one’s working memory capacity, thus one can see a link

between hypothetical thinking and working memory (idem, 50–56; Evans and Stanovich 2013, 235–36).²

A further discovery made during research in cognitive science and psychology is that humans are *cognitive misers*: our basic tendency is to deal with problems at the lowest level of computational expense, which means that, by default, we use Type 1 processing (Stanovich 2011, 29; Evans and Stanovich 2013, 237). Combined with the automaticity of Type 1 processes, our miserliness means that we give fast intuitive answers when presented with novel problems; answers which can often be inappropriate or wrong (see section 1.2 below). In order to improve performance, we need to engage in a Type 2 *override* of the intuitive Type 1 answer: using slower, more computationally expensive processing to check our Type 1 answer and, if necessary, attempt to improve it. This picture of Type 1 processes firing automatically and Type 2 processing sometimes being recruited in an attempt to intervene and override incorrect Type 1 responses is known as *default-interventionism* (Evans and Stanovich 2013, 237).³ However, cognitive miserliness can persist even if Type 2 processing is activated: humans always have a tendency to carry out the least effortful cognitive process possible. Sometimes, even though Type 2 processing is engaged, people fail to carry out cognitive decoupling: instead of simulating alternative models of the world, people revert to *serial associative processing*, in which they carry out analytic processing using an inflexible model which is provided by Type 1 heuristics. When carrying out serial associative processing people do not achieve override of their Type 1 intuition, but instead provide a rationalisation of said intuition. Cognitive miserliness can also manifest itself when people begin decoupled simulation, but do not complete it: this is known as *override failure* (Stanovich 2009, 68–69).

Thus far I have described Type 1 and 2 cognition, detailed their core features and their commonly associated attributes, and explained how they interact with one another. In the following section, I will look more closely at how our cognition is influenced by our *mental environment*, and the biases that can result.

² This point is of particular relevance for the discussion in section 4.2.

³ An alternative to the *default-interventionist* interpretation of dual-process theory is the “*parallel-competitive*” (Evans and Stanovich 2013, 237) interpretation, see for instance Smith and DeCoster 2000, esp. 112 for an account of dual memory processes operating simultaneously. I favour the default-interventionist interpretation due to its consistency with cognitive miserliness: given cognitive miserliness, Type 2 processes are unlikely to be engaged most of the time, making a parallel-competitive approach unfeasible (Evans and Stanovich 2013, 237).

1.2 Cognitive Biases and the Mental Environment

Type 1 cognition is firing all the time: it pervades all cognitive functioning, and cannot be turned off (Stanovich 2004, 112). A lot of the time this reliance on Type 1 is not an issue: Type 1 processing is powerful and fast, and many innate Type 1 cognitive modules allows us to do things like rapidly recognise faces, predict the trajectory of falling objects, and interpret the emotions of others (Mercier and Sperber 2009, 150–51, 157; Sauer 2018, 7). In the case of expert judgement, where a skill has been practiced to the point that it has become intuitive, people are able make incredibly accurate judgements almost instantaneously: from fire-fighters who intuitively know that a building is about to collapse, to chess masters who have acquired a repertoire of between 50,000 and 100,000 immediately recognisable patterns, enabling them to make expert moves without having to engage in explicit computation (Kahneman and Klein 2009, 515–16).

However, the chief limitation of Type 1 processes is that they are inflexible, and so they need what Kahneman and Gary Klein call a “high-validity environment” in order to function well (2009, 519). Essentially, Type 1 processes rigidly pick up on certain cues in the environment, and use these cues to generate an intuition that is accessible to the conscious mind. In what follows, I will refer to the elements of our environment which can influence our cognition, both Type 1 and Type 2, as our *mental environment* (Heath 2014b, 13, 17, 23). As will become clear in throughout this thesis, many elements of our environment can be considered to be part of the mental environment: thus, the term is very broad, but this breadth is reflective of just how important the external environment is for our cognition (thus it is a feature, not a bug, of the terminology).

Type 1 processes are only trustworthy in high-validity environments; otherwise, they tend to produce intuitions that systematically deviate from rationality: either from *epistemic rationality* (the measure of how well a belief is mapped onto the world) or *instrumental rationality* (behaving in the world in such a way that you get exactly what you most want given the physical and mental resources available to you) (Stanovich 2011, 6; see also Stanovich 2004, chap. 3). When Type 1 or Type 2 processes deviate from rationality in this way, they give rise to *cognitive biases* (Stanovich 2004, 98).⁴ In the

⁴ On this point there is a disagreement between those like Stanovich, Evans, and Kahneman who think that Type 1 responses often fall short of rational standards, and therefore often give rise to cognitive biases (the *Meliorists*), and those like Gerd Gigerenzer, Peter Todd, Leda Cosmides, and John Tooby who argue that there are good reasons to think that the rational standard for a task is constituted by the *modal response* for that task, which is often a Type 1 response (the *Panglossians*) (Stanovich 2011, 7–10; Cosmides and Tooby 1994; Todd and Gigerenzer 2000, 736–38). I elide

following discussion I will use the standards of epistemic and instrumental rationality, as defined by Stanovich, to measure whether cognition is normatively successful.

Because the workings of Type 1 processes are often not accessible to the conscious mind (see for example Wilson 2002), it is not usually possible to know whether one is in a high-validity mental environment or not, because one does not know what cues a given Type 1 process is picking up on. Furthermore, all intuitions tend to come with a subjective sense of confidence which Kahneman terms the “illusion of validity”, even when they are biased (2011, 209–21). This illusion of validity is partially caused by the fact the Type 1 processing is cognitively easy, and cognitive ease (also known as fluency) creates feelings of happiness, familiarity and trust (idem, 59–70).

So, we always have the potential to be thinking in biased ways, and we often cannot tell introspectively whether we are thinking in biased ways. We also often cannot tell introspectively whether we are in mental environments of high- or low-validity for our Type 1 processes: however, we have reason to think that we are increasingly living in low-validity environments. Many of our Type 1 processes are evolutionarily old, and were adaptive in the era of evolutionary adaption (EEA): an era of small hunter-gatherer communities, with group size at roughly 150 individuals (and a total number of acquaintances of maybe 500 individuals), that persisted up until the agricultural revolution, approximately 10,000 years ago (Stanovich 2011, 112; see Dunbar 2014, 78–80, on group size; see Roberts and Westad 2013, 30–34, on the dating for the emergence of agriculture). For example, Stanovich identifies four fundamental biases in Type 1 cognition, all of which would have been adaptive in the EEA:

1. The fundamental bias towards contextualising problems using as much prior knowledge as possible would have been adaptive in a natural environment where most problems are contextual and it is likely that most of our prior beliefs would be true.
2. The tendency to ‘socialise’ problems, even in situations where interpersonal cues are limited, and;
3. The fundamental bias towards seeing deliberative design where there is none are both likely side-effects of the cognitive adaptations that we evolved in

this debate primarily because some of the key political proposals I wish to assess, those found in Joseph Heath’s *Enlightenment 2.0* (2014b), are mainly informed by the Meliorist position. Furthermore, for a suggestion that Meliorists and Panglossians may be talking past each other on this issue, see Sauer 2018, 10.

order to predict the behaviour of conspecifics and co-ordinate our actions in groups.

4. The tendency towards a narrative mode of thought which imputes animate objects with intentionality is likely an adaptation that allowed us to categorise objects in our environment quickly (Stanovich 2004, 112–15).

However, the modern world makes increasing demands for de-contextualised reasoning and requires the suppression of narrative and social modes of thinking in many domains: for example, juries in a common law legal system are required to put aside their previous beliefs and theories when they assess evidence, a process which requires the override of fundamentally contextual Type 1 processing, and which many juries are unable to carry out despite legal compulsion to do so (idem, 121–23). Stanovich has an apt metaphor for the relation between the modern world and our Type 1 cognition: modern society functions like a *cognitive sodium vapour lamp*. In the same way as sodium vapour lamps (the light source used for street-lights) give off a spectrum of light that causes our evolved colour constancy mechanisms to fail (hence why everything appears monochrome under street-lighting), many elements of modern society constitute low-validity mental environments for our Type 1 processes, often because they contain situations that were simply not present during the EEA (idem, 134–39). The cognitive sodium vapour lamps we must work under in modern society include:

[T]he probabilities we must deal with; the causation we must infer from knowledge of what might have happened; the vivid advertising examples we must ignore; the unrepresentative sample we must disregard; the favored hypothesis we must not privilege; the rule we must follow that dictates we ignore a personal relationship; the narrative we must set aside because it does not square with the facts; the pattern that we must infer is not there because we know a randomizing device is involved; the sunk cost that must not affect our judgment; the judge's instructions we must follow despite their conflict with common sense; the contract we must honor despite its negative affects on a relative; the professional decision we must make because we know it is beneficial in the aggregate even if unclear in this case (idem, 136-137).

Not only are we often biased in these uniquely modern situations, but we are also *exploitable* by third parties who know how to use our biases: notably to sell us things, whether these be lottery tickets, food, coffee, or political candidates (Stanovich 2004, 128; Heath 2014b, 5–7, 200–207, 236–44, 254–57). This environmental perspective helps to set the stage for the political implications of dual-process theory: if our mental

environment has a large influence on whether we can act rationally, then the question naturally arises whether we can actively influence our own environment in order to enhance our rationality (Heath and Anderson 2010; Heath 2014b, chap. 12).

To sum up, Type 1 cognition is always active, and it in environments of low validity it produces intuitions which diverge systematically from epistemic and instrumental rationality: furthermore, such cognitive biases are generally not introspectively detectable. On top of this, we have reason to believe that our modern mental environment is a cognitive sodium vapour lamp which is increasingly hostile to our Type 1 processes, many of which evolved to be adaptive in the EEA but which are now faced with unfamiliar environments in which they produce biased responses. Furthermore, Type 2 processing, while it can lead to normatively preferable outcomes (for example, more true beliefs), can also be used in a purely algorithmic way to rationalise one's pre-existing commitments: that is to say that biases can result from Type 2 cognition as well as from Type 1 (Evans and Stanovich 2013, 229). There are many cognitive biases, and I lack the space to list them all here (for an overview see Kahneman 2011, chaps. 9–22): instead I will focus on a few cognitive biases that will be referred to frequently during this thesis, and will describe other biases in detail where necessary.

1. *Availability Bias*: when making judgements of how frequent a type of event is, or how likely a given event is to occur, people often use the ease with which instances can be brought to mind in order to produce estimations. For instance, people who are asked to estimate the divorce rate in their society are likely to recall instances of divorces. If an instance of the relevant event or type of event can be recalled easily, then this ease is taken as a proxy for that event being probable (Tversky and Kahneman 1973; Kahneman 2011, 129–36). As was discussed earlier, in a natural environment (like the EEA) the underlying heuristic can give valid results: “Availability is an ecologically valid clue for the judgment of frequency because, in general, frequent events are easier to recall or imagine than infrequent ones” (Tversky and Kahneman 1973, 209). However, frequency is not the only thing which influences how easy an event is to remember: if an event is salient (perhaps salacious, such as a celebrity divorce or a political sex scandal), dramatic (e.g. plane crashes or traffic accidents), personal, vivid, or frightening then it will be more easily recalled even if such events happen relatively infrequently. Due to these

other features which make events easy to remember, media coverage can bias our sense of how frequent events are. Media coverage is biased towards rare and emotionally engaging events, making these events overrepresented in our mental environment, and frightening thoughts and images occur to people with ease: these factors combine to give people very warped beliefs about the probability of certain risks (Kahneman 2011, 130, 138–40; Lichtenstein et al. 1978; Slovic et al. 1981, esp. 18-19).

2. *Belief Bias*: people are far more likely to accept the conclusion of an argument if it seems believable, regardless of whether the argument is valid or not. This bias results from the fundamental computational bias towards contextualisation described above: people tend to bring prior beliefs to bear when assessing an argument, and as such if the conclusion does not sound believable we will be biased in favour of rejecting the argument (Stanovich 2010, 107–13; Heath 2014b, 137).
3. *Confirmation Bias*: confirmation bias is best described as a failure to ‘think the negative’. It occurs when people only attempt to confirm their own hypothesis or belief, without thinking about how it could be wrong. This bias tends to occur only when it is one’s own hypothesis which is under consideration (see also myside bias, below): it can be alleviated if one is asked to assess somebody else’s hypothesis, at which point people use falsifying strategies as much as four times as often (Mercier and Sperber 2009, 163–64; Heath 2014b, 131–42).
4. *In-group/Out-group Bias*: people have an intuitive tendency to categorise the social world into different groups, where each group is associated with social stereotypes comprised of “category prototypes, perceived trait distributions, and implicit theories about the social meaning of the [group] category” (Brewer 2007, 696). Once this cognitive categorization has taken place, there is then the potential for affective identification with one’s own group (the in-group) which leads to various forms of in-group favouritism: seeing the limits of in-group as the bound of one’s concern for distributive justice, offering pro-social cooperative behaviour more readily to in-group members,

and being more likely to trust in-group members (*idem*, 698–701). More worryingly, people can develop attitudes of hatred towards members of other groups (out-groups), especially if the relationship between the in-group and the out-group is seen to be one of competition over scarce resources or other group goals (as Marilyn Brewer notes, the perception of the intergroup relationship is ripe for political manipulation) (*idem*, 701–6). Our tendency to be cooperative within our group and ambivalent or hostile towards members of other groups likely has an evolutionary rationale: in the EEA, bounded cooperation with one's group would secure a competitive advantage against other groups (Greene 2013, 22–25). However, in the modern world, there are opportunities for cooperation at a much larger scale than was possible in the EEA: as such, the in-group/out-group bias, if not kept in check, can undermine cooperation in the modern world (Heath 2014b, 92–101).

5. *Myside Bias*: this bias is similar to belief bias, and occurs when people evaluate evidence and test hypotheses in ways that are biased towards their prior opinions and beliefs. This is problematic for critical thinking, because to assess evidence and arguments objectively one should be able to put aside one's prior beliefs (Stanovich, West, and Toplak 2013, 259). Myside bias is very prevalent: smokers are less likely to acknowledge the harmful effects of second-hand smoke, people who consume relatively high amounts of alcohol are less likely to acknowledge the negative health consequences of alcohol consumption, people who are more religious are more likely to believe that religiosity leads to more honesty, and so on (*idem*, 260).

We have now built up a picture of human cognition as chiefly relying on Type 1 processing, which needs reliable high-validity environments in order to produce good responses rather than biased ones. However, our modern world is a mental environment in which Type 2 override is required more and more, due the fact that it contains many environments and situations which are hostile to Type 1 cognition. In the following section I will describe two of the key political developments that has resulted from the literature on dual-process theory and cognitive biases: Thaler and Sunstein's promotion of nudges and Hertwig and Grüne-Yanoff's promotion of boosts.

1.3 Political implications of Dual-Process theory: Nudges and Boosts

Thaler and Sunstein's promotion of nudges is perhaps the first large political development built off of the literature on dual-process theories and cognitive biases (beginning with a series of articles, Sunstein and Thaler 2003; Thaler and Sunstein 2003, and culminating in their book *Nudge* (2008)). Their proposal is political inasmuch as it describes and promotes a new type of public policy tool, nudges, and that it justifies these nudges with a new political philosophy called *libertarian paternalism*. These two elements are sometimes treated as inseparable in Thaler and Sunstein's work (and later in books solely authored by Sunstein), which I will argue to be an error: as such I will treat the two elements separately, aiming to show how they are both related to the cognitive science outlined in section 1. A more recent development in the literature is the promotion of boosts by Hertwig and Grüne-Yanoff: this development draws on a different interpretation of the cognitive science than that which I have outlined so far, but nonetheless it will be instructive to briefly describe this development. My treatment of nudges and boosts will be somewhat brief, as I mainly wish to get them on the table to compare them with braces in Chapter 2.

Building off of the heuristics and biases research of Tversky and Kahneman, Thaler and Sunstein recognise the power of the external environment to influence our Type 1 processing. They call the environment in which people make decisions *choice architecture* (2008, 3), and they recognise that many aspects of choice architecture can influence our behaviour via Type 1 processes: for instance, the way in which choices are framed, the bias choosers have to stick with the status quo, the fact that people are more swayed by prospective losses than prospective gains, and the fact that people often conform to the behaviour of those around them (*idem*, 19–41, 57–65). A *nudge* is a policy tool which is defined as “any aspect of choice architecture which alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives” (*idem*, 6). While the meaning of the term “significantly” in definition is somewhat vague, Thaler and Sunstein insist that nudges must be “easy and cheap to avoid” (*ibid*) in order to differentiate them from the more coercive regulation by mandates, bans, and economic incentives (e.g. fines) that make up a lot of existing public policy (Wolff 2011; Sunstein 2016, 5, 18–20). Examples of nudges include arranging the order of food in a cafeteria to affect people's meal choices (Thaler and Sunstein 2008, 1–3), signing people up for a default green energy provider unless they explicitly choose to opt out (Schubert 2017, 330), and automatically enrolling people into

saving plans at work (again, unless they explicitly choose to opt) (Thaler and Sunstein 2008, 118–19). Taken as a policy tool, nudges have two main arguments in their favour. Firstly, they are non-coercive (or, at the very least, much less coercive than other policy tools). Secondly, while governments can choose not to intervene in a certain choice architecture by nudging, this does not mean that the influencing effects of choice architecture will go away. If governments do not influence choice architecture, then other actors often will: and even if an effort is made to decide arrange the choice architecture randomly, this random selection will still nudge people. In effect, some kind of nudging is often unavoidable: the question really becomes who will be doing the nudging and in what direction people will be nudged (idem, 11; Sunstein 2014, 13–19, 21; 2016, 35–36). One crucial point about nudges is that they aim to alter people’s behaviour by utilising their existing cognitive biases, and do not attempt to directly support people’s Type 2 cognition.

Thaler and Sunstein’s justification for nudges is also inspired by dual-process theory: the justification is provided by their political philosophy, libertarian paternalism. The ‘libertarian’ part of this name stems from the fact that nudges are, qua nudges, always avoidable at no or low cost: if one is automatically enrolled into a saving plan at work, one can always opt out (and if opting out becomes too costly to the point where the freedom-of-choice-preserving libertarian credentials of the policy are put at risk, then the policy can no longer be considered a nudge). The ‘paternalist’ part begins with the insight from psychology that, due to cognitive biases, people often make very bad choices by their own lights: they drink, smoke, and eat too much, often leading to premature death, and they frequently don’t save enough for retirement (Thaler and Sunstein 2008, 7–8, 115–17). The paternalist response of Thaler and Sunstein is to try to design nudges that will encourage people to choose options that will make them “better off, *as judged by themselves*” (idem, 5). The “as judged by themselves” clause refers to the informed and reflective judgements of choosers: those judgements they make when they have the relevant information and are not in situations in which their judgement is undermined by bias or self-control problems. Essentially, the “as judged by themselves” standard reflects the judgements that people make when they override Type 1 cognition with Type 2 (Sunstein 2016, 43–48). The role of dual-process theory in this argument is to demonstrate that when people make choices in many domains they systematically deviate from what they judge to be best. This fact is then used to endorse paternalism by challenging J. S. Mill’s influential anti-paternalist argument in *On Liberty*. Mill holds that:

[T]he only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant (Mill 1859/2015, 13).

Mill's reason for holding that interference for people's own good is illegitimate is that in normal circumstances individuals have a greater interest in and knowledge of their own intentions, feelings, and circumstances than their government or society at large could hope to have (*idem*, 74), and therefore that paternalistic interventions to alter their behaviour for their own good are likely to be misguided. However, given the cognitive biases are systematic, one can argue (*contra* Mill) that there are particular instances in which third parties can intervene for people's own good: namely those instances in which people systematically diverge from their explicitly held goals due to cognitive biases (Sunstein 2014, 8–12; Heath 2014b, 323). While Thaler and Sunstein mainly discuss nudges and libertarian paternalistic justifications as one subject (Thaler and Sunstein 2008, 5–6; Sunstein 2014), in fact nudges and the justification for nudges can be teased apart: in principle it would be possible to justify nudging people towards options which produce the most overall social welfare (welfarist nudges) or towards options which promote social fairness (Rawlsian nudges) (Kelly 2013, 222–25).

Thaler and Sunstein's nudge program has issued in many concrete policy proposals, many of which have already been successfully implemented: for examples, see the 2017-18 report from the Behavioural Insights Team, a UK-based social purpose company which was founded in 2010 as a team with the UK Cabinet Office. The report abounds with examples of nudge (and nudge-like) policies which have developed in consultancy with many national governments and NGOs: including measures to reduce risky play in online gambling (2018, 11), measures to increase voter turnout (*idem*, 31), and measures to increase urgent referrals for cancer treatment in the UK NHS (*idem*, 37). What this demonstrates is that the nudge research program can produce concrete policy recommendations, and that these recommendations can be successfully implemented within current institutional frameworks.

Hertwig and Grüne-Yanoff's promotion of boosts is a more recent development in the literature (2016; 2017), which does not accept the dual-process interpretation of existing evidence in cognitive science which informs the nudge program and (so I will argue) the brace program. Instead, the proponents of the boost program do not share a single view on how the evidence from cognitive science should best be interpreted (Hertwig and Grüne-Yanoff 2017, 974–76, 980). Due to space constraints, I cannot

outline in detail the cognitive science background behind different kinds of boosts (but see *idem*, 974-976 for an overview): instead I will briefly describe what boosts are and how they differ from nudges. Those promoting boosts have more optimistic assumptions about the human capacity to be rational than those promoting nudges: perhaps because of this, boosts are designed not engineer the environment in order to steer people in certain directions by taking advantage of their existing biases (as nudges are), but rather are designed to “foster people’s cognitive and motivational competences” (*idem*, 981). The intention is to either give people competencies or enhance their existing competences in order to enable them to avoid cognitive misfires (*ibid*): this can be achieved either by training people to have particular skills (*idem*, 979) or by altering their environment (for instance, by framing information in ways that are more intuitively understandable, *idem*, 977). Like nudges, boosts are non-coercive: they do not restrict anybody’s choices, or impose significant incentive distortions. Furthermore, unlike nudges, boosts require active consent on the part of the person being boosted: they have to consent to receiving the relevant training or to having their environment restructured (*idem*, 982).

To conclude, it is apparent that the nudge program is heavily influenced by findings in psychology: crucially, it recognises the influence that our external environment can have on our cognition. Similarly, the boost program, while sometimes focussing on fostering competencies at the individual level, also recognises the power of reorganising the mental environment to support our capacity to be rational. In the next chapter, I will contrast these approach with an as-yet-unrecognised new trend in the literature, which also draws heavily from research on dual-process theory and cognitive biases, and which also places central importance on the role of the mental environment to both help and undermine our cognition.

2. Beyond Nudges and Boosts: Braces

Discussions of psychologically informed political philosophy have thus far focused on nudging (for just a few examples see: Bovens 2009; Anderson 2010; John et al. 2011; Rebonato 2013; Levy 2017, 2018; Vugts et al. 2018), with discussions of boosting emerging more recently. However, in addition to the nudge and boost program, I believe that there is a new, and as yet unidentified, trend in the literature which is similarly inspired by research in cognitive and social psychology, and similarly has a focus on the role of the mental environment in human cognition, but which recommends stronger policy interventions and has the potential to address more fundamental problems afflicting political decision-making in modern societies. I will call this trend the *Brace Program* (for reasons that will become clear in the subsequent discussion). The brace program contains recommendations for a broader set of policy responses than the nudge program: it sometimes recommends far-reaching regulatory changes or indeed coercive policy interventions in order to combat the negative effects of cognitive biases. It also provides a new theoretical term, *deceptors*, for identifying and grouping together features of our modern mental environment which routinely cause cognitive misfires. Furthermore, I will argue that the trend can be given a coherent and powerful normative framework which can be used as the basis for justifying individual policies.

