UNIVERSITY OF WARWICK

The importance of making assumptions: why confirmation is not necessarily a bias

by

Jess Whittlestone

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Behavioural Science Group
Warwick Business School

July 2017

Contents

List of Figures vi

List of Tables vii

Acknowledgements viii

Declaration of Authorship x

Abstract xii

1 Introduction 1

2 When is confirmation a bias? 7

2.1 Introduction . . . 7
2.2 Different types of confirmation bias: a review . . . 9
2.2.1 Bias in search . . . 12
2.2.1.1 Bias in hypothesis testing - Wason's original experiments 14
2.2.1.2 Subsequent research on hypothesis testing . . . 15
2.2.1.3 Selective exposure . . . 17
2.2.1.4 Bias in amount of time spent searching . . . 19
2.2.1.5 'Myside bias' in producing arguments . . . 20
2.2.1.6 Bias in search: discussion . . . 21
2.2.2 Bias in inference . . . 24
2.2.2.1 Biased interpretation of evidence . . . 26
2.2.2.2 Biased evaluation of evidence . . . 30
2.2.2.3 Bias in inference: discussion . . . 31
2.2.3 The relationship between bias in search and inference . . . 34
2.2.4 Biased beliefs: belief persistence and overconfidence . . . 35
2.2.4.1 Belief persistence . . . 36
2.2.4.2 Overconfidence . . . 39
2.2.4.3 Belief persistence and overconfidence: discussion . . . 41
2.2.4.4 Biased beliefs independent of search and inference? . . . 42
2.2.5 Summary: the evidence for confirmation bias . . . 44
2.3 The challenges to confirmation bias . . . 46

2.3.1 Strategies that look 'confirmatory' do not necessarily lead to a confirmation bias in all circumstances . . . 47
2.3.2 Strategies that look like they produce 'bias' may in fact be accurate on average . . . 48
2.3.3 Problems with experimental design . . . 50
2.3.4 Lack of normative standards . . . 51
2.3.5 Bias in search and inference cannot be studied in isolation . . . 54
2.4 What remains of confirmation bias? . . . 56
2.5 Conclusion . . . 64

3 The mixed evidence for selective exposure 65
3.1 Introduction . . . 65
3.2 Background . . . 67
3.3 Past research on selective exposure . . . 69
3.3.1 A broad overview . . . 70
3.3.2 A narrower look at selective exposure - social and political attitudes 71
3.3.2.1 Strength of opinions and dissonance . . . 72
3.3.2.2 Goals and motivations . . . 74
3.3.2.3 Personality differences . . . 75
3.3.2.4 Other factors influencing information choice . . . 76
3.3.3 Naturalistic and field experiments . . . 78
3.3.4 Summary: the state of selective exposure research . . . 80
3.4 The present research . . . 82
3.4.1 Experimental paradigm . . . 82
3.4.2 Experiment 1: setting up a selective exposure paradigm . . . 83
3.4.2.1 Background . . . 83
3.4.2.2 Design and procedure . . . 84
3.4.2.3 Handling of moderates . . . 86
3.4.2.4 Results . . . 87
3.4.2.5 Analysis of moderates . . . 89
3.4.3 Experiments 2 and 3: how robust is the evidence for selective exposure? . . . 90
3.4.3.1 Background . . . 90
3.4.3.2 Design and procedure . . . 91
3.4.3.3 Results . . . 92
3.4.4 Experiment 4: a replication of Taber and Lodge, 2006 . . . 95
3.4.4.1 Background . . . 95
3.4.4.2 Design and procedure . . . 95
3.4.4.3 Results . . . 96
3.4.5 Experiments 5 and 6 . . . 97
3.4.6 Experiment 5: information source manipulation . . . 99
3.4.6.1 Design and procedure . . . 99
3.4.6.2 Results . . . 99
3.4.6.3 Participant's reported reasons for choices . . . 101
3.4.7 Experiment 6: information source manipulation 2 . . . 102
3.4.7.1 Design and procedure . . . 103
3.4.7.2 Results . . . 103