I will argue that two notable works expounding the brace program are Joseph Heath's *Enlightenment 2.0* (2014b), and Jason Brennan's *Against Democracy* (2016), and so these are the works I will discuss at most length when describing the new trend (although I will also point out other authors who I will argue can be seen to be promoting braces). Heath and Brennan both have a political focus which I will maintain throughout this thesis: they focus on how a variety of cognitive problems are amplified or exacerbated when it comes to democratic decision-making. These problems typically have to do with the mental environment that shapes political decisions, thereby creating profound problems for the capacity agents can be expected to have in rendering competent decisions. I do not take this political focus to be fundamental to the brace program: braces do not have to be designed to improve the mental environments that shape political decisions, as is made clear in section 2.3. However, I do think that this focus is particularly interesting, hence my choice to maintain it.

That Heath and Brennan both converge on similar proposals is interesting because they hold different political positions (Heath roughly being a left-liberal with a focus on social welfare, and Brennan being a classical liberal with strong libertarian leanings) and have not been in dialogue with one another in their work.⁵ I will begin in section 2.1 by discussing Heath's *Enlightenment 2.0*, and while doing so I will pick out features which I take to be paradigmatic of the brace program. I will then turn, in section 2.2, to Brennan's work in *Against Democracy*, and argue that it shares many of these features. I will also point out other authors who fall within the trend in section 2.3, before turning to the matter of normative justification in section 2.4.

Having described what braces are and how they can be justified, I will turn in Chapter 3 to the prospects for the brace program: I will extend the work already carried out in the brace program by analysing several existing problems in our mental environment as deceptors, and arguing that they can be combatted by braces. In Chapter 4 I will examine the problems with braces: specifically the way in which braces may be undermined by the very problems that they are meant to ameliorate.

2.1 Heath's Proposals

In *Enlightenment 2.0* Heath uses the empirical literature discussed in sections 1.1 and 1.2 to sketch a theory of rationality in which the ability to reason is heavily reliant on the external environment. Observing the ways in which rationality can be strengthened and undermined by the mental environment in which cognition operates, he suggests a range of political measures to try to prevent rationality being undermined and to strengthen it.

Heath begins with an overview of dual-process theory, drawing particularly on Stanovich and Evans. He observes that one of the key features of Type 2 cognition is that it can be made *explicit*: when we have achieved a certain outcome using slow, serial reasoning, we can also generally reproduce how we arrive at that outcome (think, for instance, of checking a logical syllogism). This is in contrast to Type 1 processing. Even in cases when experts make rapid judgements that are the result of decades of experience and learning, and even when these judgements are correct, the experts cannot explain *why*

⁵ Indeed, the only cross-reference I can find between these two authors is a brief mention of Heath by Brennan in the context of a discussion about the use of political science and economics in political philosophy (Brennan 2018, 3).

they are correct, because they have no access to the underlying process that is giving rise to the judgement (Heath 2014b, chap. 1). This attention to modern psychology is an important feature of the brace program, and one that it shares with the nudge program.

Heath then observes that we often enhance and structure our Type 2 processing by utilising our mental environment. For example, we use written language to circumvent our limited working memory capacity: as in the case of using a pen and paper to carry out complicated arithmetic (the products of individual steps in the calculation can be stored on the paper, rather than having them tax our working memory) (idem, 67–68). Another example, demonstrating the flexibility with which we use the environment to enhance our abilities, is the case of a student using a brick wall to sketch walking directions for another student: in this case, the wall is used to convey special information, and it enables the students to get around the fact that natural language is a limited tool for conveying such information (Heath and Anderson 2010, 3). Heath calls these environmental tools *kluges*, borrowing the term from engineering and computer science: a kluge is a solution to a problem that gets something to work (i.e., achieves the desired outcome) without fixing the underlying problem (Heath 2014b, 62–63; see also Marcus 2008, 2–5). With this definition in mind, we can see why the examples given above involve kluges. Having the ability to write the intermediate results of a chain of arithmetic down does not solve the problem of human working memory being very limited, but it does allow the desired result: the ability to do complex arithmetic. Heath argues that rational thought in general is made possible by an enormous array of kluges which enable us to bypass our innate limitations (e.g. working memory capacity), or else turn our individual biases into a socially useful practice. An example of this latter type shows how kluges can form the basis of successful institutions. Individuals suffer from confirmation, belief, and myside biases, which means that individually we struggle to notice our own mistakes. Add into this mixture of biases the fact that people have a *bias blind spot* (they tend to think that “biased thinking on their part would be detectable by conscious introspection”, whereas in fact “most social and cognitive biases operate unconsciously” (Stanovich 2011, 112)), and individual reasoning can start to look hopelessly compromised. A very useful kluge in this case is a culture of contestation, in which we rely on other people to see the flaws in our biased reasoning (and vice versa): we can see this kluge at work in adversarial relations between lawyers in a court of law, in academic peer review, and in various institutional practices of criticism (Heath 2014b, 142–44). This institutional solution is immensely useful, but one can clearly see it is a

kluge: we have not managed to eradicate confirmation, belief, or myside bias, instead we have structured the external environment (in this case an environment largely made up of other people (see Heath and Anderson 2010, 19–21)) in order to work around these biases.

Because our ability to reason is so thoroughly scaffolded by myriad kluges, many of which have come about through long processes of cultural evolution, Heath has a Burkean scepticism about sweeping social reforms. In pursuing such reforms, we run the risk of eliminating useful scaffolding and kluges which we do not perceive. Furthermore, reason, underpowered as it is, is very unlikely to be about to comprehend the vast array of variables (“moving parts”) which constitute complex societies, and therefore it is not going to be possible or advisable to try to rationally construct large social institutions from scratch (Heath 2014b, 84–89). However, while acknowledging the limitations of human rationality, Heath still maintains that we must rely on Type 2 cognition when thinking about social and political questions. While Type 2 cognition is slow, effortful, and limited, it is the only form of cognition we have where we can introspectively determine whether we have reasoned correctly. The vast majority of Type 1 processing simply gives us intuitions or judgements without any introspective access into how these intuitions or judgements are formed: therefore, it is only by carrying out detailed empirical research (of the kind described in section 1.1) that we can work out how our Type 1 processes are actually working. Furthermore, once we have this empirical evidence, it is only from the standpoint of explicit Type 2 cognition that we can decide whether our Type 1 processes are delivering good or bad results (idem, 111–13). Think back to the list of cognitive biases outlined at the end of section 1.2: why is it that the availability bias, for example, seems sub-optimal? Because it seems to return the wrong result in many cases: it makes us think that certain events are more common than they are, due to features like their saliency and their repetition in the media. But why is *this* such a bad thing? Because it clashes with our explicitly held goal of wanting our beliefs to conform to reality (in accordance with epistemic rationality): if we’re trying to judge how dangerous air travel is, we don’t think it’s relevant that airplane crashes are extremely dramatic, because we realise that drama does not influence probability. All of these thoughts require an appreciation of one’s explicit goals in the domain of belief: and it is from the perspective of these explicit goals that the Type 1 availability heuristic often goes astray.

Relying on our Type 1 cognition is a recipe for disaster in many areas of our political and social lives. As mentioned in section 1.2, people are subject to strong in-

group bias, due to the fact that we are adapted for living in small groups. However, this groupishness can be very damaging because it “dramatically limits the scope of cooperation” to members within the group: in large modern societies which are based on expansive networks of cooperation this causes problems. Groupishness sets us up for a series of often intractable collective action problems in which groups become embroiled in intergroup conflict, which is kept alive by moral feeling within each group: that they must stay loyal to their side, and antagonistic to the other side (Heath 2014, 95–96, 150–52; see also Greene 2013, chap. 3, for a detailed account of the psychology behind intergroup conflict). Our intuitive pro-social instincts can cause other problems. Uniquely among primates, humans exhibit a behaviour called “altruistic punishment”: we are willing to punish people who violate the rules or behave non-cooperatively, even if such punishment comes at a personal cost to us. Uninvolved third parties will also often intervene to punish norm-violators or non-cooperators, again at risk to themselves: people simply have a retributive desire to see norm-violators punished (Heath 2014, 148; see also Machery and Mallon 2010, 16–19). In small groups, this behaviour is useful because it heightens the cost of free-riding behaviour, and so can stabilise systems of cooperation. However, in large groups these instincts can destabilise cooperation. As the group gets larger and more anonymous there is more likelihood that somebody will free-ride, either deliberately or simply due to an accidental mistake. Once individuals within the group begin to feel, fairly or unfairly, that others are not cooperating, they will tend to demand that the defecting individuals are punished and they will tend to withdraw their cooperation: this can lead the whole cooperative enterprise to fall apart. Understanding how our pro-social instincts can destabilise large systems of cooperation requires thinking about the problem rationally: we can then implement solutions, many of which are kluges. For instance, many large organisations (schools, armies, companies) subdivide their members into groups (school houses, squadrons, work teams) so that our pro-social instincts can work their cooperation-supporting magic on the small scale, while intergroup competition is often held in symbolic ways (competitions for house points, or work targets) (Heath 2014b, 96–99). On a much larger scale, one of the key functions of the state is its monopoly on force: this is used to prevent people from exercising their retributive instincts against one another by making the most important cases of enforcing norm-compliance the business of the state (*idem*, 152–53).

Individual irrationality, partially caused by cognitive biases, feeds into irrationality in the overall political system: as Heath (and also Bryan Caplan) observe, the ignorance

and biases of voters *are* the environment that politicians (and other elements of our political environment like the news media) must adapt to (Caplan 2007, chap. 7; Heath 2014b, chap. 9). To offer some examples, Heath points out that political tactics like talking points (in which politicians simply repeat the same phrases again and again in interviews and debates) are primarily effective because they exploit the availability heuristic (2014b, 237–44), while political campaigns that promise to be ‘tough on crime’ play into our inherent retributive instincts (idem, 248-257).

Our Type 1 thinking is generally unreliable, if not counterproductive, in dealing with political and social issues: ideally we should use our Type 2 thinking when making political decisions in democracies. We should achieve this by thinking of smart kluges that can be used to circumvent our cognitive biases and the limited capacity of our Type 2 thinking, and promote our collective rationality, particularly in the domain of politics. A further point in Heath’s argument is based on Stanovich’s metaphor of modern society being a cognitive sodium vapour lamp, likely to increasingly confuse our Type 1 cognition. Heath accepts Stanovich’s metaphor, but argues that many of these features which cause our Type 1 processes to misfire are often invented or propagate *because they cause cognitive misfires*. He calls such features *deceptors*:

Deceptor: a feature (context or object) of our mental environment that has been invented or become ubiquitous because of its ability to exploit human cognitive or social psychological processes, leading to misfires (idem, 170–71).

The concept of deceptors is a useful for discussing both features of our environment have been designed in order to cause misfires, and features which may not have been designed for the purpose of causing misfires, but which persist largely *because* they are able to. I will be making use of this concept throughout this thesis, most notably in chapter 3, where I will argue that two problematic features of our political environment which have been identified in social scientific research can be fruitfully analysed using the concept of deceptors (see, in particular, section 3.1 for a further analysis of deceptors). Several examples of deceptors, along with the Type 1 processes that they interact with and the consequences they can generate, are gathered in Table 2 (below).

Having introduced the concept, Heath then argues that deceptors are likely to proliferate in our environment (idem, 177). Deceptors such as fake news stories and conspiracy theories will spread due to people believing and then sharing them. Other times the most powerful driving force behind the proliferation of deceptors is

Table 2 *Examples of deceptors and their consequences.*

Deceptor	Type 1 cognitive process	Consequences
<p>24 hour news and saturation coverage (repeating images and news segments) (Heath 2014b, 120, 239–40) and general trends in reporting to cover negative events (e.g. terrorism) more frequently than positive ones (Dreyfuss 2017; Pinker 2018).</p>	<p>Availability heuristic</p>	<p>Increased ease of recalling events which are commonly reported in the media (Busselle and Shrum 2003). Increased perception of risk where there is very low risk: e.g. 77% of Americans thought that ISIS is a serious threat to the US, according to 2016 poll (Pinker 2018). To put this in perspective, in 2015 North America (Canada and the US combined) experienced 40 deaths as a result of all terrorism (Institute for Economics and Peace 2016, 22), while 2015 was a year in which terrorist attacks <i>increased</i> for OECD nations (of which the US is a member). Despite being a high watermark, it should go without saying that 40 deaths is miniscule: far smaller, for instance, than deaths caused by accidental injury, which accounted for 43.2 deaths per 100,000 people in the US in 2015 (Xu et al. 2016, 3).</p>
<p>Large food and drink portion sizes, both at restaurants and due to the average size of commercially available plates and bowls: number of larger portion sizes in supermarkets has increased 10-fold between 1970–2000, jumbo-sized portions in restaurants are often 250% larger than normal portions, and from 1900–2010 the average size of American dinner plates has increased by 22% (Wansink 2010, 455–56).</p>	<p>Anchoring heuristic</p>	<p>People use portion size to estimate both the norm for how much it is appropriate to eat and to estimate satiety (people judge how full they to a large extent based on how much food remains on their plate). Larger portion sizes can increase the amount that people eat by between 15% to 45% compared to experimental control groups (Wansink 2010, 454–55). The effect of long term overeating is becoming overweight or obese, with associated health risks including diabetes and heart disease. More than 1 in 2 adults in OECD nations are overweight, with average obesity in the population aged over 15 at 19.5% (OECD 2017, 3), with increasing rates expected in the near future (idem, 6).</p>
<p>Conspiracy theories are deceptors inasmuch as the theories are often constructed in such a way that disconfirming evidence is already taken into account by the theory: for example, if one believes the account given in the <i>Protocols of the Elders of Zion</i> of a vast Jewish conspiracy ruling the world, then one also has a readymade explanation for the apparent evidence that the original Russian tract is a forgery (Heath 2014b, 189–90).</p>	<p>Confirmation bias</p>	<p>Belief in conspiracy theories can have many harmful consequences, both individual and social. People who believe conspiracy theories about vaccines often harm their children by failing to vaccinate them (and threaten herd immunity in their communities), those who believe conspiracy theories about HIV/AIDS in South Africa use condoms less frequently, climate change conspiracies unsurprisingly reduce peoples' willingness to reduce their carbon footprint, and harmful social movements such as neo-Nazism are characterised by extensive conspiratorial beliefs (van Prooijen and Douglas 2018, 899–900)</p>

commercial: where there is a potential to exploit our Type 1 cognition to make money, the commercial incentive will generally ensure that this potential is seized upon. Casinos borrow techniques from supermarkets when designing their floor layouts: in both cases, the most tempting products (slot machines and confectionary respectively) are placed near the exits (idem, 183).

Heath's diagnosis is that there is an inherent "potential for a hazardous dynamic to develop in the way that cultural systems as a whole are reproduced, with irrational memes [deceptors] pooling in the population" (Heath 2014b, 357). His suggested response is to take political action in order to try to dispose of or regulate deceptors in our environment, in order to make it less hostile for our Type 1 cognition (idem, 209-210); and to try to improve our environment, making it more friendly to our Type 2 cognition by putting into practice useful kluges (idem, chap. 12, esp. 328). Heath ends his book with a "Slow Politics Manifesto", recommending collective action to achieve careful and informed deliberation about political issues and a cultivation of improved mental environments (idem, 351-352).

With Heath's project described, I can now pick out those features which I taken to be paradigmatic of the brace program. Firstly, a focus on the threat that hostile mental environments can pose to human cognition. In some sense, the problem is really with human beings as boundedly rational creatures with many cognitive bugs: however, once this is taken as a state of affairs which we largely have to put up with, the focus then turns to how our environment can either exploit or scaffold our cognition; either mitigating our cognitive biases or worsening them. With this focus established, the second feature of the brace program is that it recommends a broad range of policies which aim to make our environment more supportive for our cognition: making it easier for us to be rational (in the epistemic and instrumental senses). I will call these policies *braces* (hence the name which I have attributed to the literature trend), which I define as follows:

Braces: Policies which aim to influence our environment (using regulation, coercion, kluges, bans, mandates, uses of technology, etc.) in order to scaffold and support our Type 2 cognition and prevent deceptors from causing cognitive misfires.

I have deliberately left the measures which can count as braces quite vague: this is to allow for the many different ways in which our mental environment can be influenced. However, some precision is certainly needed on the sense in which braces are

environmental policies, and on the ways in which they differ from (and occasionally overlap with) nudges and boosts. The importance of the mental environment for our cognition (see above, and section 1.2) is due to the way in which it interacts with our on-board cognitive resources: either scaffolding or undermining them. However, this relationship can also run in reverse: as Heath and Anderson discuss, our mental environment is often one comprised of other people, where their on-board cognitive resources are part of this environment (2010, 19–21).

Not only do the environmental level and the individual level interact, but sometimes this interaction can be recursive. If a change in people’s mental environment can cause them to build cognitive competencies in their on-board cognitive resources, then there is potential for recursion: and similarly, if a change in a group of people’s individual cognitive resources can result in that group collectively being a superior mental environment for each individual, then there is the potential for recursion. Consider the following two examples of recursion: (1) a boost is implemented which improves the cognitive competencies of a group of individuals, making that group a better social environment for cognition, given that they are better adversarial reasoners. This new environment then further strengthens individual cognition, and so on. (2) A company introduces internal policies which put employees in adversarial reasoning relationships, and as a result these people become better reasoners on an individual level, due to partially internalising this adversarial form of reasoning. These better individual reasoners then constitute a further improved mental environment within the company.⁶

Note that example (1) is *individual-first*: the first move is to strengthen the cognitive competences of individuals. Example (2) is *environment-first*: it recommends firstly changing people’s environment in order to make their cognition more rational. Braces, like nudges and unlike many boosts, are environment-first: while the always interactive and sometimes recursive relationship between the environment and the individual means that environment policies will have (potentially recursive) influences on people’s individual cognition, there remains a distinction to be drawn between those policies which start with the individual and those which start with the environment. Thus I hold that braces are always environment-first policies.

One complication with this characterisation of braces is that some boosts are also environment-first: for example, Hertwig and Grüne-Yanoff argue that people’s cognitive

⁶ Whether these specific recursive relationships would obtain is of course an empirical matter: I use them to illustrate the dynamic of recursive relationships between the individual and the mental environment.

competences can often be fostered “by redesigning aspects of [their] external environment” (2017, 980). In part, my response to this is simply to admit that there may be policies which can be equally described as boosts or braces: namely policies where there is an environmental intervention which then leads to increased individual competencies. However, there are also key distinctions to be made. Many of the braces I will discuss in the rest of this thesis are environmental interventions which are intended to make people more rational but which are unlikely to build their individual competencies; still others may improve people’s competencies so long as the brace is in effect, but have no ability to create longer-lasting increases in competence. Both of these points make clear the contrast with boosts, where the intention is to create effects which “should persist once (successful) intervention is removed” (idem, 974, table 1).

To pursue another line of attack, one can ask why are braces different from nudges? There are three answers to this question, one of which also serves to distinguish braces from boosts: thus I will start with the other two. Firstly, braces, unlike nudges, sometimes have the aim of supporting people’s Type 2 cognition: of directly enabling them to reason more effectively (a good example of this is the adversarial reasoning kluge). Secondly, braces do not use the same causal pathway to influence people as nudges do: nudges take advantage of our pre-existing biases to guide our behaviour, whereas braces attempt to both support our Type 2 cognition and remove misfire-causing deceptors from our mental environment (on the connection between these two elements, see below). Neither of these aims requires using people’s biases to influence their behaviour. The third answer, which also serves to further distinguish braces from boosts, is that braces can be coercive. Particularly in the case of braces which are intended to combat deceptors (see below), braces may take the form of coercive regulations intended to be imposed on private enterprises: in stark contrast to nudges and boosts, neither of which can be coercive.

Thus far I have clarified the difference between nudges, boosts, and braces: the latter being environment-first policies which can use coercion. At the margins, there may be braces which could be equally well described as boosts or nudges, but as a category I believe it is useful to single out braces because of their aforementioned features. A further feature which is unique about braces is their aims. Braces have the dual aim of scaffolding our Type 2 cognition and of combatting deceptors: however, both of these aims are, I argue, interlinked. The ultimate aim of braces is to improve our mental environment in order to allow us to be more rational. This aim can be broken down into

two mutually supporting projects: a negative project, to rid our environment of misfiring causing deceptors which undermine our capacity to be rational, and a positive project, to support our Type 2 cognition by designing our mental environment to be conducive to reasoning (for instance, but utilising the adversarial kluge). The negative project is more likely to utilise coercive means, because it is more likely to involve regulating private enterprises, while the positive project is more likely to involve trying to create environments which interact with people's on-board cognitive resources in ways which enable them to be more rational. With the definition of braces clarified, I will now turn briefly to different types of brace policy, before discussing Jason Brennan's proposals. I will return briefly to the relations between braces, nudges, and boosts in the conclusion.

The brace policies that Heath recommends range from those which seem to be justified on paternalistic grounds to those that are justified on grounds of social welfare. For an example of the former, see his endorsement of a 2012 proposal by New York mayor Michael Bloomberg to ban the sale of sugared drinks in bottles or cups larger than 16 fluid ounces (approx. 473 ml): the rationale behind endorsing this proposal is that changing people's environment in this way will enable many people to achieve their explicitly stated goal of losing weight (idem, 314-317). For an example of the latter, see Heath's suggestion that we should enforce a ban on lying in political advertising, and also prohibit the use of images, music, and sound effects in such adverts (idem, 345). Here the goal is to prevent viewers from being misled by lies or emotionally manipulative content in political adverts, with the benefit seemingly being the public good for a democratic society of having informed voters. In section 2.4 I will argue that both types of brace can be given one underlying justification. Before turning to this issue, however, I will briefly describe Brennan's arguments in *Against Democracy*, and argue that Brennan is also arguing for braces.

2.2 Brennan's Proposals

Brennan's book falls into the tradition of works arguing that the widespread voter ignorance and irrationality discovered by political science favour more restricted government or less influence from voters on government.⁷ However, I will argue that

⁷ Other examples from this tradition include Bryan Caplan's *The Myth of the Rational Voter* (2007), Ilya Somin's *Democracy and Political Ignorance* (2013), and Christopher Achen and Larry Bartels' *Democracy for Realists* (2016).

Brennan can also be seen as promoting braces: specifically, like Heath, promoting braces designed to improve the quality of democratic decision-making. It may seem strange to argue that Brennan is arguing in favour of braces that will improve democratic decision-making, when Brennan takes himself to be arguing for forms of *epistocracy*, not for an altered form of democracy (Brennan 2016, 19–22). Brennan defines epistocracy such that “a political regime is epistocratic to the extent that political power is formally distributed according to competence, skill, and the good faith to act on that skill” (idem, 14). However, many of Brennan’s epistocratic proposals in fact seem to recommend a mixture of epistocracy and democracy: as he says himself, “[m]any forms of epistocracy worth considering have some of the same institutions we find in democracies” (idem, 208), and indeed many of his suggestions maintain democratic institutions such as public voting. As such, I will concur with Robert Talisse in counting Brennan’s proposals as epistocratic modifications to current democratic institutions (Talisse 2018, 9–11). From this perspective, Brennan is pursuing a similar project to Heath: suggesting modifications to our existing democratic institutions in order to make democratic decision-making more rational. Having tentatively placed Brennan’s proposals in the same category as Heath’s, I will now support this claim exactly with an investigation of his proposals.

Brennan reviews the vast literature on voter’s ignorance of politically relevant factual information: information which is empirically supported, and which bears on important policy questions or information about prospective candidates in elections (such as their policy positions). He concludes that the average level of politically relevant factual knowledge in democracies is low, a conclusion supported by numerous other surveys of the available evidence: the average voter has low levels of information about who the candidates in any given election are, often does not know which party controls the legislature, they have low levels of knowledge about economics, they often do not know even the rough breakdown of their national budget, they are generally ignorant or misinformed about matters of policy, and so on (Brennan 2016, chap. 2; see also Caplan 2007, chaps. 2–3; Somin 2013, chap. 1; Achen and Bartels 2016, chap. 2; Talisse 2018, 2). Brennan puts this ignorance and misinformation down to several factors. One of the key factors is that people process politically relevant factual information in biased ways, often conforming their assessments of evidence to try to fit their prior beliefs (confirmation bias) or to maintain beliefs which are important for their group identity (this phenomenon will be discussed at greater length in section 2.4 below) (2016, chap. 2).

The cognitive biases that plague voters may be inherent to human cognition, however Brennan puts their large influence in the political domain (and the influence of ignorance in general) down to the *environment* in which individual voters process this information (idem, chap. 2). Specifically, this is an environment in which the kind of incentives that can motivate people to combat their cognitive biases and think more rationally (specifically to try to achieve epistemic rationality in their beliefs about politically relevant facts) are absent. Achieving epistemic rationality in the domain of politically relevant factual knowledge would mean not indulging false beliefs (even if they flatter your political persuasions or if believing them is important for maintaining your social identity among your peers), actively seeking out accurate information in order to inform one's decisions in elections, and attempting to override one's cognitive biases. Doing this would take a lot of time and effort, which could be used to pursue people's other goals, and it comes with potential costs to one's social standing (if one acquires true beliefs which is largely held to be false, or even unpleasant, in one's community). However, the expected pay-off of engaging in all this activity is vanishingly small, because each voter only gets to cast one vote. Individual votes only change the outcome of an election in tie-break scenario, and the most optimistic estimates put the chance of this happening (in the US) at one-in-a-million: and this is only the case if the individual voter lives in a swing state and votes for a major political party. As such, individual votes have essentially zero value in terms of their ability to influence the outcome of elections, and so, in Brennan's analysis, it is obvious why the average voter does not invest the time and energy required to become informed and overcome their biases: acquiring political knowledge has no benefit (idem, 30–32). Thus, this environment is one in which it is not individually instrumentally rational for voters to be epistemically rational: when voters are epistemically irrational, they are displaying a kind of *rational irrationality* (Caplan 2007, chap. 5).

While individually this situation is instrumentally rational, it has harmful effects on the collective level. Brennan's chief concern is that citizens have a right to competent government, a requirement which he expresses in the *competence principle*:

It is presumed to be unjust and to violate a citizen's rights to forcibly deprive them of life, liberty, or property, or significantly harm their life prospects, as a result of decisions made by an incompetent deliberative body, or as a result of decisions made in an incompetent way or in bad faith (idem, 141).

The incompetence that Brennan refers to is chiefly incompetence in an epistemic sense: being ignorant or evaluating information irrationally (idem, 151-152). Policies decided upon by an incompetent body, a body which is epistemically irrational, impose a risk of harm upon citizens, hence such decision-making power should not be given to an incompetent body (idem, 154). Because voters are epistemically irrational in many cases, Brennan holds that they are an incompetent body. Like Heath, Brennan analyses the problem of epistemic irrationality in the domain of political knowledge as being the result of an environment which is hostile to rationality: hostile because it does not provide the correct incentives required to combat the influence of our innate biases. This focus on the potential of the mental environment to undermine our cognitive abilities is common to both Heath and Brennan. Common too is the environmental nature of Brennan's proposed solutions.

Several of Brennan's solutions would straightforwardly alter the institution of voting in order to encourage voters to be more epistemically rational (by changing their incentives). For example, Brennan suggests that a system of sortition could be used in lieu of elections with a full franchise, in order to try to make voters better informed. Firstly, a demographically representative sample of the population would be selected by lottery. This sample would be substantially smaller than the total population, and thus each vote would count for more than current votes count for. Secondly, the sample would be educated on the platforms of each party and would partake in deliberation with one another in order to improve their political knowledge, before then being allowed to cast their votes (idem, 214-215). This suggestion has both environmental and individualistic elements: its two crucial environmental components being the changing of incentives so that individual votes count for more than they currently do and the placing of voters into deliberation with one another, making use of the environmental kluge outlined by Heath in which our desire and ability to point out the flaws in the arguments of other people are harnessed to counteract our biases. Brennan also offers another, more realistic, environmental solution, which also fundamentally involves changing people's incentives in order to motivate them to be epistemically rational in the domain of politics. He suggests that governments could provide monetary incentives to encourage citizens to become politically informed: for example, they could offer a tax-credit of \$1000 for citizens who can pass a test of politically relevant factual knowledge

(e.g. “introductory microeconomics and introductory political science” (idem, 213)).⁸ The intention is to incentivise people to become epistemically rational in this domain by making the benefit of doing so greater. This can be thought of as a brace designed to alter voters’ environments to take advantage of the attractive power of monetary incentives in order to encourage them to become informed, due to the fact that their current environment does not provide such incentives and as such leads to a lack of epistemic rationality in the domain. Brennan’s environmental focus, in both his diagnosis of the problem of voter ignorance and irrationality, and his proposed solutions is the primarily similarity between his work and Heath’s, and the basis for his inclusion in the list of authors who can be seen to endorse the brace program.

2.3 Other Brace Proposals in the Literature

Other authors can also be seen to be endorsing braces, though not all of them have shared the political focus of Brennan and Heath. Bruno S. Frey and Alois Stutzer have argued that people systematically overestimate the happiness that they will gain from acquiring extrinsic goods (such as status, money, and possessions) and underestimate the happiness gains from intrinsic goods (such as spending time with family and friends). They further claim that this error in estimation is partly caused by the commercial environment that people inhabit: because advertising primarily promotes extrinsic goods, people can be influenced to perceive that these goods will make them happier than they in fact will. Essentially, they argue that we mispredict what will make us happy due to biases in our cognition; biases which are aggravated by commercial advertising. To combat these problems, Frey and Stutzer argue that we should endorse policies designed to encourage people to favour intrinsic goods more heavily: for instance, policies that limit total working hours or restrict opening hours for commercial enterprises (Frey and Stutzer 2006). By changing people’s environment, quite coercively in this case, Frey and Stutzer hope to correct for the predictable cognitive misfires that people fall prey to when trying to predict what will make them happy: it is this

⁸ In *Against Democracy*, Brennan offers this policy in a slightly different context. He is arguing that people could be given a tax credit for passing a test which *also* determines whether that person can vote or not: I have modified his example to exclude the idea that one’s right to vote depends on passing the test. However, Brennan has separately endorsed my modified example in a televised interview, see Paikin 2016, 23:25-23:50.

environmental intervention to combat cognitive misfires that qualifies their proposals as braces.

Neil Levy has endorsed what he calls “ecological engineering”, in order help people to manage their responses to temptation given bounded rationality and limited self control: for example, he suggests that policies to reduce the density of and number of outlets selling tempting goods (e.g., alcohol, fast food, etc.) in order to allow individuals to manage their limited cognitive resources more effectively (Levy 2012, 598–99). Like Heath and Brennan, Levy recognizes the importance of the environment in undermining or supporting our cognition, and he suggests altering this environment in order to protect people from elements which can undermine their ability to act as they would like to (in other words, to be more fully instrumentally rational). Indeed, his proposed policy is one that I would define as a brace.

To offer a final example, Regina Rini argues in favour of regulating people’s informational environment: specifically by attempting to track the trustworthiness of testifiers on social media (I will discuss this proposal at greater length in section 3.3). The aim of this proposal is to prevent people from acquiring false beliefs which can subsequently negatively affect the political process in democracies (Rini 2017). Rini’s proposal qualifies as a brace because she intentionally focuses on altering people’s environment in order to promote epistemic rationality by preventing false beliefs and attempting to spread true beliefs.

Thus far I have argued that a new literature trend, the brace program, can be identified. This program recommends the use of braces to scaffold Type 2 cognition and try to prevent or mitigate cognitive misfires, and it also offers a theoretical term, *deceptor*, for identifying problematic features of our environment which could be remedied with the use of a brace. However, as I have mentioned previously (and as is made clear by the examples of braces that I have provided), braces are at least sometimes coercive. As such, it is necessary to look at what kind of justification can be offered for braces. In the subsequent section I will argue that braces of many different kinds can be offered a unified justification.

2.4 Justifying Braces

Braces are sometimes paternalistic: such as Levy’s promotion of regulating the density of fast food restaurants, which is for the benefit of consumer’s own health.

Braces are sometimes welfarist: as with Brennan's policies intended to produce more competent government, in order to protect individuals from the incompetent decision-making of their collective fellow citizens. They are often coercive, usually because they can only achieve the required manipulation of the environment by mandating that private enterprises behave in a certain way. In this section I will argue that there is a justification available for braces which can show that paternalistic and welfarist braces both derive their normative justification from the same source, and that this source provides a justification which is quite powerful and therefore has great potential to justify coercive measures. My argument here will be partly interpretive and partly constructive. Heath and Brennan both appear to subscribe to my proposed justification, and I will be using their works to construct my argument. However, the justification is intended to be powerful regardless of whether it was intended by Heath or Brennan (or any of the other authors I have identified as promoting the use of braces).

If braces are intended to promote Type 2 reasoning and prevent cognitive misfires (where these are deviations from epistemic or instrumental rationality), then a superficial justification of braces could be that they promote rationality. This seems to be Heath's justification for his proposals: he argues that promoting reason is to be encouraged given that reason is:

[T]he basis of human freedom and autonomy. It is the set of rules that we follow when we want our beliefs to correspond to reality, when we want to avoid failure in the pursuit of our objectives, and when we want to agree on principles for living life in common (idem, 356).

The chief problem is getting from the promotion of reason to the justification of coercive policies. Sometimes the route from one to the other is quite uncomplicated: in the paternalistic cases peoples' irrationality prevents them from achieving their explicitly formulated goals, and so undermines their own instrumental rationality. Because this is a case where a significant number of people can generally be taken to *want* the outcome that the brace is intended to promote, one can offer individually compelling reasons for supporting paternalistic policies aimed to remedy this. However, in welfarist cases things are not so simple. Take Heath's proposed brace of regulating political advertising to prevent falsehoods or emotional manipulation. The intention is certainly to promote rationality, notably epistemic rationality. However, as Brennan points out, promoting epistemic rationality in this context is not necessarily in the individual interests of those who will be watching the advertisements. As described above, voters' mental

environment is one in which it can be instrumentally rational for individuals to be epistemically irrational (Caplan 2007, chap. 5). Rather than getting into a discussion about what exactly counts as rational behaviour for individuals in various contexts, I think a less controversial justification can be offered in favour of braces: one which will enable us to decide whether braces are justified in cases like the one outlined above, in which epistemic and instrumental rationality conflict.

The seeds of this justification can be found in some of Heath's other writings, and it is useful to begin with Heath's Rawlsian characterisation of society as a "cooperative venture for mutual advantage" (Heath 2014a, 147; Rawls 1999, 4). When individuals restrain their self-interest and adopt moral rules which govern interpersonal relations, the benefits produced for others generally outweigh the losses incurred by the individual. When such rules are jointly adopted, they produce mutual benefit: which is to say that they establish a system of cooperation (Heath 2014a, 148). That cooperation is mutually beneficial is made particularly stark by the fact that, often, the alternative to a system of cooperation is a collective action problem: a case where individuals can best pursue their own individual goals by imposing some kind of cost onto others, and where the aggregation of these imposed costs leads to a collectively suboptimal outcome (Heath 2001a, 170–76; 2006, 313). Furthermore, the mutual benefit of cooperation can be articulated into a normative principle.

If we imagine two individuals considering various courses of action to take in an interaction, one can identify the set of outcomes in which both individuals end up better off (as they subjectively define "better off"): these outcomes are then obviously superior (from the perspective of both participants) to those outcomes where both individuals end up worse off (as is likely to be the result from them both engaging in purely self-interested action). This basic idea can be articulated into the *Pareto-superiority criterion* (Heath 2014a, 151–52): "whenever it is possible to improve at least one person's condition without worsening anyone else's, it is better to do so than not" (Heath 2014c, 9–10). Holding this principle commits one to promoting cooperative outcomes, where these are understood to be Pareto efficient: positive-sum outcomes in which it is impossible to reallocate resources in order to make one individual better off without also making at least one other individual worse off (*idem*, 10).⁹

⁹ Beyond the efficiency of an allocation, there is also the issue of whether an allocation of goods is equal. This is a very important issue, and I believe that an account of equality can be provided in line with the contractualist account of efficiency already provided: see Heath 2014a, 152–55.

I hold that this is the fundamental normative principle which justifies the use of braces: specifically, braces are justified when their use promotes or protects positive-sum cooperation.

Brennan also endorses such positive-sum cooperation as the chief goal of public policy, echoing Heath's Rawlsian sentiment that "[s]ocieties are cooperative ventures for mutual gain" (Brennan 2011, 120). He argues that voters have a duty to vote for the common good, meaning "policies good for all" (idem, 112). He argues that there are goods such as "personal and physical integrity, mental and physical health, some wealth, some degree of education, opportunities for economic advancement, some ability to influence others" (ibid) which are useful for all regardless of their individual conceptions of the good life: promoting these goods is therefore something that is in the common interest, and promoting them requires cooperation (idem, 114).

My position is that the kinds of irrationality discussed by Heath, Brennan, and others are chiefly problematic because they threaten to undermine mechanisms of cooperative benefit. This justification also puts the importance of epistemic and instrumental rationality in perspective. Epistemic irrationality, caused, for instance, by the spread of false information, undermines cooperation because it threatens the ability of groups of individuals to coordinate their actions in cooperative ways: if one constituency in a society believes that climate change represents a serious threat to the well-being of the members of that society, but another constituency believes climate change to be a hoax, then they will be unable to effectively coordinate their actions to produce a response to this problem. Another point is implicit in the previous one: if false information spreads throughout a society, then it undermines the ability of that society to even have a productive discussion about what courses of action to take. When there are features of our mental environment which were invented or persist because they cause epistemic irrationality, we might term these features *epistemic deceptors*.

The relation between positive-sum cooperation and instrumental rationality is more complex. On a very basic level, cooperation relies on people being instrumentally rational because it relies on people fulfilling their own ends by working together: indeed, one of the attractions of arguments based on positive-sum cooperation is that it should be relatively easy to convince people to take part in it, given that doing so serves their own interests (Heath 2014a, 148). However, when dealing with more complex issues,

However, for present purposes the focus of the argument is on the ways in which efficiency gains can be lost when cooperation is undermined.

such as trying to ensure that voters are informed, cooperation is undermined because it is instrumentally rational for people to be epistemically irrational: and this epistemic irrationality undermines cooperation. As this example shows, sometimes the link between instrumental rationality and cooperation is quite complex.

Considering the factors which can sustain collective action problems demonstrates how cooperation can be undermined by epistemic factors, in that people can fail to recognise the collective action structure they are trapped in, but also by *motivational factors*. Take Heath's example of fishermen from Portugal and Spain and fishermen from Newfoundland who found themselves locked in a collective action problem in which both groups were collectively overfishing cod stocks in the Grand Banks, eventually leading the stocks to collapse (a classic "tragedy of the commons" (Hardin 1968)). We can assume that all fishermen involved would have preferred that the stocks did not collapse: that is to say, it was in their interests for the stocks not to collapse. The failure to follow through and act in order to serve this preference can have two explanations, which are non-exclusive. The first is an epistemic error, specifically the error of not achieving a rational insight into the structure of the collective action problem: namely, not realizing the likely effects of one's actions and not realizing that both your own group and the other group face identical incentives. The second is a motivational problem. It is possible, and likely, that being locked in this collective action problem will lead both groups to simply hate one another, with each wanting to 'stick it to' the others regardless of the consequences. Even if the collective action structure is recognised, and both groups realise that they must cooperate in order to prevent the stocks from collapsing, it can still be extremely difficult for each group to override their dislike of the other group in order to actually agree to a cooperative scheme (Heath 2014b, 150–51). This intergroup hatred can be maintained by a powerful mixture of "myside bias, in-group solidarity, and retributivism" (idem, 150): and it does not require great leaps of imagination to believe that elements of the mental environment, such as political campaigns and advertising, can be used to stoke such intergroup hatred, thus serving as *motivational deceptors* which make it harder for people to achieve the outcome that would be collectively optimal, and avoid the outcome that would be collectively disastrous.

I will not discuss motivational issues further in this thesis, for three reasons. Firstly, I simply lack the space to fully cover them alongside epistemic issues. Secondly it is easier in epistemic cases to say what the standards of correctness are, whereas with

motivational issues there is inherently more controversy. Thirdly, in many cases where there are motivational problems there are epistemic issues also, as in the case of the cod fishery. If the epistemic issues can be resolved, such that all parties genuinely realise the prospective benefits of cooperation (and the consequences of non-cooperation) then overcoming motivational issues will, one hopes, become substantially easier than in the case where epistemic errors can prop up motivational issues: where one can continue to think of the other actors as ‘the bad guys’, rather than recognizing the structure that both parties are trapped in.¹⁰

Justifying braces in terms of their ability to promote positive-sum cooperation also provides a unified justification for both paternalistic and welfarist braces: in both cases what is at issue is people’s welfare, which in one case is being undermined by their own inability to act on their preferences, and in the latter cases is undermined by the actions of other people. Such a justification also lends braces powerful normative credentials: arguments in terms of positive-sum cooperation are powerful primarily because positive-sum benefit is a very thin ethical concept which can be appreciated as important by people who might have different thicker conceptions of the good life. In an illustrative turn of phrase, Heath describes efficiency (understood as Pareto efficiency) as “the kind of value that allows people to get along without shared values” (2001b, 35, see also Chapter 2 in general): positive-sum cooperation can be beneficial even when people hold very different substantive values.

While I hold that arguments based on positive-sum cooperation are very powerful, any such argument in favour of braces can of course be defeated if that brace raises other significant normative issues: for instance, if it undermines autonomy, generates grievously inequalitarian outcomes, or deprives people of vital liberties. Such ethical concerns should certainly be addressed when considering any particular brace, but I will not pursue them much further in this thesis: my central focus will rather be on whether the brace program can issue in concrete policy recommendations, and on whether these braces could be implemented in our existing institutional environment.

2.5 Braces: Prospects and Problems

In the remainder of this thesis I will argue that the brace program has both prospects and significant problems. In exploring these prospects and problems, I will

¹⁰ Many thanks to Joel Anderson for discussing this point with me.

maintain Heath and Brennan's particular focus on using braces to improve democratic decision-making. While braces need not be designed to maintain cooperation in the domain of politics and public policy specifically, I find the examples designed to maintain cooperation in this domain the most interesting. As such, my subsequent arguments will be focused on the use of braces to improve the quality of decisions made by democracies.

Firstly, I will discuss the prospects for the brace program. Unlike the nudge program, which has resulted in numerous policies being suggested and implemented, the brace program has thusfar produced comparatively few concrete policy proposals. While both Heath and Brennan make some concrete suggestions, the number pales in comparison to the various nudges which have been suggested: Heath is quite open about this, pointing out that the majority of his book consists of diagnosing the problems that we face, with the more positive project emerging only briefly at the end of the book. He describes himself as being concerned not with "trying to solve problems [...], but rather [with] encouraging other people to think about how to solve them in a particular way" (Heath 2015b). I will pursue this project in Chapter 3, arguing that there are existing problems in our mental environment which can (a) be analysed as epistemic deceptors, and (b) can potentially be combatted by braces. My intention is to demonstrate that, like the nudge program, the brace program can be fruitful, and can generate concrete policy proposals.

Having argued that the brace program can issue in concrete proposals, I will then use these proposals as useful examples when discussing the problems that face the brace program: in Chapter 4 I will present two key practical problems which stand in the way of implementing braces. Neither of these problems have easy answers: instead, I will aim to make both problems as clear as possible, and suggest potential resources that proponents of braces could use to answer them.

3. Prospects: Bracing Ourselves Against Epistemic Deceptors

In this chapter I will demonstrate how the brace program can issue specific policies for combating epistemic deceptors which have arisen in democracies. It is not the case that all braces must be responses to the influence of deceptors (some braces are designed to scaffold our Type 2 cognition rather than combat deceptors), however I have chosen my examples such that they involve deceptors *and* can be potentially combatted by braces: my intention is to demonstrate that both theoretical concepts have concrete applications. My two examples are fake news and politically motivated reasoning, both of which threaten to lead to false beliefs among citizens. I will aim to show that both problems can be analysed using the concept of deceptors, and that they require environmental responses: I will outline specific braces, based on proposals in the literature. As well as demonstrating that an analysis in terms of deceptors and braces can be applied to these particular cases, the analyses in this chapter will also offer specific content which can be discussed in Chapter 4 when addressing the practical problems facing the implementation of braces.

Firstly, I will distinguish three different types of deceptors to set the stage for the analyses of fake news and politically motivated reasoning as involving deceptors. Secondly, I will describe the kind of knowledge that we should ideally want citizens in a democracy to have about policy relevant facts:¹¹ this will provide a benchmark for understanding why fake news and politically motivated reasoning are problematic. Thirdly I will describe what fake news is, why it is potentially so problematic, why (I argue) it involves deceptors, and what kind of brace policy could potentially combat the negative effects of fake news. Fourthly, I will describe what politically motivated reasoning is, clarify why it is problematic and why it involves a deceptor, and suggest several braces which could be used to combat it.

¹¹ I will focus in on exactly what is meant by ‘ideally’ in the relevant section.

3.1 Deceptors: Conditions and Types

A deceptor is a feature (context or object) of our mental environment that has been designed or become ubiquitous because of its ability to exploit human cognitive or social psychological processes, leading to misfires. For the analyses that I will carry out in this chapter, it will be useful to label different parts of the definition of deceptors. Working backwards, we can begin with the fact that deceptors must lead to misfires: call this the *misfire condition*. A misfire is a deviation from epistemic or instrumental rationality, as defined by Stanovich (see section 1.2). Secondly, a deceptor must achieve this misfire by interacting with a known human cognitive or social psychology process: this is a fairly modest requirement, which I shall call the *causality condition* on the basis that this condition simply requires that the deceptor causally interact with human psychology.

Thirdly, the deceptor must have been invented or become ubiquitous due to its ability to fulfil the causality and misfire conditions. This third condition concerns why a deceptor exists in our environment, and why, once it exists, it will proliferate. Heath argues that there are three different forces at work in producing and spreading deceptors: *pooling* (once deceptors are present they are difficult to get rid of, due to their misfire-causing properties, and so they will ‘pool’ in the environment’), *contagion* (due to their misfire-causing properties, deceptors are good at reproducing themselves), and *pumping* (deceptors can be used to ‘pump’ people for money, and thus there is an economic incentive to create them) (Heath 2014b, 177). There is an important distinction between these first two factors and the last one, and so I will analyse Heath’s list into two separate conditions. In the case of pooling and contagion, what is really doing the work of propagating the deceptor is a quasi-evolutionary process: the deceptor need not have been specifically designed to be a deceptor, but once it is present in the environment its misfire-causing qualities will lead it to either stick around or spread. No intentionality is needed here: the deceptor simply needs to exist (for whatever reason) and then (if it truly *is* a deceptor) the quasi-evolutionary process will kick in. Thus, if a deceptor is persisting or spreading through the environment due to this quasi-evolutionary process I will say that it meets the *evolution condition*. In the case of deceptors being invented or spreading because of their ability to pump money, there is a greater level of intentionality involved. For example, Heath offers the example of advertising utilising numerous techniques designed to exploit our Type 1 processing, by building emotionally resonant associations, utilising music and images, and generally cutting down on language (with its strong association with Type 2 processing). Each of these techniques was developed piecemeal,

and so there is still a strong evolutionary component to any explanation of why these advertising techniques exist, but once they exist and are shown to work, they are intentionally used in order to attempt to sell products (idem, 200–207). However, intentionally designed deceptors are not only used to sell products: as Heath later points out, they are also used to gain support for political candidates, via political advertising which utilises the same tricks as commercial advertising (idem, 344–346). Therefore, intentional deceptors can be designed to do more than simply pump money. To capture the broadness of their purposes, I will say that a deceptor which is intentionally designed to cause misfires meets the *means condition*, as a deceptor of this type is serving as a means to some other individual or group’s ends. Often a deceptor will persist or spread due to meeting both the *means condition* and the *evolution condition*, and so I will hold that something is a deceptor if it meets the *misfire condition*, the *causality condition*, and an inclusive disjunction of the *means condition* and/or the *evolution condition*.

I have already distinguished (in the previous chapter) different ways that deceptors can be problematic: either epistemically, by leading us to have false beliefs, or motivationally, by fostering non-cooperative motivations. However, for the analyses that I will carry out later in this chapter, it is useful to distinguish some further types of deceptor. These types of deceptor are not distinguished based on the kind of problem which they cause, but rather on what kind of thing the deceptor in question is. While I do not claim that the following list (see Table 3 below) is exhaustive, I think at least three different types of deceptors can be distinguished.