3.5 Summary and discussion . . . 105
3.5.1 Summary . . . 105
3.5.2 Discussion: just another failed replication? . . . 106
3.6 General discussion . . . 110
3.6.1 Design issues in selective exposure research . . . 110
3.6.2 Different ways of measuring 'selective exposure' . . . 111
3.6.3 Selective exposure and bias . . . 113
3.6.4 Selective exposure in the 'real world' . . . 115
3.6.5 Final thoughts and implications . . . 116

4 Bias, rationality and improving human reasoning 118
4.1 Introduction . . . 118
4.2 What does it mean to be biased? . . . 120
4.2.1 'Bias' in different areas of research . . . 120
4.2.2 'Bias' in the confirmation bias literature . . . 124
4.3 Normative models . . . 127
4.3.1 The use of normative models in psychology . . . 127
4.3.2 Normative models in the study of confirmation bias . . . 129
4.4 What does it mean to be 'rational'? . . . 132
4.4.1 The relationship between bias and rationality . . . 132
4.4.2 Different types of rationality . . . 134
4.4.2.1 Summary of different types of rationality . . . 135
4.4.2.2 Links and overlaps between different types of rationality 138
4.4.3 Disagreements about rationality: summary . . . 140
4.4.3.1 Substantive disagreement or terminological confusion? . . 141
4.4.3.2 Disagreements about rationality in the confirmation bias literature . . . 144
4.4.4 Why does this matter? Rationality and improving human reasoning 147
4.5 Summary . . . 150
4.5.1 What does it mean to be biased? . . . 150
4.5.2 Normative models in psychology . . . 151
4.5.3 Different types of rationality . . . 151
4.5.4 Bias, rationality, and improving reasoning . . . 153
4.6 Implications for confirmation bias . . . 153

5 Open-mindedness 157
5.1 Introduction . . . 157
5.2 What is open-mindedness? . . . 159
5.2.1 An intuitive picture of open-mindedness . . . 159
5.2.2 Open-mindedness in psychology . . . 161
5.2.2.1 Early accounts of open-mindedness . . . 161
5.2.2.2 Openness as a personality dimension in the five-factor model . . . 161
5.2.2.3 Open-mindedness as behaviour . . . 164
5.2.2.4 Open-mindedness in psychology: summary . . . 165
5.2.3 Open-mindedness in philosophy . . . 166
5.2.3.1 Open-mindedness as uncertainty . . . 166

5.2.3.2 Open-mindedness as intellectual humility . . . 167
5.2.3.3 Open-mindedness as detachment or engagement . . . 169
5.2.3.4 Open-mindedness in philosophy: summary . . . 171
5.2.4 Summary: what is open-mindedness? . . . 172
5.3 Should we be more open-minded? . . . 173
5.3.1 The costs of open-mindedness . . . 174
5.3.2 Why open-mindedness is not a normative concept . . . 175
5.3.3 Why open-mindedness does not need to be a normative concept . 177
5.3.4 Open-mindedness as an explore-exploit tradeoff . . . 179
5.3.5 Open-mindedness and science . . . 180
5.3.6 Summary: would more open-mindedness be better? . . . 183
5.4 Implications . . . 184
5.4.1 Implications for the psychological study of open-mindedness . . . 185
5.4.2 Implications for promoting open-mindedness . . . 188
5.5 Conclusion . . . 190