As can be seen in Table 3, I will argue that fake news, particularly in the form of online articles and videos, is a media deceptor. I will also argue that the particular social dynamic which exacerbates politically motivated reasoning is a social deceptor. Before turning to these examples themselves, I will first briefly discuss what kind of knowledge we should ideally want citizens in democracies to have regarding politically relevant facts. I will discuss this ideal in order to make clear the ways in which fake news and politically motivated reasoning are problematic.

Table 3 *Deceptors distinguished by type.*

Type of deceptor	Description
Object deceptor	A physical object which is a deceptor. For example, Heath offers the case of detergent caps. These caps tend to be much too large (carrying up to six loads of detergent), which exploits the anchoring bias causing people to use too much detergent. They have also been redesigned over the years to become gradually shorter and wider: this exploits a perceptual bias that causes people to underestimate the amount of liquid in short wide containers, as opposed to in tall narrow ones (Heath 2014b, 170–71).
Media deceptor	Content in either a written or audio-visual medium which is a deceptor. Deceptors of this kind include 24 hour news and saturation coverage, which exploit the availability heuristic (see Table 2), and fake news, which exploits other bugs in the way that human memory functions (see section 3.3 below).
Social deceptor	A social dynamic which functions as a deceptor. Deceptors of this kind include the social dynamic which exacerbates politically motivated reasoning (see section 3.4 below).

3.2 Ideal Public Knowledge

In the previous chapter I have argued that epistemic irrationality is problematic for systems of positive-sum cooperation. In this section I will sharpen that analysis, by providing an assessment of what politically relevant factual knowledge we should ideally want citizens in a democracy to know: while I sketch this as an ideal, I will also try to make this ideal as realistic as possible.

For clarity, I shall first define what I mean by ‘ideal’. After all, in a truly ideal world (for instance, one in which citizens have very large amounts of leisure time and education, are motivated to research political issues, and are free of cognitive biases when it comes to processing information) we might want citizens to all spend as much time as they need to be completely informed about all the facts that are relevant for assessing the policy positions of competing parties in elections: how likely these policies are to work, whether they are likely to have certain unintended side effects, whether these side effects will be mild or serious, and so on. However, this is clearly unrealistic: many citizens do not have the time, background knowledge, or expertise to fully assess the likely outcomes of the policy positions of competing parties at elections (see sections 2.2 and 2.3). These problems with citizens political knowledge have been recorded for a

long time, and they appear quite intractable: in the US, despite substantial increases in educational attainment in the latter half of the 20th century, the public's level of politically relevant factual knowledge in 1996 was little different than it had been 50 years prior (Achen and Bartels 2016, 37; Brennan 2016, 30).

When discussing what ideal public knowledge of politically relevant facts would be, I will take into account this widespread ignorance, and the cognitive biases which plague information processing among voters: I will then ask, under these conditions, what kind of knowledge we should ideally want citizens to have about politically relevant factual matters. The kind of ideal sketched in the previous paragraph is, to borrow Heath's terminology, a *first best* ideal theory: it specifies what ideal public knowledge should be, while not taking into account empirical evidence for the limits of human psychology. The ideal I will now sketch can be considered a *second best* ideal theory: specifying the best kind of public knowledge given the limits of human psychology in terms of limited knowledge, available time and attention, and cognitive biases, and also given our current democratic systems and media environment, where the primary route from public opinion to public policy is via elections (Heath 2013, 164–66). I will only sketch this ideal minimally, in order to give an idea of why fake news and politically motivated reasoning are problematic.

To begin with, it is necessary to observe that all societies, but modern societies in particular, are characterised by an epistemic division of labour (A. I. Goldman 1999, 3–4, 103–9; Goldberg 2011, 112). All of us rely on others for much of our information about the world, both because we directly receive testimony from other people which provides us with information, and because norms of truth-telling and testimony policing exist in our society (often these norms are institutionally enforced) which lowers the cost for each individual when it comes to trying to discern what testimony is true and what is false (Goldberg 2011, 113–16, 119–22). Take the institutionalised pursuit of scientific knowledge as an example. As discussed in the previous chapter, the process of peer review is a way of policing the testimony of scientists, which helps to ensure that plainly misleading or false testimony is weeded out before such testimony can reach the public. This reduces the epistemic cost for public consumers of scientific knowledge, because they do not have to exert individual effort in order to verify any given scientific finding: this verification is already carried out at the institutional level. We should exploit this epistemic division of labour, and the reduction in individual epistemic costs that it brings about, by recommending that members of the public attempt to identify relevant experts

and take their opinions into account when confronting policy questions which involve technical scientific reasoning: for example, questions about climate policy, energy policy, or public health initiatives (E. Anderson 2011).¹²

The notion of expert testimony is quite broad: it can be taken to include all sincere pronouncements from experts within their domain of expertise. However, I do not want to argue that members of the public should attempt to take account of all such expert testimony. Firstly, not all such testimony is relevant to policy questions: some expert disputes over, say, the correct interpretation of quantum mechanics, are unrelated to any pressing policy issues. Secondly, even when a domain of expertise is relevant to policy questions, experts often disagree with one another. When they do so they may, on both sides of the discussion, be sincere and have well-evidenced arguments for their own position. In such a situation, it is not usually possible for lay members of the public to decide which position is the correct one: after all, even experts with the correct training and evidence cannot come to a consensus on which position is correct, so it is highly unlikely that untrained members of the public will come to the correct position.

Instead, we should ideally want citizens to have knowledge of positions of scientific consensus that are relevant to policy questions. Consensus implies that the overwhelming majority of experts in the domain endorse a particular belief: while I think it would be foolish to suggest a particular percentage of relevant experts that must be reached in order to constitute consensus, a clear example of consensus would be the position of climate scientists on anthropogenic climate change. According to a large meta-analysis carried out in 2016, between 90 to 100% of publishing climate scientists (depending on the exact question, timing, and sampling methodology) support the theory that humans are causing recent global warming (Cook et al. 2016): I would count this as a paradigmatic instance of a scientific consensus that is relevant for policy questions. Other examples include the fact that vaccinations are safe (notably that they do not cause autism, see Taylor, Swerdfeger, and Eslick 2014), that nuclear power is much safer and cleaner than many available alternatives (Markandya and Wilkinson 2007, 982–983, esp.

¹² There is ongoing debate in social epistemology between two positions on the extent to which laypersons should take expert beliefs into account when forming their own beliefs: the *proper-basing view* (or *total evidence view*), in which the testimony of experts should be weighed as evidence alongside the subject's other evidence on the matter (see Jäger 2016; Dormandy 2018), and the *pre-emption view* in which subjects should update their beliefs to match those of the relevant experts (see Zagzebski 2013; Napolitano 2018). I am more sympathetic to the latter view, but for my argument here I will remain agnostic: according to either theory, lay subjects should seek out the testimony of relevant experts when gathering information, and should update their own beliefs based on these expert beliefs.

fig 3), and that terrorism poses a comparatively tiny threat to American lives (see Table 2). Other facts which fall within this category would be more prosaic: for instance, the rate of unemployment in the run up to an election, the percentage of the national budget spent on key policies, etc. To sum up, one of the main categories of knowledge that we should want democratic citizens to possess is knowledge of expert consensus that are relevant to policy issues.

We should ideally also want democratic citizens to be aware of the policy positions of the various parties that they can vote for in a given election. This may seem like a minimal requirement, but as we have seen it is one that is often violated in reality: indeed, this requirement may already be stretching the second best ideal framework.

The ideal of public knowledge that I have sketched above is minimal, including only two conditions: that democratic citizens aware of (i) expert consensus where this information is relevant to deciding policy questions and (ii) the policy positions of competing parties in elections. Clearly it is not the case that if these conditions are met then the 'correct' policy positions simply emerge. Even if citizens are equipped with the above factual information there is still great potential for disagreement based on value disputes between citizens, or even differential weighting of the same values among different groups of citizens. At most, fulfilling the above two conditions would eliminate certain policy choices. Assuming, for instance, that people value the sustainability of human life on the planet, knowledge of anthropogenic climate change would at least eliminate policies which recommend taking no action to reduce greenhouse gas emissions. Furthermore, even if policy choices are not eliminated, having accurate factual knowledge that is relevant to understanding the effects of available policies will enable a more sophisticated discussion regarding how difficult trade-offs should be made: for example, should we build coal-fired power plants, which generate relatively large amounts of jobs, but which are relatively dangerous and polluting, or should we build nuclear power plants, which might generate a smaller number of more highly-skilled jobs, but which are relatively much safer and cleaner than coal-fired power plants? Such a decision is not *decided* by the accurate information concerning the safety, but with such information the trade-offs involved become clearer, thus enabling a decision to be made which accurately takes into account the costs and benefits of each option. While there may still be positions where trade-offs must be made, having accurate information at least eliminates those positions that are based on an inaccurate understanding of the costs and benefits involved. Eliminating positions based on misinformation opens up

greater potential for positive-sum agreements to be reached between different political groups.

In order to meet conditions (i) and (ii) (outlined above), and given our current media environment it is necessary that democratic citizens are able to obtain accurate news about the policy positions of competing parties and the opinions of large groups of experts, and that they are able to process such information accurately, such that when they encounter expert consensus they respond by incorporating these consensus into their beliefs, rather than ignoring or denying them. Both of these necessities for meeting conditions (i) and (ii) can be threatened however, particularly by the effects of fake or biased news in the former case, and the effects of politically motivated reasoning in the latter.

3.3 Fake News (and other Falsehoods): Problems and Brave Solutions

“A lie gets halfway around the world before
the truth has a chance to get its pants on.”
- Winston Churchill (misattributed)¹³

A number of definitions of fake news have been offered in the literature (see Dentith 2016, 66; Rini 2017; Allcott and Gentzkow 2017, 213; Mukerji 2018, 929, 933–36; Lazer et al. 2018, 1094), but all have the following features in common: fake news is information, presented in the form of a news article (including video and audio news), which purports to be a story about the real world (thus excluding sites like *The Onion* which are openly satirical), but which is known by its creators to be significantly false (thus differentiating fake news from honest reporting errors, ideologically distorted reporting which is believed to be true by its authors, and cases of bullshit in which the purveyor of news simply do not care whether it is true or not)¹⁴, and which is disseminated with the goal of being widely shared and believed by its audience. Being believed helps to facilitate a piece of fake news being shared. In contrast to the definition that I have just offered, Lazer et al. identify fake news at the level of publishers rather than at the level of individual stories (Lazer et al. 2018, 1095; Grinberg et al. 2019, 374), and this is understandable: in a news organisation with a robust editorial policy designed

¹³ See Gillin 2016.

¹⁴ I will come back to these other kinds of false news when discussing whether fake news meets the conditions required to be considered a deceptor, below.

to weed out false information, individual journalists will most likely be unable to get fake news stories (as defined above) published. This empirical connection between the publication of individual fake news stories and poor, complicit, or absent editorial policy on the part of publishers will be significant when discussing brace policies to address fake news. When discussing the creators of fake news I will be referring both to individual writers and to the organisations that they work for, and which publish and/or commission their material. In addition, as most of the recent academic debate on fake news has focused on fake news on the internet, this will be the kind of fake news I discuss.

The authors of fake news have two primary incentives: pecuniary and ideological (Allcott and Gentzkow 2017, 217; Persily 2017, 67–68; Rini 2017, E-44-E-45; Mukerji 2018, 928–29). The pecuniary incentive is the drive to get as many clicks as possible in order to generate advertising revenue: this was the case with the fake news purveyors from the Macedonian town of Veles. Over 100 IP addresses from Veles were linked to the dissemination of fake news: most of the articles, which were overwhelmingly pro-Donald Trump, were copied from fringe American conservative websites, then dressed up to look like reliable news articles and posted to pro-Trump Facebook groups. The motivation for doing this was to earn revenue from adverts placed alongside the fake news articles (Subramanian 2017; Tynan 2016). For those following the pecuniary incentive, the main aim is simply to secure clicks: getting people to believe the fake news is useful only derivatively, in that it makes the articles more likely to be shared. The ideological incentive is the drive to promote and advance the candidates or parties favoured by the fake news creators. This was the case with the owner of the fake news site *endthefed.org*, who said after the 2016 election that he was a Trump supporter, and was proud to have played a role in building support for Trump (Townsend 2016). Of course, even those who are motivated by the ideological incentive also stand to gain substantial amounts of money through advertising revenue.

Whether motivated by the pecuniary incentive, the ideological incentive, or a blend of the two, it is in the interests of fake news creators to create or copy and disseminate content which is likely to be believed and shared. One of the key features of fake news which makes it likely to be believed and shared is that it often panders to people's biases: specifically confirmation bias and in-group mentality. There is evidence that people are more likely to believe and share fake news stories if those stories are congruent with their political outlook: that is, if the story is something which, given their

partisan loyalties, they would like to be true (Schwarz, Newman, and Leach 2016, 88–89; Allcott and Gentzkow 2017, 228–31; Lazer et al. 2018, 1095; Grinberg et al. 2019, 376). People are also more likely to believe information if that information is presented as coming from a source which they have a favourable and warm opinion of (Briony et al. 2017; see also Levy 2017, 25). Such methods of judging the truth of news stories can lead to bad epistemic outcomes: namely to false beliefs.

Given the prior analyses, it is now possible to argue that fake news constitutes a media deceptor. If fake news were a deceptor it would by definition be a media deceptor (see Table 3). To establish whether it is a deceptor at all, let us revisit the basic definition again: a deceptor is a feature (context or object) of our mental environment that has been invented or become ubiquitous because of its ability to exploit human cognitive or social psychological processes, leading to misfires. In this case, the *misfire condition* is met: misfires in this case are the false beliefs which can result from being exposed to fake news. Fake news stories are designed to take advantage of human cognitive processes (such as confirmation bias) and human social psychological processes (such as group identity, in the form of partisan loyalty): they successfully interact with these processes, and thus the *causality condition* is met. Finally, successful fake news stories are created for the purpose of exploiting such processes, either in order to secure revenue for their creators (percupiary incentive) or to secure political support for the creators' favoured party or candidate (ideological incentive): as such, fake news meets the *means condition*. Fake news thus meets the misfire, causality, and means conditions, and so can be classified as a deceptor.

While we are on the subject of media deceptors, what should be made of other types of false news (fake news being a member of the overall category of false news): that is, news that is factually incorrect, but which does not meet the criteria of fake news? In this category are honest reporting errors, ideologically partisan reports in which false information is given, but is not necessarily known to be false by its providers, and factually incorrect bullshit news, in which the purveyors of the false information simply do not care whether the information is true or false. Compared to these other types of false news, fake news is relatively easy to categorise as a deceptor. However, here I will briefly discuss how partisan news can be seen to be a deceptor.¹⁵ Partisan news providers often meet the *misfire condition*: 59% of statements from the right-leaning FOX News

¹⁵ I will deem honest reporting errors to not be deceptors, as the *means condition* is not met and I do not believe the *evolution condition* is met either. As for false bullshit news, I will remain agnostic: the *means condition* may not be met, but I remain unsure about the *evolution condition*.

broadcasts are rated either “mostly false”, “false”, or “pants on fire” (meaning both false and ridiculous) by *Politifact* (Punditfact 2019a), while 42% of statements made on the left-leaning MSNBC had the same ratings (Punditfact 2019b). Consumers are, perhaps unsurprisingly, more likely to perceive stories with partisan leanings which align with their own as informative and interesting (Coe et al. 2008, 215–16): as with fake news, partisan news takes advantage of people’s propensity towards confirmation bias and group loyalty, with robust evidence being found that individuals select their news sources based on partisan affiliation, and discount news that comes from sources that do not share their partisan affiliation (Arceneaux, Johnson, and Murphy 2012, 174, 183–85; S. K. Goldman and Mutz 2011, 55–59). Thus partisan news meets the *causality condition*. Moreover, the fact that people enjoy partisan news creates, as it were, a demand side incentive for its production, creating a “trend toward polarization” among news sources (Coe et al. 2008, 216): partisan news sticks around because it is so effective at gathering viewers (for a stark demonstration of this, see *Statistica’s* recent survey (2019) of cable network viewership in the US, with Fox News and MSNBC at the top of the poll). Partisan news pools in our environment because it is so effective at engaging our groupish partisan loyalties: thus partisan news meets the *evolution condition*. Where partisan news is also false, it arguably fulfils the conditions of a media deceptor.

Turning back to fake news, while it fulfils the criteria of a deceptor one can still ask whether it is a particularly worrisome one. Much has been made of the prevalence of fake news on social media, and the resultant influence on the 2016 US Presidential Election (Silverman 2016; Persily 2017, 68–71; Blake 2018; Kurtzleben 2018; Pennycook and Rand 2018, 1) and the 2016 Brexit Vote and its political aftermath (Cadwalladr 2017; Grice 2017; Toynbee 2019; Wright et al. 2019). However, hard evidence of the influence of fake news on these votes is still being gathered, and the evidence that we have thus far suggests a minimal role is being played by fake news (although this evidence only measures the incidence of fake news and the sharing of fake news, and cannot otherwise measure its impact). Two recent studies by Grinberg et al. (2019) and Bovet & Makse (2019) explored the impact of fake news on Twitter during the 2016 US Presidential Election, and both found that the impact was quite small. Grinberg et al.’s survey of 16422 twitter users (who were also registered voters) found that during the six month period between 1st August – 6th December 2016 the average individuals’ political news exposure consisted of only 1.18% fake news, and that 80% of fake news exposure was experienced by just 1% of their sample. Bovet & Makse surveyed tweets collected in the

5 months preceding the election, and found that 10% of the tweets linking to news media were links to fake news sites (Bovet and Makse 2019, 10). Combined with the fact that, in 2016, only 34% of American web-using adults trusted information obtained from social media “some” or “a lot” (Allcott and Gentzkow 2017, 223), these more recent studies appear to corroborate Allcott & Gentzkow’s claim that the impact of fake news was unlikely to have been large enough to account for Trump’s margin of victory in the 2016 Election (idem, 232).

While we should not overestimate the impact of fake news on democratic elections thus far, there are compelling reasons to view fake news as a deceptor with a lot of harmful potential: and thus motivation to consider brace policies in order to combat it. Firstly and simply, we may consider ourselves lucky that, at present, the density of fake news in our media environment has not reached the point at which it would start being able to meaningfully affect elections.

Secondly, although only 34% of American web-using adults claim to trust news that they read via social media, we should be wary when considering such introspective reports: in particular we should not adopt the view that fake news is only harmful if it leads to a false belief at the point of consumption. Fake news can be problematic even if the person who consumes the fake news does not believe the claims in the fake news report at the point of consumption, due to the rather counterintuitive way in which people form and recall beliefs. In a review of a wealth of empirical evidence, Levy points out the flaws in what he calls the *naïve view*: the view that “mental representations are *reliably* and *enduringly* categorized into kinds: beliefs, desires, fantasies and fictions, and that we automatically or easily reclassify them given sufficient reason to do so” (2017b, 22). He first presents evidence against the view that mental states are reliably and enduringly categorised. Humans continually form representational states of the world around them and of internal states: these representational states include beliefs (or belief-like states), desires, and imaginings, and different states have different causal powers (for example, beliefs are used as premises in reasoning and inference). Levy calls these states *ground-level representations* (ibid). Crucially for the forthcoming discussion, even if ground-level representations are such that the individual would deny belief in them they can still remain in that individual’s memory, where they can influence subsequent cognition (Levy and Mandelbaum 2014, 23–24; Mandelbaum and Quilty-Dunn 2015, 45, 49). When people are called to self-ascribe beliefs, for instance in response to questions, ground-level representations powerfully shape the self-ascription, whether effortfully or

automatically. However, there are instances where the available ground-level representations underdetermine how a person self-ascribes belief. In these situations, other internal or external cues are used to guide self-ascription: for our current purposes, internal cues are of most interest.¹⁶ The key internal cue is *fluency* (also known as *cognitive ease*). When somebody is prompted to decide whether they believe *X*, such as when they are reading an article presenting *X* or when they are asked if they believe *X*, *X* will intuitively seem more plausible if it feels easy to understand or if it is easily recalled: in both cases, this ease of processing is called fluency. Many factors influence how fluent a piece of information feels to process, including how consistent the information is with what people already believe (here we can see some of the processes behind confirmation bias), but importantly for this current discussion information is processed more fluently if people have *encountered that information before* (Kahneman 2011, 60; Schwarz, Newman, and Leach 2016, 86–88). As Schwarz, Newman, and Leach describe in a recent review article on the subject, even if people encounter information and explicitly disavow belief in it at the time (even if they are explicitly told when reading the information that it is false), this information remains in their memory (to use Levy’s terminology, it remains as a ground-level representation) and as little as 30 minutes after learning the explicitly false information people are likely to label at least some of the false messages as true, simply because they feel familiar and this familiarity is used as a proxy for truth. After 3 days people are more likely to rate any claim that they saw previously as true, regardless of whether it was a false claim (Schwarz, Newman, and Leach 2016, 90–91; for an example of an experimental study see Skurnik et al. 2005). This phenomenon is known as the *sleeping effect*, and it is particularly worrying that it happens even in seemingly ideal contexts where the false claims that people are encountering are explicitly labelled as such. This should count as a discounting cue when people remember the claim, but it seems that this discounting cue is easily forgotten or else dissociated from the claim itself (Tarcu Kumkale and Albarracín 2004). The upshot is that even when people know, at the time, that a claim is false, this claim still remains in their memory and means that in subsequent

¹⁶ The influence of external cues on belief self-ascription can be very powerful. For example, when students are asked to write an essay defending the claim that their tuition fees should be raised (a claim which most of them profess to disagree with) they subsequently (i.e. after having written the essay) agree with the claim: the dominant interpretation of this result is that students alter their belief self-ascription in order to explain their own behaviour (writing the essay in favour of tuition fee raising) to themselves (Levy 2017b, 23; for another example see Hall et al. 2013). However, examples of external cues influencing belief ascription tend to involve some sort of manipulation or sleight of hand, and so it is unclear how often they will occur outside of experimental conditions.

encounters the claim is processed more fluently and thus seems more true than it otherwise would had it not been previously encountered.

The increased fluency that results from repeated exposure to a claim doesn't only apply to that particular claim itself: once false information is present in one's memory, it also makes inferentially related information more fluent to process. For example, if an agent succeeds in recalling a claim such as "Hillary Clinton is a criminal" then, even if they are aware that this claim comes from a fake news source, they will also find it easier (more fluent) to process claims such as "Hillary Clinton is concerned only with her own self-interest" because the semantic content of the former claim makes the latter claim seem more familiar, and thus more plausible (Levy 2017b, 31). Thus being exposed to fake news claims not only makes those claims seem more plausible in subsequent encounters, but also makes related claims seem more plausible: in both cases because fluency is increased (Mandelbaum and Quilty-Dunn 2015, 49–50). Given that repeatedly encountering claims, regardless of whether one believes them, lends the claims themselves and related claims fluency, we should not automatically assume that fake news disseminated on social media has no effect on American adults, despite the fact that many of them claim to distrust news on social media. Fake news can affect subsequent cognition even if people explicitly disbelieve the fake news when they encounter it.

A third reason to be concerned about the effects of fake news is the way in which a profusion of fake news in people's media environments can lead to a generalised distrust of all media, including veridical media. Lazer et al. have expressed this concern, observing that fake news "is parasitic on standard news outlets, simultaneously benefiting from and undermining their credibility" (2018, 1094). As recorded in a 2016 Gallup poll, in 2016 surveyed Americans displayed record levels of distrust in the mass media, with one potential explanation being their wariness to place "trust on the work of media institutions that have less rigorous reporting criteria than in the past" (Swift 2016). The fear that a profusion of fake news will lead to reduced trust in mainstream media is largely hypothetical at present, but it does seem to be a plausible contributing factor to distrust against media in general: and such distrust would lead to false beliefs, or to the failure to acquire as beliefs easily accessible true information.

Given the three concerns outlined above, it is reasonable to conclude that fake news has harmful potential. How should we attempt to combat this potential? One natural suggestion is to try to make newsreaders more vigilant, perhaps by encouraging

them to seek out and read fact-checks when confronted with suspicious sounding information. However, this more individual approach is misguided. Firstly it assumes that people will be willing and able to spend the time required to fact-check each and every piece of news media they are exposed to. This is unrealistic: firstly it assumes that people are motivated to spend their time engaging in fact-checking, and secondly it assumes that people are naturally good at spotting suspicious news articles. In a study by Pennycook & Rand they found that a key predictor of how good individuals are at spotting fake news is their performance on the cognitive reflection test (CRT), a three-question measure which indicates one's disposition to engage in analytic and critical thinking (Pennycook and Rand 2018, 8–10): however, performance on the CRT is quite low on average, and it is not unusual for a high proportion of a general population sample to get none of the answers correct (Kahan 2013b, 410). Secondly, seeking out fact-checks can contribute to the harmful effects of fake news: because fact checks often begin with a description of the falsehood, this can further enhance the falsehood's subsequent fluency because it counts as an instance of repetition. Unfortunately, fact-checks can often increase the falsehood's believability (Schwarz, Newman, and Leach 2016, 90–93). Instead of seeking out such individual solutions, we should instead address the problem of fake news at the environmental level using a brace.

Rini offers an interesting example of a brace policy to combat fake news on social media (2017), a suggestion which is in large part an extension of measures which had been taken by Facebook to combat fake news in 2017. Following the 2016 US Election, Facebook set up a fact-checking system in which users can flag news stories as fake in order to trigger a fact-checking process which was carried out by 3rd parties, including Snopes, Factcheck.org, ABC News, AP, and Politifact. If two of these 3rd party organisations considered a story false, then it was flagged in people's news feed as "Disputed by 3rd Party Factcheckers", with a link to the relevant fact-checking article. When users go to share the news story, they are warned prior to sharing that the story is disputed (Persily 2017, 72–75). This system, and the subsequent variants that I will now discuss, has the advantage of not only combatting fake news (as defined at the beginning of this section) but also of combatting other false news, regardless of whether that false news is categorised as a deceptor: thus, if it is successful, it reduces misfires from a broad range of news sources.

Rini acknowledges that the aforementioned measures could be useful when combatting fake news, but she, like Nathaniel Persily, points out that these measures

move too slowly to keep up with fake news: a story could have been read by the majority of its total lifetime audience before a fact-check article can be written in response, and thus the fake news would have done the majority of its damage already (let alone the problems with fact-checking outlined above) (Rini 2017, E-56-5-57; Persily 2017, 73). To get around this problem, Rini suggests that social media platforms like Facebook use the same infrastructure that they currently use to flag individual stories in order to track the testimonial reputation of individual users:

Facebook already knows exactly what each user chooses to share. It will also soon have a database of disputed stories [...]. It would be computationally simple, then, for Facebook to calculate a Reputation Score for individual users, based upon the frequency with which each user chose to share disputed stories. Reputation Scores could be displayed in a subtle way, perhaps with a colored icon beside user photos (2017, E-57).

Rini's proposed system would get around the slowness of fact-checking by creating a more long-lasting record of how trustworthy individual users are as sharers of news. Such a system would offload the task of researching claims in news articles and keeping track of which news articles have already been flagged as fake off of the individual and onto the institutional structure: it is this environmental change that qualifies it as a brace. Rini is correct in pointing out that her proposal does not count as censorship: it does not restrict speech on social media, because it does not prevent anyone from sharing any news story, it simply maintains a record of what people have shared and of whether these shared stories are disputed or not. However, she also points out an unsavoury aspect of her proposal: the ranking of individual users by social media platforms has dystopian undertones, featuring in both dystopian science fiction (such as the *Black Mirror* episode 'Nosedive') and in the policies of repressive governments (such as the social-credit score being considered in local authorities in China, which allegedly ranks people according to their politically related social media behaviour) (idem, E-58). While I do not think Rini's proposal is anywhere near as problematic as the fictional and real examples she proffers (her proposal tracks neither likability nor political leaning, just factual accuracy), I think with a single tweak her proposed brace would be less open to criticism. Rather than tracking the reputation score of individual users, one could instead suggest that social media platforms track the reputation score of individual news sites, putting an icon next to each story published by that site recording the number (or percentage) of disputed stories that have been published by that site. This is arguably fairer than Rini's proposal, because news sites (as opposed to private individuals) can be

thought to have a strong ethical duty qua news sites to exercise care in order to confirm that what they publish is factually accurate: therefore it is more appropriate to hold them directly accountable. This proposal maintains the advantage of Rini's original proposal of creating long-lasting records of trust-worthiness, and of off-loading cognitive demands onto the institutional infrastructure. A potential problem with this new proposal is how to check enough news stories in order to maintain reputation scores for individual sites: potentially what is required is occasional audits on those news sites that produce a certain number of flagged disputed stories in a given period of time, so that a reputation score can be produced. While there are many details to work out in such a proposal,¹⁷ it is nonetheless an instructive example of the kinds of braces that could be used to combat fake news.

3.4 Politically Motivated Reasoning: Problems and Brace Solutions

Above I have addressed the problem of ensuring that the power of fake news to spread false beliefs is curtailed. However, even if people have access to correct information, there is a further problem of ensuring that they take this information seriously and attempt to incorporate it into their beliefs. A serious obstacle to this aim is the phenomenon of *politically motivated reasoning*. When people are presented with evidence which goes against their political commitments (for instance, when conservatives are presented with evidence of anthropogenic climate change, or liberals are presented with evidence of the safety of nuclear power) they often engage in motivated reasoning: counter-arguing, looking for flaws, or simply rejecting the evidence, whereas they readily accept evidence that is congruent with their political commitments (Kahan et al. 2006, 1083–84; Taber and Lodge 2006, 755–57; Kahan 2013b, 407, 2015, 11–14; Kraft, Lodge, and Taber 2015, 121–25; Kahan et al. 2017, 56–58). I will describe the conditions under which politically motivated reasoning tends to occur, and argue that these conditions constitute a deceptor (specifically a social deceptor). I will further argue that this

¹⁷ Indeed, whether or not a reputation score would be effective is of course a matter of empirical investigation. One reason to doubt its effectiveness is its similarity to the aforementioned 'disputed' tag: Facebook chose to remove this feature of its fact-checking program because, unfortunately, labeling news articles with a 'disputed' tag often had the unwanted effect of *increasing* people's belief in the disputed articles (J. Smith 2017). The idiosyncrasies of human psychology must always be fully taken into account when designing braces.

deceptor can only be effectively combatted at the environmental level, with brace policies: and I will offer examples of the kind of brace policies required.

Motivated reasoning is the tendency of individuals to process and assess information in conformity with a goal that is “*collateral* to determining its *truth*”, for instance to maintain positive self-image or to avoid the stress and anxiety of unwelcome news (Kahan 2016a, 2). Motivated reasoning clearly involves a departure from epistemic rationality: if you reason with a goal that is collateral to reaching the truth then you will only get true beliefs in the lucky case that your non-truth-seeking goal aligns with obtaining true beliefs. In the case of *politically motivated reasoning* (PMR), the collateral goal is identity protection: one forms one’s beliefs in order to protect and maintain one’s identity as a member of an identity-defining group whose members have a set of shared values (Kahan 2013b, 417–18, 2015, 11–12, 2016a, 3; Kahan et al. 2017, 56–57; Nyhan and Reifler 2019). PMR can have truly dramatic effects. To offer just one example, an experiment by Kahan, Jenkins-Smith, and Braman (2011) demonstrated that PMR shapes people’s opinion of what experts believe, such that they believe that experts agree with the position that is dominant in their cultural group: this thereby affects who they are willing to consider as experts (*idem*, 166–67). In the experiment, a representative sample of US adults were categorised according to their positions on a hierarchy-egalitarianism axis and an individualist-communitarian axis (for examples of the questions participants were asked in order to categorise them, see *idem*, 151). Participants were then asked to consider that they were advising a friend who is uncertain about the risks associated with climate change, geologic isolation of nuclear waste, or concealed carry laws, and is planning to read a book by an expert in the field to help make up their minds. The participants were shown a brief biography of the expert author (including their qualifications and their membership in internationally recognised expert bodies) and an excerpt from their book: this excerpt was randomly manipulated to show that the expert either viewed the matter in question (e.g. climate change) as low risk or high risk (see *idem*, 155 for examples of the excerpts that participants read). Participants were then asked to indicate on a 6-point scale how much they agreed with the statement that the author was a trustworthy and knowledgeable expert in their field. Whether they agreed that the author was an expert in their field was determined by whether the purported expert’s position cohered with their cultural identity. On the issue of climate change, 78% of those who ranked as egalitarian communitarians perceived that experts agree that climate change is occurring, and 68% perceived that experts agree that climate change

has an anthropogenic source. By comparison, 56% of hierarchical individualists perceived that experts were divided on the issue of whether climate change was occurring, and another 25% perceived that scientists disagreed that climate change was occurring; similarly, 55% perceived that scientists were divided over whether climate change has an anthropogenic source, and 32% perceived that most scientists disagree with climate change having an anthropogenic source (*idem*, 156-157). The results were similar on the issues of nuclear waste management and concealed carry laws, with participants judging the proposed authors to be experts when their testimony coincided with the dominant view among the participant's cultural group, and rejecting their expertise when the expert's view was incongruent with the dominant view. Results of this kind, showing the strong influence of PMR, have emerged in numerous experiments, demonstrating that people tend to estimate risk, perceive scientific consensus, and even misinterpret numerical evidence in order to conform their beliefs to those that dominate in their cultural group (Kahan 2013b; 2015; Kahan et al. 2017). PMR likely has both a cognitive and affective component, with people's prior beliefs being largely determined by the predominant views in their cultural group (cognitive) and their having an affective connection to these priors which motivates them to attend to new information in a way which reinforces them (affective) (Kahan, Jenkins-Smith, and Braman 2011, 168; Kraft, Lodge, and Taber 2015, 125–29). Before moving on to my analysis of PMR as involving a social deceptor, there are two further points that must be made about the phenomenon.

Firstly, people are *more* prone to PMR the better they are at analytic (Type 2) thinking and more general scientific knowledge they possess. When people of above average religiosity are asked to judge the statement “Human beings, as we know them today, developed from earlier species of animals” they unsurprisingly answer this question wrongly, on average, at a greater rate than people who are of below average religiosity. Surprisingly, however, the more highly people of above average religiosity score on a test of ordinary science intelligence (OSI, a test which measures their ability to proficiently reason in scientific ways, see Kahan 2015, 3) the more likely they are to produce the incorrect answer in response to the question. This is in contrast to those of below average religiosity: for this group, higher scores on the OSI correlate with a greater chance of answering correctly (*idem*, 6). When Conservative Republicans (Republican in the sense of supporting the Republican Party) are asked to assess the truth of the statement “There is solid evidence of recent global warming due mostly to human

activity such as burning fossil fuels” they too become more likely to give the incorrect answer the higher they score on the OSI test. For Liberal Democrats (in the sense that they support the Democratic Party), higher OSI scores correlate with a greater chance of answering the question correctly (idem, 12). After taking a cognitive reflection test (CRT), people were then presented with different (experimentally manipulated) facts about the correlation between high CRT scores and belief in climate change, either implying that those who believed in climate change were more open-minded, or that those who are sceptical about climate change are more open-minded: they were then asked to judge whether these facts were accurate. Liberal Democrats with higher CRT scores were more likely to disbelieve the facts that implied that climate change skeptics were open-minded than Liberal Democrats with lower CRT scores, whereas Conservative Republicans with higher CRT scores were more likely to disbelieve the facts that implied that climate change believers were open-minded than Conservative Republicans with lower CRT scores (Kahan 2013b, 412–15). Being more analytic, more intelligent, even specifically being more proficient at scientific reasoning, seems to *increase* people’s ability to reason effectively in motivated ways, not protect them from the effects of PMR. Indeed, even experts (who, after all, have cultural identities as much as everyone else does) are only able to avoid PMR when they are dealing with issues which are squarely within their domain of expertise (Kahan 2016b, 8–10). They seem to be able to achieve this because of their particular “habits of mind, acquired through training and experience, distinctively suited to specialized decision-making” (idem, 8).

Secondly, PMR does not occur with every factual issue that is relevant to politics: as Kahan et al. point out, “[t]he conditions that trigger [politically motivated reasoning] are rare” due to the fact that “[v]ery few facts amenable to empirical investigation ever become symbols of group identity” such that their acceptance or rejection becomes a badge of cultural meaning (2017, 57). This is borne out by many tests of factual knowledge in which both Liberal Democrats and Conservative Republicans answer in the same ways, with the chance of getting the correct answer simply increasing with the individual’s score on the OSI: for instance, giving the correct answer to the question “Climate scientists believe that nuclear power generation contributes to global warming” (the correct answer being that this is false) is predicted almost entirely by one’s OSI score, with minimal effects related to cultural identity (Kahan 2015, 28). What is going on here? Ordinarily, people seek to answer factual questions by looking for credible experts (Kahan 2010, 296), but that who they deem to be a credible expert is highly influenced

by whether a prospective expert shares their cultural commitments, in part because by spending more time with such people they are better at reading such individuals, and ascertaining if they really have the relevant knowledge or if they are “bull shitting” (Kahan 2015, 32). Thus, while the process is insular, and this insularity is partly a result of group loyalty, it also has a rational defence available to it. Every cultural group has a sufficient number of individuals with high-science comprehension (for example, there were both Liberal Democrats and Conservative Republicans who scored highly on the OSI), and so if these individuals are sought after for their opinions by other members of the cultural group who have lower levels of scientific knowledge then the group as a whole, in most circumstances, will likely converge on the correct answers to factual questions over time (ibid). However, this process is distorted, and produces false beliefs, when “positions on risks and other policy-relevant facts become entangled with antagonistic cultural meanings that transform them into badges of membership in, and loyalty to, opposing groups” (ibid). Once this has become the case, then ordinary individuals have a larger stake in believing their group’s dominant position (as believing this position maintains their status in this group, and their connection to others in the group) than they do in going against their group by having accurate beliefs on the issue. Thus, they will use their reasoning abilities in motivated ways to argue against evidence that threatens their group identity. Furthermore, those with the most effective reasoning abilities (such as those with high OSI scores) will use their reasoning abilities in ways that will generate even more polarised positions: others in the group will then look to these individuals and take their positions as a cue to form their own. Thus the most polarised positions (those formed by the most intelligent members of the group) radiate outwards, further polarising the entire group and further entrenching the polarised position as a badge of group identity (ibid).

How then do positions on risks and other policy-relevant facts become entangled with cultural meanings? Sometimes the entanglement is inevitable: if the reality of anthropogenic climate change makes it likely that policies must be adopted to cut carbon emissions and thus cut down on heavy industry, and if individuals who rate more highly on the individualist and hierarchical scales tend to see industry as an admirable activity, then accepting the reality of climate change naturally comes at a cost to their cultural identity (Kahan 2010, 296). However, deliberate actions can create entanglement or exacerbate it. For example, in 2006 the pharmaceutical company Merck had produced a new HPV vaccine, and, in a bid to establish themselves as a market leader, they

sponsored a nationwide lobbying campaign in the US in order to secure approval for their vaccine prior to their rivals GlaxoSmithKlein. The result of this extensive lobbying was that most people heard about the new HPV vaccine from the (often partisan) news media before they heard about it from their physicians or insurers: as a result, the new HPV vaccine became embroiled in controversy, with perceptions of the vaccine's risks being associated with partisan identity (Kahan 2013a).

Now that these two further observations are out of the way, it is possible to inquire as to whether PMR is a deceptor. PMR itself is not a deceptor, however the social dynamic which creates it arguably is: the social dynamic in which positions on facts become entangled with cultural identity. The end result is PMR, which results in false beliefs: thus, the *misfire condition* is met. The entanglement of facts with cultural identity causes this misfire by triggering motivated reasoning in individuals: thus the *causality condition* is met. Once positions on facts have become entangled with cultural identity, the normal process by which information spreads through a cultural group becomes problematic: extremely polarised views spread from the most informed and intelligent through the group, thus connecting the polarised belief even more with group identity. Thus the initial entanglement between factual belief and cultural identity becomes amplified with time due to the way that it interacts with the social process of knowledge dissemination within the group (Kahan 2015, 32). Once the initial entanglement is in place it has the potential to reinforce itself through its own effects: thus, it meets the *evolution condition*.¹⁸ Entanglement therefore counts as a deceptor: specifically as a social deceptor.

What can be done about entanglement, and the resultant PMR that it causes? It seems clear that individual solutions will not be sufficient: once entanglement is in place, trying to further educate people about science or improve their reasoning abilities will likely be either ineffective or else increase polarisation. There are cases in which individual solutions can work, because, as we have seen, experts seem resistant to PMR when they are reasoning about issues in their domain of expertise due to their training: so people can be trained, on an individual level, to avoid PMR. However, the cost in terms of time and resources to give laypeople this level of training renders such a proposal

¹⁸ It is also possible that entanglement could meet the *means condition*. This would be the case when an individual or group with, for instance, a political and economic interest in preventing informed public discussion deliberately attempt to create entanglement between certain facts and certain cultural identities. This possibility seems eminently plausible, however at present I will remain agnostic as to whether any particular instance of entanglement meets the *means condition*.

unrealistic. Instead, we require an environmental solution to prevent the influence of the deceptor: we need a brace. Three examples come to mind.

Firstly, we can follow Kahan's advice in trying to prevent science communication being delivered by partisan news outlets rather than by physicians or other scientific experts, as happened with the launch of the new HPV vaccine in 2006. This could potentially be achieved by a cap on how much money can be poured into the kind of nationwide lobbying which brought the HPV virus into the news media prior to its emergence into the public health system (Kahan 2013a).

Secondly, if entanglement has already occurred, institutions which attempt to communicate science can try to avoid the pitfalls which can exacerbate such entanglement. Kahan points out that a widely disseminated video, produced by the advocacy group Organizing for Action, which was meant to emphasize the overwhelming consensus on the reality of anthropogenic climate change among climate scientists was in fact likely to increase polarisation (and indeed had no noticeable effect on people's acceptance of climate change). It relentlessly associated accepting climate change with Democrats and rejecting it with Republicans, and so likely stoked existing partisan loyalties rather than promoting a sober reflection on the facts (Kahan 2015, 14–19). Such counterproductive science communication should obviously be avoided, in order to create an environment in which facts and cultural identities are less entangled.

Thirdly, even in cases where entanglement and polarisation due to PMR have occurred, one can attempt to frame issues in such a way that members of the public can deliberate about them without becoming overwhelmingly polarised. Kahan gives the example of participants in public deliberation in Southeast Florida. The citizens in the four counties concerned are politically diverse, and so share the same deep polarisation over the issue of climate change as in the rest of the US. However, in 2011 these four counties agreed, after a public and highly participatory process of meetings, to a Regional Climate Action Plan which set out 110 concrete action items to be achieved in the subsequent five years. Kahan's analysis of why this public deliberation worked is that questions were framed in the correct way. Citizens are politically divided on the question "is human activity causing the temperature of the Earth to increase?", but they can agree to the proposition that "local and state officials should be involved in identifying steps that local communities can take to reduce the risk posed by rising sea levels." Thus they can effectively agree to the use of science in government decision-making that will directly affect their lives, even if they do not agree on what the scientific consensus is

(Kahan 2015, 33–36). This does not directly achieve the goal, outlined in section 3.2, of ensuring that the citizens themselves have correct information about scientific consensus: but it does allow scientific consensus to inform policy, which crucially delivers the desired effect of having public policy be scientifically informed in ways that are likely to lead to positive-sum outcomes. These three examples show the potential power of brace policies to combat PMR.

4. Problems: Braces and Bootstrapping

In the previous two chapters I have argued that there are a number of authors in the literature who have promoted the use of braces. I have shown how braces are inspired by work in psychology, much like nudges, but that they go beyond nudges by recommending interventions that are sometimes coercive. I have further argued that braces can be given a coherent and powerful (though defeasible) normative justification in terms of the benefits of positive-sum cooperation. I have argued that there are concrete applications for braces designed to improve the quality of democratic decision-making, and that the brace program provides a key concept, the concept of deceptors, which can be used to produce analyses of problematic elements in the mental environment surround politics and democratic decision-making. In this final chapter I will take a somewhat pessimistic turn, by explicating what I take to be two of the largest problems facing the use of such braces. I will argue that these problems must be addressed when proposing brace policies to improve democratic decision-making.

Before continuing, I should note that Heath and Brennan, my chief examples of authors supporting the brace program, are themselves aware of the dim prospects facing their proposals. In informal contexts, they have both expressed doubts that their proposals could actually be implemented in our current social and political environment. Heath has written that the part of his book offering solutions is notably “half-hearted” (2014d), and that “[s]ubstantively, [*Enlightenment 2.0*] is actually a work of profound pessimism” due to his doubts that workable solutions to public irrationality can actually be achieved (2015b). Brennan, meanwhile, concedes in a televised interview that any possibility of having any of his epistocratic proposals implemented “is a long way off” (Paikin 2016, 23:50) due to the fact that even his most moderate suggestions (such as the voter test in which voters can be remunerated for knowing certain political relevant factual information) are not in effect anywhere in the world. Neither Heath nor Brennan are blindly optimistic, as both admit that serious obstacles stand in the way of their proposals being implemented. My intention in this chapter is to get some clarity on the precise nature of the obstacles which are likely to prevent braces from being successfully implemented.

These problems are chiefly practical problems, rather than ethical ones: I will explain what I mean by this distinction in section 4.1. However, ethical problems will

also arise in the discussion. These two practical problems are, to be more specific, *bootstrapping problems*, by which I mean that they each suggest that the problems which braces are meant to address have the potential to be problematic *for* braces themselves: thus, in trying to implement brace policies, one runs the risk of attempting to pull oneself up by one's bootstraps because the problems one wants to address *with* the brace are also problems *for* the brace. These problems can be briefly sketched as follows:

1. The Implementation Problem – The cognitive misfires that braces are meant to ameliorate can potentially afflict those who are supposed to be implementing the brace (e.g., fact-checkers, science-communicators, civil servants, etc.), thus creating the potential for braces to be undermined at the implementation level by the cognitive misfires that they are supposed to address.

2. The Democratic Problem – In order for any kinds of braces to be implemented in democracies they will often need to secure democratic support: i.e., political parties (or other political entities, such as local governments) which plan to implement braces will need to secure support from those people who will be subject to the brace. However, due to the cognitive misfires that braces are intended to address, it will potentially be difficult to secure such support: e.g., if people already have false and politically motivated beliefs, it may be difficult to secure public agreement on the use of fact-checking organisations or on the use of brace policies to promote knowledge of policy-relevant facts. This difficulty could be circumvented by carrying out brace policies *without* explicit public support, but this then opens up the potential for braces to become democratically illegitimate.

I have chosen to list these problems in ascending order of how much of a threat each of them is to the use of braces: both those suggested by Heath and those suggested by Brennan. I believe the former problem has more potential of being resolved than the latter. In the relevant sections I will further explicate these bootstrapping problems and argue that they must be addressed in each instance that specific brace policies are proposed. The explication of these bootstrapping problems will be more speculative than the discussions in the previous chapters, due to the fact that brace policies are still largely theoretical and thus the problems that I sketch here are those that I believe would likely arise were brace policies to be implemented. As such, there is little empirical evidence that bears on whether these problems would in fact arise, or how serious they would be. Instead, I will try to draw relevant evidence from other debates in order to inform my discussion of the bootstrapping problems. My aim will be to make the case that those who propose braces to improve democratic decision-making, such as Heath and Brennan, must reckon with these bootstrapping problems and that there is an onus on proponents of individual braces to explain how they can circumvent these problems.