6 Summary and discussion 192
6.1 Summary . . . 192
6.1.1 When is confirmation a bias? . . . 192
6.1.2 The mixed evidence for selective exposure . . . 193
6.1.3 Bias, rationality, and confirmation . . . 194
6.1.4 Open-mindedness . . . 195
6.1.5 Confirmation bias, open-mindedness, and Bayes' rule . . . 196
6.2 Discussion . . . 198
6.2.1 Why is the belief in confirmation bias so pervasive? . . . 198
6.2.1.1 A theoretical case for confirmation bias? . . . 199
6.2.1.2 Confirmatory reasoning as a (not necessarily irrational) tendency . . . 200
6.2.1.3 Confirmation bias as a convenient explanation . . . 201
6.2.1.4 Confusing confirmation bias for something else . . . 201
6.2.1.5 The role of emotions . . . 202
6.2.1.6 The role of trust in sources . . . 203
6.2.2 Confirmation bias and real-world problems . . . 205
6.2.2.1 Confirmation bias and politics . . . 208
6.2.2.2 Improving forecasting accuracy . . . 209
6.2.2.3 A new way of thinking about bias . . . 210
6.2.3 Implications for future research . . . 211
6.2.3.1 More attention and clarity around normative issues . . . 211
6.2.3.2 Studying different stages of reasoning in conjunction, not as isolated phenomena . . . 212
6.2.3.3 Clarifying the relationship between confirmation bias and related concepts . . . 213
6.2.3.4 More specific research directions . . . 213

7 Conclusion: the costs and benefits of making assumptions 216

8 Final reflections 218

A Factors influencing selective exposure **223**

B Selective exposure studies of social/political attitudes **225**

C Ideology-based measures of selective exposure **232**

D Materials used in experiments **234**
D.1 Experiment 1 234
D.1.1 Opinion measures 234
D.1.2 Arguments used 235
D.2 Experiments 2 and 3 246
D.3 Experiments 4-6 248
D.3.1 Attitude measures 250
D.3.2 Arguments used 251

Bibliography **256**

List of Figures

2.1 How confirmation bias can arise at different stages of reasoning 8

4.1 An illustration of the bias-variance tradeoff 123

List of Tables

2.1 Phenomena that have been associated with confirmation bias in the psychological literature . . . 11
2.2 Different types of confirmation bias in search . . . 13
2.3 Different types of confirmation bias in inference . . . 25
2.4 Different types of 'biased beliefs' associated with confirmation bias . . . 36
2.5 Probabilities of observing positive test results given diseases A and B . . . 55
2.6 How strong is each piece of evidence for confirmation bias? . . . 57

3.1 Demographic characteristics, experiment 1 . . . 87
3.2 Bias scores by condition and topic, experiment 1 . . . 89
3.3 Average number of 'pro' articles selected by initial opinion, experiment 1 . . . 90
3.4 Bias scores by topic, experiment 2 . . . 92
3.5 Correlation between initial views and choices, experiment 2 . . . 93
3.6 Bias scores by topic, experiment 3 (Bilendi) . . . 94
3.7 Correlation between initial views and choices, experiment 3 . . . 94
3.8 Bias scores and Pearson's r, experiment 4 . . . 96
3.9 Bias scores by condition, experiment 5 . . . 100
3.10 Pearsons's r by condition, experiment 5 . . . 100
3.11 Demographic characteristics, experiment 6 . . . 103
3.12 Political views, experiment 6 . . . 103
3.13 Bias scores by condition, experiment 6 . . . 104
3.14 Pearsons's r by condition, experiment 6 . . . 104

4.1 Ways reasoning can deviate from Bayes' theorem - adapted from Fischoff and Beyth-Marom (1983) . . . 131
4.2 Different types of rationality . . . 135

5.1 Individual difference variables related to openness . . . 163

A.1 Factors influencing selective exposure . . . 224

B.1 Selective exposure studies of social and political attitudes . . . 226

C.1 Correlation between political ideology and arguments selected, experiment 5 . . . 233
C.2 Correlation between political ideology and arguments selected, experiment 6 . . . 233

Acknowledgements

I've been lucky enough to have three different supervisors who have helped me in different ways over the past few years.

Thank you first to Nick Chater: for encouraging me and making me want to do research from the beginning, and for giving me the freedom to pursue what most interested me.

Jerker Denrell: for keeping me focused these past six months; for helping me to avoid perfectionism, and for helping me to clarify a number of disparate thoughts and ideas.

Gerard Hodgkinson: for challenging and pushing me at the beginning when I was lacking direction, and for giving me a different perspective.