4.1 Ethical vs. Practical Problems for Braces

In this chapter I plan to discuss two of the largest practical problems facing braces. Before examining these problems, I will explain how I am differentiating them from the general ethical problems which braces could pose. As I argued in section 2.4, I take braces to be primarily justified because of their potential to protect and scaffold systems of mutually-beneficial cooperation. However, as I mentioned in that section, this justification is of course defeasible. Any particular brace can in principle be objected to, and objected to successfully, on grounds of autonomy, freedom, well-being, or any number of other values: this applies to the braces that Heath and Brennan have proposed, and those which I outlined in the last chapter. Furthermore, it must always be borne in mind that even if a brace would be ethically acceptable and likely to be practically effective, it could always be misused. Braces are often coercive, and, as with all coercive policies, there is always the potential for their misuse: once you have established the administrative capacity to carry out a certain brace, this capacity could always be abused. However, pursuing all of the potential ethical issues with braces, both those that would occur if the implementation of braces went to plan and those which would occur if they were misused, is beyond the scope of this thesis: as such, I will not attempt to pursue them here except where they arise specifically in relation to the practical problems that I will discuss.

How are the problems I will discuss practical problems? I deem them to be practical problems because they concern whether braces could actually be carried out in existing institutional frameworks in democracies. Both of the problems concern ways in which braces could fail to be effective, either because they could not be competently carried out by those charged with implementing them (the Implementation Problem) or because they could not receive the popular support necessary to be implemented in modern democracies (the Democratic Problem). While I am labelling them as practical problems, both of them clearly have ethical dimensions: in the first case, braces being carried out incompetently is an ethical problem because it could be harmful; in the second case, I will argue that the practical problem of securing popular support for braces leaves one with an ethical quandary about the extent to which one is willing to circumvent democratic procedures in order to secure welfare gains. However, these ethical problems are secondary effects of practical problems: one would not have to be concerned about the harmful effects of incompetent braces if one thought that they could be carried out competently, and, similarly, if one believed that cognitive biases

would not impede public support for braces then one would not have to face the ethical issue of balancing the value of democratic procedures against welfare gains.

A further reason for addressing these practical problems is that I believe they would arise regardless of how one intended to justify braces. I have argued that braces are chiefly justified because they can support systems of mutually-beneficial cooperation: I take this to be a strong justification, because it is a relatively thin ethical concept, and I have argued that it is the justification implicit in Heath and Brennan's proposals. However, one could argue in favour of braces on other grounds: perhaps on the grounds that braces enhance autonomy by scaffolding our rational (Type 2) cognition. Regardless of how one justifies the use of braces, one should be worried about practical problems which threaten their implementation: and, indeed, if it transpires that they cannot be implemented then ethical issues which would apply were they implemented fall by the wayside. With the distinction between ethical and practical problems, and importance of the latter, established, I will now turn to the first bootstrapping problem.

4.2 The Implementation Problem

Whenever one intends to use a brace to address problems in the mental environment, this brace will need to be implemented by a particular group of people. In the case of Brennan's suggested brace of offering a tax rebate for citizens who can pass a test of factual knowledge, the implementing group would likely be civil servants. Rini's proposal to combat fake news would be implemented by people who work for fact-checking agencies. Kahan's proposals to ameliorate the effects of PMR would be carried out by science communicators, or, in the case of the public deliberation in South Florida, was carried out by civil servants and local politicians.

The Implementation Problem can be stated as follows: in each of these cases, we must ensure that the group implementing the brace intending to combat cognitive biases is not susceptible to the same or similar biases itself. If it were subject to such biases, then one would have strong reason to doubt the effectiveness of the brace: for example, if the fact-checkers who are charged with determining if news articles contain falsehoods are themselves extremely biased reasoners then we would have strong reasons to doubt that their activities would effectively combat fake news: if anything, through the power they wield to mark news articles as false, they could make the situation worse. Indeed, if the problem is as diagnosed, they *will* make the situation worse.

The threat that the implementation problem poses to braces can be made particularly clear via a comparison with a structurally similar problem raised by Christopher Freiman: the problem of government failure as applied to Rawls' political theory. Freiman points out that there is an important asymmetry between the state and the market in Rawls' theory. Rawls assumes that people will be less than fully compliant with justice in civil society and the market: this deviation from full compliance creates the conditions in which the state must intervene to resolve market failures and ensure that actors behave justly. However, Rawls assumes that the actors within the state who will be carrying out these interventions *are* fully compliant with justice: indeed, this assumption enables him to have confidence that state interventions will not make things worse due to corruption or other unjust behaviour on the part of state officials (Freiman 2017, 25–28; 2018, 301–3). Freiman points out that Rawls is endorsing a kind of behavioural asymmetry: he is applying one model of behaviour, a model in which people deviate from the demands of justice, to market actors, and other model, in which they are compliant with justice, in the institutional context of the state. Freiman argues that this is unjustified, as it violates a requirement of institutional analysis which he, following Geoffrey Brennan and James Buchanan, calls *behavioural symmetry*: that one should apply one's behavioural model of actors across different institutional contexts *unless* one has a good reason to believe that people will behave differently in different institutional contexts (Freiman 2017, 26–27).

When thinking about braces, I hold that one should apply a requirement of *cognitive symmetry*: assuming that one's model of people's cognition applies across different institutional contexts *unless* one has a good reason to believe that their cognition will be different in different institutional contexts. Unless one has good reason to believe that the people's cognition as members of the general public (which, as discussed a length above, is subject to numerous biases) is different in the institutional contexts of (for example) a civil service or a fact-checking agency, then one should be concerned that those charged with implementing braces will be subject to the same cognitive limitations as those for whom the braces are being implemented. If this is the case then one has reason to doubt that the brace will be correctly designed and implemented, in which case one has reason to doubt that the brace will be successful in scaffolding rationality and promoting positive-sum cooperation.

One can now see how the Implementation Problem is a threat to the use of braces: even if one could secure political support for braces (on which see section 4.3),

one would have serious concerns about whether braces could be carried out as intended. In order to overcome the Implementation Problem, one must be able to satisfy the latter condition of the requirement of cognitive symmetry: one must be able to give a positive reason why people's cognition would be different in the institutional context in which a brace is being designed and implemented (e.g., a government department or a fact-checking agency) than in civil society. If this condition can be met, then we can justify a form of cognitive asymmetry which allows us to defuse the Implementation Problem. So, can this condition be met? I think that working out the fine details of whether this condition can be met is something that must be done on a case-by-case basis: however, I want to offer two general resources one could draw upon in order to argue that a brace can avoid the Implementation Problem.

Firstly, one can argue that the individuals who will be charged with implementing the brace have particular cognitive abilities which insulate them from cognitive biases: this would justify cognitive asymmetry. As we saw in section 3.4, there are some cases in which people's expertise insulates them against certain cognitive biases. People who are experts in a given domain do not engage in PMR when considering questions in that domain (Kahan 2016b, 8–10). More generally, there are “knowledge bases, rules, procedures and strategies” (Stanovich 2018, 429) which people can learn, and which enable them to avoid certain kinds of cognitive biases. These are collectively known as mindware: key examples of mindware include knowledge of probabilistic reasoning, causal reasoning, and numeracy (*ibid*). Simplifying the issue somewhat, we can say that mindware enables people to avoid cognitive biases in two key ways. Sometimes mindware is available during cognitive override: one has some Type 1 intuition, which one then effortfully overrides and replaces with a response derived from explicitly represented mindware which is only accessible during effortful cognitive decoupling. In other cases, mindware can be successfully overlearned and can migrate into Type 1 processing: in this case, people's intuitive judgements are informed by their mindware (*idem*, 430-432, see especially Figure 1). If one can successfully demonstrate that the people who are charged with implementing a brace possess the relevant expertise or mindware to avoid potential cognitive biases which could undermine the brace's implementation then one can potentially defuse the Implementation Problem.

This first response has mixed prospects. On the one hand, mindware is measurable, and so it is possible to discover whether a prospective individual or group of individuals possess the requisite mindware to carry out a task in a non-biased way: for

examples of mindware measurement see the experimental designs in Stanovich and West 2008. This measurability means that arguments for cognitive asymmetry can be robustly evidenced. However, a criticism of this response is that even when the appropriate mindware is present to carry out a task in a non-biased way, this is no guarantee that the task will in fact be carried out in a non-biased way. This is due to the fact that in cases where mindware must be effortfully accessed during cognitive decoupling there is always the possibility cognitive biases persisting: either due to detection failure, where the biased Type 1 response is not detected as being in error and so decoupling is never initiated, or due to override failure, where the error is detected and cognitive decoupling is initiated, but this decoupling fails and so the Type 1 response is not overridden by a response generated by the relevant mindware (Stanovich 2018, 432–39, see especially Figure 4). In cases where mindware has migrated to Type 1 processing, these error possibilities are avoided: in these cases, the intuitive response is directly generated by the overlearned mindware, and so is likely to be non-biased (*idem*, 435, Figure 3). At least in cases where one has reason to believe that mindware is not overlearned, one should be concerned about the rate at which detection and override errors are likely to occur: to the extent that they are likely to occur, the first response against the Implementation Problem is weakened. Happily, proponents of braces have a second line of argument available to them.

The second potential way of defusing the Implementation Problem focuses not on the individual cognition of the people tasked with implementing the brace, but rather on the mental environment in which they are implementing it. As Heath points out, there is something confusing about the fact that private enterprises often take advantage of the cognitive biases of consumers (recall the example of the changing shape of commercial detergent caps in table 3), but this rarely happens in reverse (Heath 2014b, 302–3). This is surprising if one considers the presumption of cognitive symmetry: why are corporations seemingly more rational than consumers? Heath’s explanation is that decisions made in private enterprises are often made in highly scaffolded mental environments with many helpful kluges to avoid cognitive biases and promote rationality: decisions are made in groups, proposals are discussed in meetings, the relevant factors have all been made explicit as part of a cost-benefit calculation, and so on. Decisions within corporations are “made in a highly structured social environment, with all the assumptions and inferences made explicit and with multiple points of contestation and correction” (*idem*, 304). On the other hand, consumers often face purchasing decisions

using only their onboard cognitive resources: as such, cognitive biases can play a much greater role in their decisions than in the collective decisions of corporations (ibid; Clark 1997, 271–73). The reason why the requirement of cognitive symmetry gets broken in this case “is not because the *people* are more rational, it is because they are operating in an institutional environment that is more conducive to rational thought and planning” (Heath 2014b, 304; for an extensive discussion of the role of the environment in scaffolding rationality see Clark 1997). The suggestion that arises from this discussion is that cognitive asymmetry can be justified if one can show that the environment in the institution that will be carrying out the brace is one which is likely to promote rationality: essentially, you can successfully design and implement a brace if you inhabit an environment where one is already present.

This may sound problematic: if implementing braces requires braces, then how are these latter braces to be implemented? Wouldn't the Implementation Problem simply reassert itself at this stage? However, I think that this suggested response can be defended. Firstly, many useful braces that are already present within institutions (e.g. procedures whereby people have to justify their ideas to others in an adversarial context) are not the product of individuals attempting to design good mental environments, but are rather the products of quasi-evolutionary processes: as Clark points out, many firms have developed institutional practices that promote rational decision-making due to the pressures of a competitive marketplace, in which firms which do not develop such strategies are eliminated by firms which do (Clark 1997, 272). Under such conditions of competition, people will naturally try to find ways to keep their firms afloat: and, once institutional practices are found which promote rationality and thus enhance a firm's competitiveness, these practices are likely to spread simply because any firm which does not adopt them will be outcompeted (see Brooks et al. 2018, 3 for a description of the mechanism by which such practices can be transmitted). Many useful institutional practices are extremely old: for instance, Heath traces back the kluge of using adversarial reasoning back to Aristotle, and notes that it has been used heavily in the European intellectual tradition, including by the Catholic Church in the 16th century in their practice of appointing a “devil's advocate” to present arguments against potential candidates for canonization (2014b, 142–44). Given the longevity of this kluge, one can see how it would naturally occur in many different institutions: its long history of success ensures that it is continually reproduced when people are looking for ways to make more rational decisions within institutions. Secondly, once these institutional practices have evolved,

they form an environment in which people who are attempting to design further positive environments can work in: this opens the potential for a virtuous cycle in which being in an already braced environment may make one better able to further brace one's environment. However, while these responses have merit, we should be sanguine about them: the cultural evolutionary process which has provided our current repertoire of environmental kluges is slow, and so we should expect the rate of discovery of new kluges, if indeed there are a large number still to be found, to be similarly slow. However, where we already have institutional kluges which function as braces, they can be used to answer the Implementation Problem.

Many institutions use the adversarial reasoning kluge, ensuring that decisions are checked by other individuals before they can be implemented: private corporations have such procedures, as do state bureaucracies (Heath 2014e).¹⁹ For further examples, think back to Rini's suggested brace, which utilises the efforts of fact-checking organisations. Many of these organisations have editorial policies which ensure that when individuals write fact-checks these are subsequently subject to editorial review by at least one other person. For example, *Snopes* often have multiple members of editorial staff work on a single fact-check, and all fact-checks pass through at least one editor before they can be published (Snopes 2019); all fact-checking articles on *Full Fact* are reviewed by at least one additional researcher before they can be published, and often go through the Director if the story is politically sensitive (Full Fact 2019); all reporters for *PolitiFact* have to take their fact-check to an assigning editor for discussion, and subsequently the fact-check is taken to two additional editors for review, after which the three editors collectively vote on whether the piece should be published as is (Holan 2018); and fact-checking stories published on *FactCheck.org* go through as many as four editors before the piece can be published (FactCheck.org 2019). This is a level of scaffolding that is simply not applied to most people's cognition in everyday life: as a result, we should expect the outcome of people's cognition to be different within a fact-checking organisation than it

¹⁹ In addition to the adversarial reasoning kluge, there is the simple fact that within private corporations, fact-checking organisations, or government bureaucracies, employees are remunerated for their work, which incentivises them to both exercise effort to carry out their tasks competently, and also to participate in the other rationality-scaffolding practices of the institution, such as group decision-making procedures. Furthermore, there is a stick behind this carrot: if people fail to carry out their tasks competently, they can be punished or fired, further incentivising people to attempt to carry out their tasks competently. This is in contrast to the position of members of the general public in their role as democratic citizens: there is no remuneration to incentivise people to become more informed or to exercise greater effort to override their cognitive biases (to the extent that this is possible in the first place).

would be if carried out in environments without the same scaffolding (such as the environment which most individuals inhabit most of the time).

This second line of argumentation against the Implementation Problem does not strictly establish cognitive asymmetry in the same way as the first line of argumentation sets out to: people in heavily scaffolded mental environments are not necessarily less likely to suffer from cognitive biases on an individual level than people in the general public (in contrast to cases where people possess appropriate mindware), instead their environment already contains a brace, so that their cognitive biases will likely be corrected for by other individuals before these biases can influence any outcomes: this serves as a resource to refute the Implementation Problem.

With those two lines of argument combined, proponents of braces are able to mount quite a strong case against the Implementation Problem. One useful feature of these two lines of argument is that they are mutually supporting. If one can show that the people implementing a given brace intended to improve democratic decision-making *both* possess certain relevant kinds of mindware and will be working in an environment which is conducive to rational decision-making then one can attempt to address some of the worries about detection and override failure that are problematic for the first line of argument: one can admit that people may suffer from detection or override failure, but then argue that there is still a strong chance of cognitive biases being caught because even if people individually fall prey to detection or override failures their cognitive biases may still be recognised by their peers. In the reverse direction, one can argue that systems of adversarial reasoning will likely be made more effective if the individuals involved all possess appropriate mindware, as this will make it more likely that cognitive biases will be recognised and addressed.

To conclude, I believe that the Implementation Problem can be successfully argued against provided that the right conditions are present in the institution which is carrying out the brace: these conditions include whether the relevant mindware is present and whether there are institutional practices present which can scaffold the rationality of the decision-making within the institution. There are resources for proponents of braces to use to argue against the Implementation Problem, but this argument must be made in every case where a brace is proposed. Furthermore, this line of response is limited by the finite number of environmental kluges that we currently have at our disposal, and the slow rate at which we can expect new kluges to arise via the cultural evolutionary process.

4.3 The Democratic Problem

I have thus far argued that braces can be justified, that concrete examples of braces can be identified, and that arguments can be made that certain groups will sometimes be able to competently carry out the implementation of braces. Now I will turn to the second bootstrapping problem. Where braces are designed to improve democratic decision-making, there is a possibility that the environment which braces would be introduced into, (an environment in which people are subject to cognitive biases, in which many people are uninformed or misinformed about political matters, and in which there are many existing deceptors which exploit these cognitive biases and worsen problems of misinformation), the environment which braces are designed to ameliorate, will lead to widespread resistance to brace policies, making it unlikely that such policies will be able to receive democratic support. This is the Democratic Problem.

Here, as when I introduced the Implementation Problem, I think it is useful to get a grip on the Democratic Problem by first discussing a very similar problem diagnosed in the literature: one of Robert Talisse's criticisms of Brennan's epistocratic proposals from *Against Democracy*, which we can dub the 'problem of Democratic Personae', to paraphrase the subtitle which Talisse uses in his paper (2018, 7–9). As discussed in sections 2.2 and 3.2, the average voter has low levels of politically relevant factual information; but when discussing the entire voting population, Brennan divides voters into three archetypes: Hobbits (those who are largely apathetic about politics and lack fixed views and beliefs on political matters), Hooligans (those whose personal identity is bound up with their political leaning, who have strong (but often inaccurate) views on politics, and who often process information in biased ways), and Vulcans (those who process political information rationally, are well versed in the relevant social science, and who are only as confident in their political beliefs and opinions as the evidence allows). Brennan argues that most voters are either Hobbits or Hooligans, and that practically nobody is a Vulcan (*idem*, 4-5). Given that Vulcans only exist in small numbers, Hobbits and Hooligans make up the majority of democratic citizens.

Talisse points out that whereas Hobbits or Vulcans may first-personally understand themselves as being Hobbits or Vulcans according to Brennan's descriptions, Brennan's description of Hooligans "is strictly second- or third-personal" (Talisse 2018, 7). No Hooligan considers himself or herself to be a Hooligan, because if their sincere reflective understanding of themselves were of being a Hooligan (i.e., of being misinformed and biased) then they would "lose confidence in [their] reasoning and its

products” (idem, 8). If anything, Hooligans likely consider themselves to be Vulcans, while considering their political adversaries to be Hooligans. What is the upshot of this for Brennan’s epistocratic proposals? Talisse argues that Hooligans may welcome a more epistocratic system of government in the abstract, because they (in most cases mistakenly) think that they already have good epistemic credentials, and so epistocratic institutions would support ‘their side’ politically. However, once a decision-making body has been formed which has an “epistemologically responsible selection” of members, Hooligans are likely to start disagreeing with the epistocratic body: they will not accept decisions based on evidence which they (wrongly) believe is mistaken (idem, 9). Due to the fact that Hooligans process factual claims in biased ways, it would be very unlikely that any particular group of Hooligans has accurate beliefs about all politically-relevant facts: as such, over time, the epistocratic body would draw the ire of politically diverse groups of Hooligans by recommending decisions based on evidence that the Hooligans do not accept. Talisse points out that such a situation is very unstable: because Hooligans are Hooligans (“the rabid sports fans of politics” (Brennan 2016, 5)) “their opposition [to the epistocratic body] will take unwelcome forms that will destabilize the social order” (Talisse 2018, 9). In the extreme, epistocratic proposals may be “met with insurgence” (idem, 10), because the biases, misinformation, and ignorance that they are designed to combat will prevent large swathes of the general public from accepting the pronouncements of an epistocratic body.

Two things are of note about the problem of Democratic Personae. Firstly, it is a bootstrapping problem for epistocratic proposals: Talisse is suggesting that Brennan’s proposals will be undermined due to resistance caused by the very problem (political hooliganism) that his proposals are designed to deal with. Secondly, Talisse’s argument is not based on any hard evidence of what would happen if epistocratic proposals were introduced: there is no empirical evidence that citizens would reject Brennan’s epistocratic proposals, or that they would engage in “insurgence” against epistocratic bodies if these bodies existed. This is not intended to be a damning criticism of Talisse: he does not provide such evidence because it does not exist, as epistocratic bodies of the kind Brennan promotes have not been tried. Instead, Talisse is engaging in informed speculation: he is essentially taking Brennan’s own description of political Hooligans as biased and motivated reasoners, and as misinformed on many issues, and extrapolating what their likely response would be to an epistocratic body which issues decisions based

on evidence which, though correct, they do not accept. This seems hard to reject for Brennan, even in the absence of positive evidence.

Using the problem of Democratic Personae as a starting point, we can, with a little tinkering, describe the Democratic Problem for braces. This latter problem shares the above pair of characteristics with the problem of Democratic Personae. Firstly, like the problem of Democratic Personae (and the Implementation Problem), the Democratic Problem is a bootstrapping problem: I will be arguing that there is a possibility that the problems in the mental environment that braces are meant to address will prevent braces from gaining political support in existing democracies. Secondly, the Democratic Problem is an exercise in informed speculation: because there is little evidence of braces being trialled, there is also minimal evidence of whether they would be able to secure political support. I will use the previously discussed descriptions of the problems that braces are meant to combat to paint a picture of the kind of environment in which braces would have to be proposed, and then argue that there is reason to suspect that this environment would be hostile to proposals for some kinds of braces: they would face political backlash which would make it unlikely that they would be implemented in current democracies. Luckily, we have at least one example of a brace which has *already* been implemented: and I will attempt to assess which qualities of this braces made its implementation possible. Ultimately I will suggest that braces are more likely to be successfully implemented to extent that they can either circumvent the need for public support and/or to the extent that they are implemented on a small scale. Naturally, the first of these features raises an ethical issue about the extent to which implementing braces in a way that avoids the need for public support violates some conception of the need for policies to be democratically legitimate: I will raise this issue, and make clear the trade-offs involved, but ultimately this is simply one more question that must be assessed when considering the use of braces.

Let us begin with a brief re-description of the kind of environment that braces will be proposed in: the environment surrounding politics in contemporary democracies. Firstly, this is an environment in which the general public are largely uninformed or misinformed when it comes to politically relevant facts, due to the fact that they lack sufficient incentives to become informed, whereas they often possess incentives to hold beliefs which conform with those of their cultural group. Secondly, it is an environment in which everyone is subject to a range of cognitive biases which can distort their perception of the facts (as with the availability heuristic or the framing effect) and

hamper their ability to reason accurately (as with myside bias, confirmation bias, and in-group bias) (see Chapter 1, particularly section 1.2). Thirdly, it is an environment which is, increasingly, a cognitive sodium vapour lamp: where increasing demands to use Type 2 cognition are made on scarce cognitive resources, with the potential to overtax these resources (see section 1.2). Fourthly, it is an environment which is populated with numerous kinds of deceptors: most relevantly fake and biased news, misleading political advertising, and entanglement between factual matters and cultural identities, leading to politically motivated reasoning (see Chapter 3).

What are likely to be the consequences of such an environment for the prospects of introducing braces? Firstly, we should not expect demand for braces to come from the public: simply put, it is not likely that sufficiently large numbers of the public have a sufficient incentive to spend their time campaigning for the introduction of braces. This is due to the fact that, as has been discussed previously, individual actions on the part of democratic citizens tend to have no effect on collective outcomes: one could be informed about all the problems outlined in this thesis (although this in itself is unlikely, given levels of public ignorance), but one would still likely lack the motivation to dedicate any significant amount of one's time to calling for brace policies, because this would likely be futile. Moreover, if voters were able engage in informed collective action, across ideological lines, to implement braces then many of the problems that braces are meant to solve (such as the public lack of knowledge about politics, or the general irrationality and bias in public political discourse) could be taken to be largely solved already. The public's ability to collectively act in this way would be proof that the problems that braces are meant to address had been largely overcome (Heath 2015a). Thus, if braces are implemented at all, we will likely have to rely on political actors, such as political parties, to attempt to introduce them: but, without public demand, we should also expect this to be quite unlikely.

Even if a political party did attempt to introduce a brace, the existing environment would likely pose further problems. One basic problem is that the rationale for why we need braces (because we are all subject to wide-ranging cognitive biases) may be a difficult one to convince people of: as mentioned previously (in section 2.1), people have a bias blind spot, where they mistakenly think that if they were biased they would know about it, whereas in fact "most social and cognitive biases operate unconsciously" (Stanovich 2011, 112). The evidence that the brace program is grounded upon is quite unintuitive, which is a handicap in an environment in which unintuitive policies are a

hard sell. A more specific problem, for epistemic braces, is that they are generally premised upon privileging certain information because it is true. For instance, Heath's proposed brace to impose "prohibition on outright falsehood" in political advertising (idem 2014b, 345) would require some agency to decide whether statements are or are not false. This agency would be something akin to existing fact-checking agencies, as were discussed in section 3.3. Whether something is an outright falsehood is not, in principle, difficult to establish: independent fact-checkers routinely establish whether statements made by politicians or news outlets are true or false, using publicly available sources as evidence. However, whether something is *accepted* by the public or by competing political parties as being true or false can be a different matter: where facts are already politically controversial (for instance, in cases where there is a lot of politically motivated reasoning involved), there is the potential for the very epistemic irrationality that the brace was intended to solve to undermine support for it. To the extent that existing problematic elements of the environment distort any public debate about whether to implement particular brace policies, it is a distinct possibility that any political party that was intending to implement them may drop the policy in response to political controversy.

One potential way around the Democratic Problem is simply to insulate the decision to implement braces from the need for public support as much as possible, so that biases in our political environment are largely kept out of the decision-making process. As was pointed out in Talisse's example, one cannot completely insulate policies from public support: at the limits, one needs a sufficient level of public support to avoid a dangerous public backlash against a policy. However, within this threshold, there are ways to insulate policies from public oversight. For instance, one can create separate administrative agencies to deal with certain types of public policy in areas where the need to secure public support would have deleterious effects on policy: existing examples of such arrangements include the use of central banks in modern democracies, which are often staffed with unelected officials and which operate with large amounts of independence from the government of the day, due to the fact that they often have to make unpopular but necessary decisions, such as raising interest rates to counter inflation (Heath 2014b, 337–38).²⁰ One could theoretically create institutions like this to design and implement braces. However, while insulating such decisions from public control

²⁰ As Heath points out, "[t]here is no justification for this arrangement other than the recognition that if the public did have control over the central bank [...] they would make terrible decisions" (2014b, 338).

would likely avoid the Democratic Problem to some extent, it is obviously ethically problematic: by designing and implementing braces in a way which partially insulates them from the need for public support, decisions are being made which are intended to affect democratic citizens without input from democratic citizens. Such decisions cannot claim democratic legitimacy, at least according to two of the major theories of democracy.²¹ On an *aggregative* view of democracy, in which the purpose of democracy is to aggregate the preferences of voters, such decisions cannot claim legitimacy because the preferences of democratic citizens are not taken into account, and on a *deliberative* view of democracy, in which the purpose of democracy is to allow citizens to collectively deliberate about their common good, such decisions cannot claim legitimacy because democratic citizens are not party to the discussions which lead to the decisions (Shapiro 2003, chap. 1; Heath, n.d.). I do not think this ethical problem can be elided: instead we are faced with a trade-off.²² Braces are intended to contribute towards the maintenance of systems of positive-sum cooperation by removing misinformation from the public debate and allowing collective agreement on pressing issues facing communities.²³ Perhaps these benefits are worth the democratic deficit required to secure them. A further reason to think that this trade-off is worthwhile is that, although these braces initially involve a lack of democratic legitimacy, their effects are arguably desirable from the perspective of aggregative and deliberative democrats: from an aggregative perspective, these braces made citizen's preferences more informed (or at least less misinformed), and from a deliberative perspective these braces have the potential to improve public deliberation, again by making it more informed. Of course, if one places an absolute value on democratic legitimacy then these arguments will fall on deaf ears.

²¹ I have omitted here to mention the third major theory, the *competitive* theory, mainly because its focus is more on the ability of democracies to produce “strong, capable leadership”, and then replace these leaders periodically. Therefore, it does not have as strong a stance on the way in which particular democratic decisions are made (Heath, n.d.).

²² The ethical dimension of this problem would likely be more troubling to Heath, who acknowledges a need for “public control of decision making” (2014b, 338), than to Brennan, who has a purely instrumental view on the value of democracy in which it is useful to the extent that it produces independently good outcomes (2016, 10–14). However, as is made clear by Talisse, there is a purely practical limit to the extent that liberal governments can make policy independently of the public's wishes without risking a dangerous public backlash: and Brennan would likely be concerned about this practical worry, given that, as a classical liberal, he has no desire for an illiberal government

²³ On the latter point, a good example is the fact that, as a result of the Regional Climate Action Plan in Southeast Florida (discussed in section 3.4 and below), roads and seawalls in Miami are being raised according to new minimums enforced by County Governments, in order to mitigate the effects of flooding (Ruggeri 2017). It is hard to think of a better example of public goods which provide positive-sum benefits.

However, I believe that if one is open to trade-offs between democratic legitimacy and social welfare, or trade-offs between democratic legitimacy in the short-term and improved democracy in the long-term, then arguments can be made, on a case by case basis, in favour of insulating the design and implementation of certain braces from the need for public support, to a greater or lesser extent. Of course, if there is another way to avoid the Democratic Problem, one which is similarly effective but which raises fewer ethical issues, then we should on balance prefer it.

I will now turn to more optimistic ground by arguing that there is another way. Although evidence of braces being implemented is very limited, we do have evidence from at least one successful case: the public deliberation sessions carried out in Southeast Florida, which enabled the politically polarised population to collectively agree to the Regional Climate Action Plan. I will now briefly re-describe this brace, and investigate which features made its implementation successful, in order to produce another suggestion for how to enable braces to avoid the Democratic Problem.

The meetings in Southeast Florida were attended by citizens from across the political spectrum. Like most US citizens, they were sharply divided on the issue of whether anthropogenic climate change is a real phenomenon: and, as is the case with politically motivated reasoning, citizens were more polarised the more scientifically literate they were. However, by framing the questions that citizens had to answer away from issues of whether anthropogenic climate change is real (a question which is entangled with their cultural identities) and towards issues of whether “local and state officials should be involved in identifying steps that local communities can take to reduce the risk posed by rising sea levels”, on which politically diverse citizens can agree, the meetings were able to avoid the negative effects of politically motivated reasoning, and instead generate a consensus sufficient to lead to the approval of a Regional Climate Action Plan (Kahan 2015, 33–36). The design and implementation of this brace can therefore be taken to be quite successful: what enabled this brace to be carried out so successfully? The public meetings were carried out after four local counties agreed to form the Southeast Florida Regional Climate Change Compact in between late 2009 and early 2010 (see Broward County Government et al. 2010 for details of their agreement). The agreement was ratified by the boards of County Commissioners in each county: these are small groups of local elected officials, with between 5 and 13 members depending on the county (Broward County Government 2018; Monroe County Government 2019; Palm Beach County 2018; Miami-Dade County Government 2018).

Therefore, one likely reason why this brace was implemented so smoothly is that only a very small number of individuals had to agree to form the Compact, meaning that all that was required was for the small group of elected Commissioners to agree, and for no significantly large group of citizens to attend the meetings of the Commission and argue against the Compact. I would not blame the reader for thinking that these fine details about boards of County Commissioners somewhere in Florida are overwhelmingly boring; however, I think that this very lack of newsworthiness is another factor which enabled this brace to be successfully implemented. The extremely local nature of the policy meant that it was not reported on by any major news organisations at the time:²⁴ thereby completely removing any potential for media deceptors to stoke any opposition to the policy, simply because it was not significant enough to be of any interest. Therefore, carrying out braces at the local level seems to have the advantage of avoiding many of the factors in our current political environment which can lead to the Democratic Problem.

This seems to be the key feature of this example which enabled it to avoid the Democratic Problem: by being carried out on a local level, its implementation was not strongly exposed to harmful elements in our current political environment (notably media deceptors). Thankfully, this feature is far less ethically problematic than the previously discussed feature. On purely pragmatic grounds, we should favour braces which are implemented at the local level because they are less likely to have their implementation undermined by harmful elements of our existing political environment. In a sense, this second feature also lessens the need for public support, because one only needs the support of a smaller section of the public to achieve results on a local scale, but it is nevertheless that case that the example from Southeast Florida was carried out by elected officials and the public meetings in which the Compact was agreed to were open to the public: so there is still a large degree of public support. One downside of this second feature is that implementation at the local level naturally limits the impact that braces can have: nevertheless there is some room for optimism. Once braces get

²⁴ One can verify this by searching for news articles containing “Southeast Florida Climate Compact” on Google, and restricting one’s date range for articles. In the two years after the Compact was created (from the 10th January 2010 to the 10th January 2012) there is only one mention of the Compact, on a clean energy blog (the article in question being published on the 10th January 2012). See the following link for the relevant search results (last accessed 11/06/2019):

https://www.google.com/search?q=%22southeast+florida+climate+compact%22&biw=1280&bih=566&source=ln&tbs=cdr%3A1%2Ccd_min%3A1%2F10%2F2010%2Ccd_max%3A1%2F10%2F2012&tbm=news.

established at the local level, and their beneficial effects become apparent, it may then be possible to make a stronger case for their implementation elsewhere: either in other localities or on larger scales. One can see some evidence of this happening in the case of the Compact. While it was scarcely reported on when it was established, its successes drew national attention within 5 years: a notable example is the fact that one of the local politicians involved in the Compact received a Presidential commendation for their work to combat the effects of climate change (Southeast Florida Regional Climate Compact 2015). Whether such national prestige can effectively promote the establishment of braces elsewhere is of course an empirical question: notably it relies on lawmakers at the national level having at least the requisite levels of rationality to take such evidence into account; a condition which we should not have undue confidence will be met. On the other hand, the fact that local lawmakers seem to have the requisite rationality to put a brace policy into effect suggests that finding lawmakers with sufficient rationality to appreciate the consequences of such policies is not impossible. Without being overly optimistic, the fact that local braces can achieve successes which are great enough to draw national attention at least provides proponents of braces with more argumentative resources to draw upon when trying to promote them. One further point in favour of implementing at the local level is an ethical one: if one is trialling a new and largely untested policy which may have unforeseen side-effects, it is simply safer from a moral perspective to trial such a policy on a smaller level than a larger one.

To summarise the content of this chapter, braces are subject to large (but not necessarily insurmountable) practical problems. Working in reverse order, when one attempts to promote a brace to fix some element of our existing political environment, one must consider the risk that the existing elements of this environment will themselves undermine public support for the brace. This is the Democratic Problem: the problem being that a combination of public bias, ignorance, misinformation, and the existence of deceptors in the political environment may cause the discussion of whether to implement a brace to be riven by irrationality, ultimately undermining public support. Firstly, if one can meaningfully quantify this risk then this should be one's first concern. Secondly, if there is reason to believe that this risk is present, then one should opt to either attempt to insulate the implementation of the brace from public discussion (while bearing in mind the ethical problems with such an approach), or should attempt to carry out the brace at a local level. If one can overcome the Democratic Problem, one must then consider the Implementation Problem. The Implementation Problem must be addressed

whenever one turns to the actual business of implementing a particular brace, and to overcome this problem one must be able to show that the people tasked with implementing the brace (e.g. fact-checkers, civil servants, etc.) are either equipped with the appropriate mindware and/or situated in the right kind of mental environment to avoid falling prey to exactly the same kinds of cognitive biases that generate the problems that braces are intended to solve.

Conclusion

The outcome of this thesis is, I believe, a mixed one for the brace program. Firstly, I have argued that braces have a powerful normative justification in terms of promoting positive-sum cooperation: this is no doubt a positive for the program. Secondly, and again positively, my analyses in Chapter 3 have demonstrated that the theoretical term deceptor has applications in contemporary social science, and that examples of braces to deal with such deceptors can be given. Indeed, I believe that analyses like those in Chapter 3 could be carried out on other social phenomena: for example, online echo chambers could potentially be argued to be a form of social deceptor. There is also room for further analysis on my chosen examples. Take the discussion of partisan news as a media deceptor, in section 3.3. Because falsehood is non-controversially epistemically problematic, I chose to argue that the misfire condition for partisan news is fulfilled when such news delivers falsehoods (as is often the case). However, there is clearly something epistemically problematic about partisan news even when the information it is delivering is not strictly false, and this is precisely that it is partisan: where it actually presents a view of the facts, it presents a one-sided view. If a sufficiently convincing and non-controversial analysis can be given of why partisanship is epistemically problematic, then this could serve as the basis for a further analysis of why partisan news is a media deceptor, and could potentially generate further recommendations for addressing it.

Thirdly, I have addressed some of the key problems facing the brace program: or, at least, facing Heath and Brennan's intended use of braces to improve the quality of democratic decision-making. As I have argued above, I do not think these problems can necessarily be solved: however, I have attempted to offer empirically grounded suggestions for how to attempt to solve these problems. I believe that my suggestions in the case of the Implementation Problem have some promise, but must admit that I have less optimism when it comes to the Democratic Problem: while local interventions may be possible, I am not yet sure that these would then lead to implementation of braces on a larger scale. The persistence of the problems I have identified is why I describe the brace program's prospects as mixed.

Looking towards another avenue for potential research, I believe that there is a one-sidedness in this thesis which could be rectified by future research. I have chosen to

focus almost exclusively on epistemic problems in this thesis and have avoided getting into motivational problems (see section 2.4), primarily because issues of the truth or falsity of information are less controversial than issues of how people should be motivated to behave. However, I think it is important that such motivational issues are addressed: particularly given my chosen justification for braces. As I have argued in favour of using braces to promote positive-sum cooperation, there is unquestionably a large motivational element involved: people can be made aware of the potential benefits of cooperation, but unless they can then be successfully motivated to engage in such cooperation the positive-sum gains will not be realised. Where there are elements of our mental environment which function to undermine such motivations (e.g. motivational deceptors), it seems clear that the brace program should have something to say about what these elements are and what environmental solutions could be potentially used to address them. To achieve this, one must decide on the standards of correctness for people's motivations: I think the most promising route would be to start by arguing that people should be motivated to achieve positive-sum gains, supplemented with an egalitarian constraint that the costs and benefits of cooperation should be shared in such a way that they at least do not undermine people's motivation to cooperate in the first place. These requirements are quite minimal, however one can already see that this argument is more controversial than the argument that people should not believe falsehoods. In many ways, the question of how to deal with motivational issues exposes something about my handling of epistemic issues: while I explicitly tied the benefits of true beliefs and the harms of false beliefs to their potential to (respectively) facilitate or threaten positive-sum cooperation, my discussion of epistemic issues was no doubt helped along by the common agreement (particularly among philosophers) that having false beliefs is obviously bad. An analysis of motivational issues which followed the lead set in this thesis would likely be more controversial than the analysis of epistemic issues not because the underlying standards of correctness are different, but rather because the same underlying standards of correctness would play a more noticeable role in the argument.

Finally, it is worth 'zooming out' to look at where the brace program stands in relation to its closest intellectual relatives, the nudge and boost programs. I have largely been following Heath and Brennan in my focus on using braces to address problems which undermine the quality of decision-making in democracies. However, this is just one application for braces: as was made clear in section 2.3, there are people who have

suggested brace-like policies to promote paternalistic ends, such as enabling people to avoid making unhealthy food choices or enabling them to spend their time in ways which are more conducive to their overall happiness. Particularly in the case of these paternalistic braces, which have the clearest parallels to the early nudge policies recommended by Thaler and Sunstein, but also with nudges and braces intended to promote social welfare, one can always ask why one would prefer a brace, which is coercive, to a nudge, which is not? I believe that ultimately both policy tools are useful. The question of which to use in particular cases is partly empirical and partly ethical. The empirical question is whether a brace will be more effective than a nudge at achieving the desired outcome: due to their often coercive nature, I would presume that often braces will be more effective, but if it transpires that a nudge would be as effective as a brace in a particular case then it seems fairly obvious from an ethical point of view that the non-coercive nudge should be preferred. The ethical question arises in cases where a brace would be a more effective way of achieving the desired outcome than a nudge: this question is simply whether the welfare gain (or other ethically desirable end that the brace is intended to promote) is worth the coercive imposition associated with the brace. This is hardly a new question, as it must be considered when thinking about implementing any coercive policy, but it is nonetheless an important one.

Situations where both boosts and braces would be useful are chiefly those where we are aiming to support people's capacity to be rational. In the choice of whether to use boosts or braces to achieve this aim, there is a key factor to take into account. If one has confidence that one can equip people with long-lasting competencies in a particular domain, and that these competencies (much like overlearned mindware) will be brought to bear in all relevant situations where cognitive misfires are likely, then it seems that a boost would be a good intervention. However, given that people have limited cognitive resources, we should expect the effectiveness of boosts to have a limit: there will come a point where people's on-board cognitive resources cannot be enhanced further at the individual level. This is where braces, with their distinctly environmental focus, come into play. Braces and boosts can be used together, and I suspect that the choice of which to use in which situation comes down to how we can best direct people's limited cognitive resources: where there are available resources we should consider boosts, where we believe such resources are already taxed, braces have the upper hand.

Acknowledgements

I would like to thank Austin Vanderburgh for several helpful discussions which helped me during the early stages of the thesis. Special thanks must go to Pepijn Al for his support, useful comments, and invaluable discussion both during the writing of this thesis and throughout the Master's programme. I would like to thank my second reader, Joel Anderson, for being a source of stimulating conversations both during this thesis and throughout the programme. Finally, I am very thankful to my supervisor Hanno Sauer, whose insightful comments and suggestions have improved this thesis a great deal.

References²⁵

- Achen, Christopher H., and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, New Jersey: Princeton University Press.
- Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36. <https://doi.org/10.1257/jep.31.2.211>.
- Anderson, Elizabeth. 2011. "Democracy, Public Policy, and Lay Assessments of Scientific Testimony." *Episteme* 8 (2): 144–64. <https://doi.org/10.3366/epi.2011.0013>.
- Anderson, Joel. 2010. "Review of Nudge." *Economics and Philosophy* 26: 369–76. <https://doi.org/10.1017/s1474747209990175>.
- Arceneaux, Kevin, Martin Johnson, and Chad Murphy. 2012. "Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure." *The Journal of Politics* 74 (1): 174–86. <https://doi.org/10.1017/S002238161100123X>.
- Behavioural Insights Team. 2018. "Annual Report 2017-18." <https://www.bi.team/wp-content/uploads/2019/01/Annual-update-report-BIT-2017-2018.pdf>.
- Blake, Aaron. 2018. "A New Study Suggests Fake News Might Have Won Donald Trump the 2016 Election." *The Washington Post*. 2018. https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/?utm_term=.6287a6b8cc96.
- Bovens, Luc. 2009. "The Ethics of Nudge." In *Preference Change: Approaches from Philosophy, Economics, and Psychology*, edited by Till Grüne-Yanoff and Sven Ove Hansson, 207–20. Berlin: Springer.
- Bovet, Alexandre, and Hernán A. Makse. 2019. "Influence of Fake News in Twitter during the 2016 US Presidential Election." *Nature Communications* 10 (1): 1–14. <https://doi.org/10.1038/s41467-018-07761-2>.
- Brennan, Jason. 2011. *The Ethics of Voting*. Princeton, New Jersey: Princeton University Press.
- . 2016. *Against Democracy*. Princeton, New Jersey: Princeton University Press.
- . 2018. "Libertarianism after Nozick." *Philosophy Compass* 13 (2): 1–11. <https://doi.org/10.1111/phc3.12485>.
- Brewer, Marilyn B. 2007. "The Social Psychology of Intergroup Relations: Social Categorization, Ingroup Bias, and Outgroup Prejudice." In *Social Psychology: Handbook of Basic Principles*, edited by Arie W. Kruglanski and E. Tory Higgins, 2nd ed., 695–715. New York: The Guilford Press.
- Briony, Swire, Berinsky Adam J., Lewandowsky Stephan, and Ecker Ullrich K. H. 2017. "Processing Political Misinformation: Comprehending the Trump Phenomenon." *Royal Society Open Science* 4 (3): 160802. <https://doi.org/10.1098/rsos.160802>.
- Brooks, Jeremy S., Timothy M. Waring, Monique Borgerhoff Mulder, and Peter J. Richerson. 2018. "Applying Cultural Evolution to Sustainability Challenges: An Introduction to the Special Issue." *Sustainability Science* 13 (1): 1–8. <https://doi.org/10.1007/s11625-017-0516-3>.
- Broward County Government. 2018. "County Commission." <http://www.broward.org/Commission/Pages/default.aspx>.
- Busselle, Rick W., and L. J. Shrum. 2003. "Media Exposure and Exemplar Accessibility." *Media Psychology* 5 (3): 255–82. <https://doi.org/10.1207/S1532785XMEP0503>.
- Cadwalladr, Carole. 2017. "The Great British Brexit Robbery: How Our Democracy Was

²⁵ All references with attached URLs were last accessed 12/06/2019.

- Hijacked.” *The Guardian*. 2017.
<https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy>.
- Caplan, Bryan. 2007. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton, New Jersey: Princeton University Press.
- Clark, Andy. 1997. “Economic Reason: The Interplay of Individual Learning and External Structure.” In *The Frontiers of the New Institutional Economics*, edited by John N. Drobak and John V.C. Nye, 269–90. Academic Press.
- Coe, Kevin, David Tewksbury, Bradley J. Bond, Kristin L. Drogos, Robert W. Porter, Ashley Yahn, and Yuanyuan Zhang. 2008. “Hostile News: Partisan Use and Perceptions of Cable News Programming.” *Journal of Communication* 58 (2): 201–19.
<https://doi.org/10.1111/j.1460-2466.2008.00381.x>.
- Connolly, Kate. 2019. “Germany’s AfD Turns on Greta Thunberg as It Embraces Climate Denial.” *The Guardian*, May 14, 2019.
<https://www.theguardian.com/environment/2019/may/14/germanys-afd-attacks-greta-thunberg-as-it-embraces-climate-denial>.
- Cook, John, Naomi Oreskes, Peter T. Doran, William R. L. Anderegg, Bart Verheggen, Ed W. Maibach, J. Stuart Carlton, et al. 2016. “Consensus on Consensus: A Synthesis of Consensus Estimates on Human-Caused Global Warming.” *Environmental Research Letters* 11 (4): 048002. <https://doi.org/10.1088/1748-9326/11/4/048002>.
- Cosmides, Leda, and John Tooby. 1994. “Better than Rational: Evolutionary Psychology and the Invisible Hand.” *The American Economic Review* 84 (2): 327–32.
- Dentith, M. R. X. 2016. “The Problem of Fake News.” *Public Reason* 8 (1–2): 65–79.
- Dormandy, Katherine. 2018. “Epistemic Authority: Preemption or Proper Basing?” *Erkenntnis* 83 (4): 773–91. <https://doi.org/10.1007/s10670-017-9913-3>.
- Dreyfuss, Emily. 2017. “The Cognitive Bias President Trump Understands Better than You.” *Wired*, February 18, 2017. <https://www.wired.com/2017/02/cognitive-bias-president-trump-understands-better/>.
- Dunbar, Robin. 2014. *Human Evolution*. St Ives: Pelican Books.
- Evans, Jonathan St. B. T. 2006a. “Dual System Theories of Cognition: Some Issues.” *Proceedings of the Annual Meeting of the Cognitive Science Society* 28: 202–7.
- . 2006b. “The Heuristic-Analytic Theory of Reasoning: Extension and Evaluation.” *Psychonomic Bulletin & Review* 13 (3): 378–95.
<https://doi.org/https://doi.org/10.3758/BF03193858>.
- . 2008. “Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition.” *Annual Review of Psychology* 59: 255–78.
<https://doi.org/10.1146/annurev.psych.59.103006.093629>.
- Evans, Jonathan St. B. T., and Keith Frankish, eds. 2009. *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof>.
- Evans, Jonathan St. B. T., and Keith E. Stanovich. 2013. “Dual-Process Theories of Higher Cognition: Advancing the Debate.” *Perspectives on Psychological Science* 8 (3): 223–41. <https://doi.org/10.1177/1745691612460685>.
- FactCheck.org. 2019. “Our Process.” FactCheck.Org. 2019.
<https://www.factcheck.org/our-process/>.
- Frankish, Keith. 2010. “Dual-Process and Dual-System Theories of Reasoning.” *Philosophy Compass* 5 (10): 914–26. <https://doi.org/10.1111/j.1747-9991.2010.00330.x>.
- Frankish, Keith, and Jonathan St. B. T. Evans. 2009. “The Duality of Mind: An Historical Perspective.” In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 1–32. Oxford: Oxford University