I'm also incredibly grateful to a number of others who have gone above and beyond to support me throughout this process:

First and foremost, Michelle Hutchinson, for practically hand-holding me through the last months with weekly Skype calls, for somehow always knowing the questions I needed to ask myself, for helping me avoid the ubiquitous planning fallacy.

My dad, Kim Whittlestone, for being just about the only person I could actually ask to read through this entire thesis, and for knowing exactly the kind of feedback I needed (and didn't need!)

My fellow PhD students in the Behavioural Science Group at Warwick and beyond, for useful conversations and making me feel I wasnt alone in sometimes finding this impossibly hard: thanks to Marsha Kirichek and Alex Surdina in particular.

Michael Sanders and Simon Day for making me feel welcomed and valued at the Behavioural Insights Team, and showing me what 'applying behavioural science' really looks like in a range of fascinating scenarios. Special thanks to Simon for doing more than anyone else to ensure I got to work on the most interesting projects, and to Michael for making it his mission to help me make progress on my thesis while working a demanding job.

Spencer Greenberg, for helping me believe I could actually do something useful in a PhD thesis (or that I should at least try!), and for always having time to help no matter how ridiculously busy his own life seemed to be.

Suzy Moat and Tobias Preis, for encouraging me when I was at my least confident.

I'm grateful to various friends for helpful questions, interesting ideas, and feedback on drafts - particularly Julia Galef, Peter Hartree, Nick Hare, and Jonah Sinick.

Finally, the friends who have simply been there to support me at various stages over the past few years - particular thanks to:

Nick Robinson, for listening to me in a way no-one else does.

Roman Duda, for being there for me throughout this whole journey, and for often being more interested in what I was doing than I myself felt.

Everyone at the CEA and FHI offices in Oxford for letting me hang around so much, and for being just the right balance of competitive and supportive.

Uri Bram, for encouraging me to write more, for telling me when something was 'good enough', and for helping me to see I was trying to do too much at a crucial point.

My housemates, Alex Foster and Martin Foley, for putting up with me even during my most stressed writing-up periods.

Natalie Cargill, for 'legally' ordering me to take time off when I was feeling burned out.

Finally, thanks to my family, for always being (about the right level of) open-minded, for being consistently supportive in everything I do.

Declaration of Authorship

I, Jess Whittlestone, declare that this thesis titled, *'The importance of making assumptions: why confirmation is not necessarily a bias'* and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- This thesis has not been submitted for a degree at any other University.

In memory of my grandfathers, Peter Dawson and Peter Whittlestone - both who had PhDs themselves, and who I wish could have lived to see me get mine.

Abstract

The idea of a 'confirmation bias' - that people reason in ways that lead them to irrationally confirm whatever it is they already believe - is one of the most widely accepted psychological findings. In this thesis, I argue that the evidence for confirmation bias is much weaker than is often supposed, and that this raises some challenging questions about what it means for beliefs to influence reasoning in an irrational way.

I suggest that the literature on confirmation bias faces three challenges. First, the term 'confirmation bias' has been used to refer to multiple different things by different people, creating a literature of disparate findings that are not well unified. Second, many of the tendencies commonly referred to as confirmation bias are either not robust, or do not lead to confirmation under all circumstances. Third, most findings of 'confirmation bias' do not do enough to demonstrate a genuine bias or irrationality, and do not adequately address the complex associated normative issues.

I discuss the link between confirmation bias and the broader concept of 'open-mindedness', suggesting that existing research on both these topics fails to recognise the necessity and benefits of making assumptions as we navigate our lives. Instead of making claims like "people fall prey to a confirmation bias" or "people should be more open-minded", I suggest that research should focus on understanding how people navigate tradeoffs - between the benefits of having firm beliefs and of making assumptions, and the benefits of being 'detached' from prior beliefs and able to change one's mind.

Chapter 1

Introduction

How should we update our beliefs in light of new evidence? How should our existing beliefs influence how we seek out and interpret new information? To what extent is it helpful to make assumptions when navigating and making sense of the world, and when can doing so constrain and bias us?

These are important questions central to the study of human reasoning and rationality. It is widely accepted that we fall prey to a *confirmation bias*: an irrational tendency to reason in ways that confirm whatever it is we already believe. In this thesis, I will argue that understanding confirmation bias - and answering the questions above - is much more complex than it might first seem. As a result, the claim that "people exhibit a confirmation bias" is not necessarily as robust as has often been supposed.

Confirmation bias has been studied by psychologists in various forms, in both abstract and applied contexts, and is one of the most widely accepted biases in psychology. It is one of the three main biases which Kahneman discusses in *Thinking Fast and Slow*, the book which introduced his Nobel prize-winning work (with Amos Tversky) on cognitive biases to a wider audience (Kahneman, 2011). Many introductory psychology textbooks now cover confirmation bias as one of the few basic flaws in reasoning it is important for students to be aware of (e.g. Plous, 1993, Stangor and Walinga, 2010). In discussing the most common logical fallacies, Risen and Gilovich mention confirmation bias before any other, as "a common and particularly powerful bias." (Risen and Gilovich, 2007, p.112) Nickerson even begins his 1998 review of the subject with the verging-on-hyperbolic claim, "If one were to attempt to identify a single problematic aspect of human reasoning

that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration." (Nickerson, 1998, p.175)

The idea of a confirmation bias also frequently arises in more popular discussion: there is a widespread impression that it is incredibly difficult to change a person's mind once it is made up, because they will simply seek out and interpret information in ways that reinforce what they already believe. A confirmation bias has often been claimed to be at the root of serious real-world problems: including ideological extremism, political polarization and conflict, and many errors of judgement and overconfidence. For example, Justin Wolfers writes in the *New York Times* in 2014: "The problem is that you seek out information that confirms your existing views, a mistake that psychologists call confirmation bias. And your confirmation bias may be the reason that our political debates remain intractable." (Wolfers, 2014)

Maria Konnikova of *The New Yorker* similarly suggests that confirmation bias "can help explain why Trump supporters remain supportive no matter what evidence one puts to them - and why Trump's opponents are unlikely to be convinced of his worth even if he ends up doing something actually positive." (Konnikova, 2016). Lilienfeld et al. suggest that "the bias most pivotal to ideological extremism and inter- and intra-group conflict is confirmation bias." (Lilienfeld et al., 2009, p.391)

Despite (or perhaps because of) being widely studied, the literature on confirmation bias is highly confused. The term 'confirmation bias' has been used to refer to various different tendencies by different people, and it's not clear how these different tendencies relate to one another. Some supposedly well-established effects do not appear to hold up under closer scrutiny (including selective exposure - the tendency to seek out belief-confirming information - as I will discuss later). Other findings - such as those from Wason's seminal work on hypothesis testing - have been interpreted as providing evidence for confirmation bias when they in fact show something subtly but importantly different. Going deeper still, there is a real lack of clarity around what it means to be 'biased' or 'irrational', from which subtle but important disagreements arise. All this casts doubt upon the questions of how, and even whether, we should be trying to 'improve' human reasoning by reducing confirmation bias.

Naively, it seems that letting what we already believe influence our subsequent reasoning

processes is clearly a mistake to be avoided. But things are not quite this straightforward. In the simplest sense, making assumptions and having expectations is crucial and unavoidable: when I get up every morning, I assume that the sun will rise as it did yesterday, that my door will be in the same place it was when I closed it last night, that my breakfast will taste about as delicious as it did yesterday. If I were to start with a totally 'blank slate' every day, my life would be impossibly cognitively demanding. To give a more nuanced example: if I've previously experienced someone as being very untrustworthy, then it seems very reasonable for me to give little weight to their opinions in future. Here, my prior beliefs about this person guide how I interpret any new information I might get from them, but in a way that does not necessarily seem to be irrational. As Kuhn puts it, "nature is vastly too complex to be explored even approximately at random... Something must tell the scientist where to look and what to look for." (Kuhn, 1963, p.363)

People often talk about confirmation bias as if we should simply set aside whatever it is we already believe when considering new information, just 'be unbiased.' Hopefully these examples make it clear that things are not that simple: that my beliefs can often helpfully guide how I think about an otherwise-confusing world, and it's far from possible to approach all new information with a totally 'open mind.' The question of when, and to what extent, it is a 'bias' for what I already believe to influence how I seek out and interpret new information, does not have a straightforward answer.