- Press.
- Freiman, Christopher. 2017. *Unequivocal Justice*. New York: Routledge.
- . 2018. “Ideal Theory.” In *The Routledge Handbook of Libertarianism*, edited by Jason Brennan, Bas van der Vossen, and David Schmidtz, 301–11. New York: Routledge.
- Frey, Bruno S., and Alois Stutzer. 2006. “Mispredicting Utility and the Political Process.” In *Behavioral Public Finance*, edited by Edward J. McCafferey and Joel Slemrod, 113–40. New York: Russell Sage.
- Full Fact. 2019. “Frequently Asked Questions.” Full Fact. 2019. <https://fullfact.org/about/frequently-asked-questions/>.
- Gillin, Joshua. 2016. “NFL’s Colin Kaepernick Incorrectly Credits Winston Churchill for Quote about Lies.” Politifact. 2016. <https://www.politifact.com/punditfact/statements/2017/oct/09/colin-kaepernick/nfls-colin-kaepernick-incorrectly-credits-winston-/>.
- Goldberg, Sandy. 2011. “The Division of Epistemic Labor.” *Episteme* 8 (1): 112–25. <https://doi.org/10.3366/epi.2011.0010>.
- Goldman, Alvin I. 1999. *Knowledge in a Social World*. Oxford: Oxford University Press.
- Goldman, Seth K., and Diana C. Mutz. 2011. “The Friendly Media Phenomenon: A Cross-National Analysis of Cross-Cutting Exposure.” *Political Communication* 28 (1): 42–66. <https://doi.org/10.1080/10584609.2010.544280>.
- Greene, Joshua D. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. London: Atlantic Books.
- Grice, Andrew. 2017. “Fake News Handed Brexiteers the Referendum – and Now They Have No Idea What They’re Doing.” *The Independent*. 2017. <https://www.independent.co.uk/voices/michael-gove-boris-johnson-brexit-euro-sceptic-press-theresa-may-a7533806.html>.
- Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David M. J. Lazer. 2019. “Fake News on Twitter during the 2016 U.S. Presidential Election.” *Science* 363 (6425): 374–78. <https://doi.org/10.1126/science.aau2706>.
- Grüne-Yanoff, Till, and Ralph Hertwig. 2016. “Nudge Versus Boost: How Coherent Are Policy and Theory?” *Minds and Machines* 26 (1–2): 149–83. <https://doi.org/10.1007/s11023-015-9367-9>.
- Hall, Lars, Thomas Strandberg, Philip Pärnamets, Andreas Lind, Betty Tärning, and Petter Johansson. 2013. “How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions.” *PLoS ONE* 8 (4): e60554. <https://doi.org/10.1371/journal.pone.0060554>.
- Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science* 162 (3859): 1243–48. <https://doi.org/10.1126/science.162.3859.1243>.
- Heath, Joseph. n.d. “The Democracy Deficit in Canada.” <http://homes.chass.utoronto.ca/~jheath/democracy.pdf>.
- . 2001a. “Problems in the Theory of Ideology.” In *Pragmatism and Critical Theory*, edited by James Bohman and William Rehg, 163–90. Cambridge, Massachusetts: MIT Press.
- . 2001b. *The Efficient Society: Why Canada Is as Close to Utopia as It Gets*. Toronto: Penguin Books.
- . 2006. “The Benefits of Cooperation.” *Philosophy & Public Affairs* 34 (4): 313–51. <https://doi.org/https://doi.org/10.1111/j.1088-4963.2006.00073.x>.
- . 2013. “Ideal Theory in an Nth-Best World: The Case of Pauper Labor.” *Journal of Global Ethics* 9 (2): 159–72. <https://doi.org/10.1080/17449626.2013.818455>.
- . 2014a. “Contractualism: Micro and Macro.” In *Morality, Competition, and the Firm*, 145–72. New York: Oxford University Press.
- . 2014b. *Enlightenment 2.0*. Toronto: Harper Collins Publishers Ltd.

- . 2014c. *Morality, Competition, and the Firm*. New York: Oxford University Press.
- . 2014d. “Restoring Sanity to Politics.” In *Due Course*. 2014.
<http://induecourse.ca/restoring-sanity-to-politics/>.
- . 2014e. “Waldron, Sunstein, and Nudge Paternalism.” In *Due Course*. 2014.
<http://induecourse.ca/waldron-sunstein-and-nudge-paternalism/>.
- . 2015a. “Graham Steele: What I Learned About Politics.” In *Due Course*. 2015.
<http://induecourse.ca/graham-steele-what-i-learned-about-politics/>.
- . 2015b. “Response to Tabarrok.” In *Due Course*. 2015.
<http://induecourse.ca/response-to-tabarrok/>.
- Heath, Joseph, and Joel Anderson. 2010. “Procrastination and the Extended Will: Extended Cut.” <http://homes.chass.utoronto.ca/~jheath/extended-will-lv.pdf>.
- Hertwig, Ralph, and Till Grüne-Yanoff. 2017. “Nudging and Boosting: Steering or Empowering Good Decisions.” *Perspectives on Psychological Science* 12 (6): 973–86.
<https://doi.org/10.1177/1745691617702496>.
- Holan, Angie Drobnic. 2018. “The Principles of the Truth-O-Meter: PolitiFact’s Methodology for Independent Fact-Checking.” *Politifact*. 2018.
<https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/>.
- Institute for Economics and Peace. 2016. “Global Terrorism Index 2016.”
<http://economicsandpeace.org/wp-content/uploads/2016/11/Global-Terrorism-Index-2016.2.pdf>.
- Jäger, Christoph. 2016. “Epistemic Authority, Preemptive Reasons, and Understanding.” *Episteme* 13 (2): 167–85. <https://doi.org/10.1017/epi.2015.38>.
- John, Peter, Sarah Cotterill, Alice Moseley, Liz Richardson, Graham Smith, Gerry Stoker, and Corinne Wales. 2011. *Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour*. London: Bloomsbury Academic.
- Kahan, Dan M. 2010. “Fixing the Communications Failure.” *Nature* 463 (21): 296–97.
<https://doi.org/https://doi.org/10.1038/463296a>.
- . 2013a. “A Risky Science Communication Environment for Vaccines.” *Science* 342 (6154): 53–54. <https://doi.org/10.1126/science.1245724>.
- . 2013b. “Ideology, Motivated Reasoning, and Cognitive Reflection.” *Judgment and Decision Making* 8 (4): 407–24.
- . 2015. “Climate-Science Communication and the Measurement Problem.” *Political Psychology* 36 (S1): 1–43. <https://doi.org/10.1111/pops.12244>.
- . 2016a. “The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It.” *Emerging Trends in the Social and Behavioral Sciences*, 1–16. <https://doi.org/10.1002/9781118900772.etrds0417>.
- . 2016b. “The Politically Motivated Reasoning Paradigm, Part 2: Unanswered Questions.” *Emerging Trends in the Social and Behavioral Sciences*, 1–15.
<https://doi.org/10.1002/9781118900772.etrds0418>.
- Kahan, Dan M., Hank Jenkins-Smith, and Donald Braman. 2011. “Cultural Cognition of Scientific Consensus.” *Journal of Risk Research* 14 (2): 147–74.
<https://doi.org/10.1080/13669877.2010.511246>.
- Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. “Motivated Numeracy and Enlightened Self-Government.” *Behavioural Public Policy* 1 (1): 54–86.
<https://doi.org/10.2139/ssrn.2319992>.
- Kahan, Dan M., Paul Slovic, Donald Braman, and John Gastil. 2006. “Review: Fear of Democracy: A Cultural Evaluation of Sunstein on Risk.” *Harvard Law Review* 119 (4): 1071–1109.
- Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Penguin Books.
- Kahneman, Daniel, and Shane Frederick. 2005. “A Model of Heuristic Judgment.” In *The*

- Cambridge Handbook of Thinking and Reasoning*, edited by Keith J. Holyoak and Robert G. Morrison, 267–93. Cambridge: Cambridge University Press.
<https://doi.org/10.1111/cogs.12119>.
- Kahneman, Daniel, and Gary Klein. 2009. “Conditions for Intuitive Expertise: A Failure to Disagree.” *American Psychologist* 64 (6): 515–26.
<https://doi.org/10.1037/a0016755>.
- Kelly, Jamie. 2013. “Libertarian Paternalism, Utilitarianism, and Justice.” In *Paternalism: Theory and Practice*, edited by Christian Coons and Michael Weber, 216–30. Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9781139179003.012>.
- Kraft, Patrick W., Milton Lodge, and Charles S. Taber. 2015. “Why People ‘Don’t Trust the Evidence.’” *The ANNALS of the American Academy of Political and Social Science* 658: 121–33. <https://doi.org/10.1177/0002716214554758>.
- Kurtzleben, Danielle. 2018. “Did Fake News On Facebook Help Elect Trump? Here’s What We Know.” NPR. 2018. <https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election>.
- Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. “The Science of Fake News.” *Science* 359 (6380): 1094–96. <https://doi.org/10.1126/science.aao2998>.
- Levy, Neil. 2012. “Ecological Engineering: Reshaping Our Environments to Achieve Our Goals.” *Philosophy & Technology* 25: 589–604. <https://doi.org/10.1007/s13347-012-0065-8>.
- . 2017a. “Nudges in a Post-Truth World.” *Journal of Medical Ethics* 43: 495–500. <https://doi.org/10.1136/medethics-2017-104153>.
- . 2017b. “The Bad News About Fake News.” *Social Epistemology Review and Reply Collective* 6 (8): 20–36.
- . 2018. “Why Nudging Is No More Paternalistic Than Arguing.” *The Philosopher’s Magazine* 83: 53–59.
- Levy, Neil, and Eric Mandelbaum. 2014. “The Powers That Bind: Doxastic Voluntarism and Epistemic Obligation.” In *The Ethics of Belief*, edited by Jonathan Matheson and Rico Vitz, 15–32. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof>.
- Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. 1978. “Judged Frequency of Lethal Events.” *Journal of Experimental Psychology: Human Learning and Memory* 4 (6): 551–78.
- Machery, Edouard, and Ron Mallon. 2010. “The Evolution of Morality.” In *The Moral Psychology Handbook*, edited by John M. Doris, 3–46. Oxford: Oxford University Press.
- Mandelbaum, Eric, and Jake Quilty-Dunn. 2015. “Believing without Reason, or: Why Liberals Shouldn’t Watch Fox News.” *The Harvard Review of Philosophy* XXII: 42–52.
- Marcus, Gary. 2008. *Kluge: The Haphazard Construction of the Human Mind*. New York: Houghton Mifflin Books.
- Markandya, Anil, and Paul Wilkinson. 2007. “Electricity Generation and Health.” *Lancet* 370 (9591): 979–90. [https://doi.org/10.1016/S0140-6736\(07\)61253-7](https://doi.org/10.1016/S0140-6736(07)61253-7).
- Melnikoff, David E., and John A. Bargh. 2018. “The Mythical Number Two.” *Trends in Cognitive Sciences* 22 (4): 280–93. <https://doi.org/10.1016/j.tics.2018.02.001>.
- Mercier, Hugo, and Dan Sperber. 2009. “Intuitive and Reflective Inferences.” In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 149–70. Oxford: Oxford University Press.
- Miami-Dade County Government. 2018. “Board of County Commissioners.” <https://www.miamidade.gov/commission/>.

- Mill, John Stuart. 1859. "On Liberty." In *On Liberty, Utilitarianism, and Other Essays*, edited by Mark Philip and Frederick Rosen, 1–112. Oxford: Oxford University Press.
- Millies, Stephen P., Stuart Gottlieb, James H. Lee, Margaret F. Nichols, Don Shipp, David Golub, Peter Schmidt, Mac Brachman, and Bernard Langs. 2016. "Irrationality in Politics." *New York Times*, 2016. <https://www.nytimes.com/2016/02/07/opinion/sunday/irrationality-in-politics.html>.
- Monroe County Government. 2019. "Board of County Commissioners." <https://www.monroecounty-fl.gov/1015/Board-of-County-Commissioners>.
- Mukerji, Nikil. 2018. "What Is Fake News?" *Ergo* 5 (35): 923–46. <https://doi.org/10.5840/tpm20188399>.
- Napolitano, Maria Giulia. 2018. "Epistemic Dependence on Official Experts." Utrecht University. https://dspace.library.uu.nl/bitstream/handle/1874/364664/thesis_final.pdf?sequence=2&isAllowed=y.
- Nyhan, Brendan, and Jason Reifler. 2019. "The Roles of Information Deficits and Identity Threat in the Prevalence of Misperceptions." *Journal of Elections, Public Opinion and Parties* 29 (2): 222–44. <https://doi.org/10.1080/17457289.2018.1465061>.
- OECD. 2017. "Obesity Update 2017." <https://doi.org/10.1007/s11428-017-0241-7>.
- Paikin, Steven. 2016. "The Agenda with Steven Paikin - In Search of Better Voters (with Jason Brennan)." Canada: TVOntario. https://www.youtube.com/watch?v=_f1NXUI_K_I.
- Palm Beach County. 2018. "County Commissioners." <http://discover.pbcgov.org/countycommissioners/Pages/default.aspx>.
- Pennycook, Gordon, Wim De Neys, Jonathan St. B.T. Evans, Keith E. Stanovich, and Valerie A. Thompson. 2018. "The Mythical Dual-Process Typology." *Trends in Cognitive Sciences* 22 (8): 667–68. <https://doi.org/10.1016/j.tics.2018.04.008>.
- Pennycook, Gordon, and David G. Rand. 2018. "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning." *Cognition* In press: 1–12. <https://doi.org/10.1016/j.cognition.2018.06.011>.
- Persily, Nathaniel. 2017. "The 2016 U.S. Election: Can Democracy Survive the Internet." *Journal of Democracy* 28 (2): 63–76. <https://doi.org/10.13140/RG.2.1.2972.6489>.
- Pfeiffer, Chloe. 2016. "Brexit Is a Perfect Example of Irrational Behavior." *Business Insider*, June 21, 2016. <https://www.businessinsider.com/thaler-brexit-shows-irrational-behavior-2016-6?international=true&r=US&IR=T>.
- Pinker, Steven. 2018. "The Media Exaggerates Negative News. This Distortion Has Consequences." *The Guardian*, February 17, 2018. <https://www.theguardian.com/commentisfree/2018/feb/17/steven-pinker-media-negative-news>.
- Prooijen, Jan Willem van, and Karen M. Douglas. 2018. "Belief in Conspiracy Theories: Basic Principles of an Emerging Research Domain." *European Journal of Social Psychology* 48 (7): 897–908. <https://doi.org/10.1002/ejsp.2530>.
- Punditfact. 2019a. "Fox's File." Politifact. 2019. <https://www.politifact.com/punditfact/tv/fox/>.
- . 2019b. "NBC's File." Politifact. 2019. <https://www.politifact.com/punditfact/tv/nbc/>.
- Rawls, John. 1999. *A Theory of Justice: Revised Edition*. Cambridge, Massachusetts: Harvard University Press.
- Read, Max, and David Wallace-Wells. 2019. "8 Predictions for What the World Will Look Like in 20 Years." *New York Magazine*, January 6, 2019.

- <http://nymag.com/intelligencer/2019/01/2038-podcast-predictions.html>.
- Rebonato, Riccardo. 2013. "A Critical Assessment of Libertarian Paternalism." <https://doi.org/10.1007/s10603-014-9265-1>.
- Rini, Regina. 2017. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27 (2S): E-43-E-64. <https://doi.org/10.1353/ken.2017.0025>.
- Roberts, J. M., and Odd Arne Westad. 2013. *The Penguin History of the World*. 6th ed. London: Penguin Books.
- Ruggeri, Amanda. 2017. "Miami's Fight against Rising Seas." *BBC Future*, April 4, 2017. <http://www.bbc.com/future/story/20170403-miamis-fight-against-sea-level-rise>.
- Sauer, Hanno. 2018. *Moral Thinking, Fast and Slow*. Abingdon, UK: Routledge.
- Schubert, Christian. 2017. "Green Nudges: Do They Work? Are They Ethical?" *Ecological Economics* 132: 329–42. <https://doi.org/10.1016/j.ecolecon.2016.11.009>.
- Schwarz, Norbert, Eryn Newman, and William Leach. 2016. "Making the Truth Stick & the Myths Fade: Lessons from Cognitive Psychology." *Behavioral Science & Policy* 2 (1): 85–95. <https://doi.org/10.1353/bsp.2016.0009>.
- Shapiro, Ian. 2003. *The State of Democratic Theory*. Princeton, New Jersey: Princeton University Press.
- Silverman, Craig. 2016. "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook." BuzzFeed. 2016. <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>.
- Skurnik, Ian, Carolyn Yoon, Denise C. Park, and Norbert Schwarz. 2005. "How Warnings about False Claims Become Recommendations." *Journal of Consumer Research* 31 (4): 713–24. <https://doi.org/10.1086/426605>.
- Slovic, Paul, Baruch Fischhoff, Sarah Lichtenstein, and F. J. C. Roe. 1981. "Perceived Risk: Psychological Factors and Social Implications [and Discussion]." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 376 (1764): 17–34.
- Smith, Eliot R., and Jamie DeCoster. 2000. "Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems." *Personality and Social Psychology Review* 4 (2): 108–31.
- Smith, Jeff. 2017. "Designing Against Misinformation." *Medium*, December 20, 2017. <https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2>.
- Snopes. 2019. "What Is Your Fact-Checking Process?" Snopes. 2019. <https://www.snopes.com/faq/fact-checking-process/>.
- Somin, Ilya. 2013. *Democracy and Political Ignorance: Why Smaller Government Is Smarter*. Stanford, California: Stanford University Press.
- Southeast Florida Regional Climate Compact. 2010. "Southeast Florida Regional Climate Change Compact." <http://southeastfloridaclimatecompact.org/wp-content/uploads/2014/09/compact.pdf>.
- . 2015. "President Obama to Florida Keys "Good Job Addressing Climate Change"." 2015. <http://southeastfloridaclimatecompact.org/news/president-obama-to-florida-keys-good-job-addressing-climate-change/>.
- Stanovich, Keith E. 2004. *The Robot's Rebellion*. Chicago: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226771199.001.0001>.
- . 2009. "Distinguishing the Reflective, Algorithmic and Autonomous Minds: Is It Time for a Tri-Process Theory?" In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 55–88. Oxford: Oxford University Press.
- . 2010. *Decision Making and Rationality in the Modern World*. Oxford: Oxford University Press.

- . 2011. *Rationality and the Reflective Mind*. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780195341140.001.0001>.
- . 2018. “Miserliness in Human Cognition: The Interaction of Detection, Override and Mindware.” *Thinking and Reasoning* 24 (4): 423–44.
<https://doi.org/10.1080/13546783.2018.1459314>.
- Stanovich, Keith E., and Richard F. West. 2008. “On the Relative Independence of Thinking Biases and Cognitive Ability.” *Journal of Personality and Social Psychology* 94 (4): 672–95. <https://doi.org/10.1037/0022-3514.94.4.672>.
- Stanovich, Keith E., Richard F. West, and Maggie E. Toplak. 2013. “Myside Bias, Rational Thinking, and Intelligence.” *Current Directions in Psychological Science* 22 (4): 259–64. <https://doi.org/10.1177/0963721413480174>.
- Statista. 2019. “Leading Cable Networks in the United States as of April 29, 2019, by Number of Total Day Viewers (in Thousands).”
<https://www.statista.com/statistics/347040/cable-networks-viewers-usa/>.
- Subramanian, Samanth. 2017. “Inside the Macedonian Fake News Complex.” *Wired*. 2017. <https://www.wired.com/2017/02/veles-macedonia-fake-news/>.
- Sunstein, Cass R. 2014. *Why Nudge? The Politics of Libertarian Paternalism*. Yale University Press.
- . 2016. *The Ethics of Influence*. Cambridge: Cambridge University Press.
- Sunstein, Cass R., and Richard H. Thaler. 2003. “Libertarian Paternalism Is Not an Oxymoron.” *The University of Chicago Law Review* 70 (4): 1159–1202.
- Swift, Art. 2016. “Americans’ Trust in Mass Media Sinks to New Low.” Gallup. 2016. <https://news.gallup.com/poll/195542/americans-trust-mass-media-sinks-new-low.aspx>.
- Taber, Charles S., and Milton Lodge. 2006. “Motivated Skepticism in the Evaluation of Political Beliefs.” *American Journal of Political Science* 50 (3): 755–69.
<https://doi.org/10.1111/j.1540-5907.2006.00214.x>.
- Talisse, Robert B. 2018. “The Trouble with Hooligans.” *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–12. <https://doi.org/10.1080/0020174X.2018.1502933>.
- Tarcan Kumkale, G., and Dolores Albarracín. 2004. “The Sleeper Effect in Persuasion: A Meta-Analytic Review.” *Psychological Bulletin* 130 (1): 143–72.
<https://doi.org/10.1037/0033-2909.130.1.143>.
- Taylor, Luke E., Amy L. Swerdfeger, and Guy D. Eslick. 2014. “Vaccines Are Not Associated with Autism: An Evidence-Based Meta-Analysis of Case-Control and Cohort Studies.” *Vaccine* 32 (29): 3623–29.
<https://doi.org/10.1016/j.vaccine.2014.04.085>.
- Thaler, Richard H., and Cass R. Sunstein. 2003. “Libertarian Paternalism.” *American Economic Review* 93 (2): 175–79. <https://doi.org/10.1257/000282803321947001>.
- . 2008. *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin Books.
- The Economist. 2019. “A Surge for the FVD, a New Right-Wing Dutch Party.” *The Economist*, March 30, 2019. <https://www.economist.com/europe/2019/03/30/a-surge-for-the-fvd-a-new-right-wing-dutch-party>.
- Todd, Peter M., and Gerd Gigerenzer. 2000. “Précis of Simple Heuristics That Make Us Smart.” *Behavioral and Brain Sciences* 23: 727–80.
- Townsend, Tess. 2016. “The Bizarre Truth Behind the Biggest Pro-Trump Facebook Hoaxes.” *Inc*. 2016. <https://www.inc.com/tess-townsend/ending-fed-trump-facebook.html>.
- Toynbee, Polly. 2019. “The Anti-EU Lies Are Back to Exploit Britain’s Weak Spot Again.” *The Guardian*. 2019.
<https://www.theguardian.com/commentisfree/2019/mar/04/anti-eu-brexite-fake->

- news.
- Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5: 207–32.
- Tynan, Dan. 2016. "How Facebook Powers Money Machines for Obscure Political 'news' Sites." *The Guardian*. 2016.
<https://www.theguardian.com/technology/2016/aug/24/facebook-clickbait-political-news-sites-us-election-trump>.
- Vugts, Anastasia, Mariëtte van den Hoven, Emely de Vet, and Marcel Verweij. 2018. "How Autonomy Is Understood in Discussions on the Ethics of Nudging." *Behavioural Public Policy*, 1–16. <https://doi.org/10.1017/bpp.2018.5>.
- Wansink, Brian. 2010. "From Mindless Eating to Mindlessly Eating Better." *Physiology and Behavior* 100 (5): 454–63. <https://doi.org/10.1016/j.physbeh.2010.05.003>.
- Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, Massachusetts: Belknap Press.
- Wolff, Jonathan. 2011. *Ethics and Public Policy: A Philosophical Inquiry*. Abingdon: Routledge.
- Wright, Mike, Robert Mendrick, Christopher Hope, and Gordon Rayner. 2019. "Facebook Paid Hundreds of Thousands to Host Anti-Brexit 'Fake News.'" *The Telegraph*. 2019. <https://www.telegraph.co.uk/news/2019/01/18/facebook-accused-pumping-fake-news-running-ads-claiming-endangered/>.
- Xu, Jiaquan, Sherry L. Murphy, Kenneth D. Kochanek, and Elizabeth Arias. 2016. "Mortality in the United States, 2015."
<https://www.cdc.gov/nchs/data/databriefs/db267.pdf>.
- Zagzebski, Linda. 2013. "A Defense of Epistemic Authority." *Res Philosophica* 90 (2): 293–306. <https://doi.org/http://dx.doi.org/10.11612/resphil.2013.90.2.12?c>.