In this thesis I'll therefore outline some of the issues that the literature on confirmation bias faces, looking at how prior beliefs affect reasoning from a range of perspectives: theoretical, experimental, philosophical, and psychological. My aim is to make two main contributions to existing research in this area. The first is a more narrowly focused contribution to the literature on 'selective exposure': a specific tendency that has generally been considered a type of confirmation bias. I'll report some of my own experiments, alongside a thorough discussion of past research in this area, in order to help explain why findings on selective exposure have been so mixed - and in particular, how to align these mixed findings with the generally pervasive impression that people prefer to read things they agree with. I'll argue that the evidence for selective exposure is even weaker than has been supposed, and that this may be because 'selective exposure' is not actually a particularly good measure of what we're really interested in: the broader phenomenon

of confirmation bias or closed-mindedness. [1]

The second (and perhaps more novel) contribution of this thesis is an integrative one: bringing together different areas of research and different perspectives on similar issues, to try to enhance our understanding of the relevant questions. This integration takes place on multiple different levels:

1. Bringing together the varied and disparate phenomena that have been considered forms of 'confirmation bias'[2];

2. Discussing some of the classic confirmation bias findings in light of the specific criticisms and 'reinterpretations' that have been launched against them in recent decades;

3. Understanding all of this in the context of a wider debate about what it really means to be 'rational' (and why it matters); and finally,

4. Connecting the literature on confirmation bias with discussion of the importance of 'open-mindedness' in both the psychology and philosophy literature - as two areas of research that seem closely connected but are rarely explicitly linked with one another.

In doing so I hope to reorient the study of confirmation bias, so that future research is able to provide better answers to questions like those I posed at the beginning - and perhaps more importantly, to generate better, psychologically-informed solutions to related practical problems. Because despite the numerous issues this thesis discusses, I *do* think we face a challenge as humans within the realms of confirmation bias - that we face some tradeoff between the benefits of making assumptions and believing things with certainty, and the costs of letting those beliefs and assumptions too heavily influence our subsequent thinking. What's important, however, is to acknowledge and carefully

[1] To expand on this slightly, though this is something I will discuss later in much more detail: it's often implicitly assumed that reading arguments one agrees with demonstrates a kind of bias or closed-mindedness. However, simply capturing people's choice of what to read doesn't actually give us enough information to draw this conclusion, because this also depends on someone's motivation in reading a specific argument, and what inferences they draw from it.

[2] This is much needed, given that the most recent review of the literature on confirmation bias was almost two decades ago - Nickerson's 1998 paper, "Confirmation bias: a ubiquitous phenomenon in many guises" Nickerson (1998)

consider what those tradeoffs are and how to navigate them - and to avoid broad, overly-simplistic claims that we should be 'more open-minded', 'reduce confirmation bias' or 'hold beliefs less strongly.'

This thesis is structured in four main sections:

1. **A comprehensive review of the literature on confirmation bias, highlighting several issues that it faces, and arguing that the case for a confirmation bias is much less conclusive than is often assumed.** I argue that the literature faces three main problems. First, the term 'confirmation bias' has been used to refer to multiple different things by different people, creating a literature of disparate findings that are not well unified. Second, many of the tendencies commonly referred to as confirmation bias are either not robust or do not lead to confirmation under all circumstances. Third, normative issues around what it really means to say that confirmation is biased are complex, disputed, and not sufficiently dealt with in the literature.

2. **A closer look at selective exposure, a type of 'confirmation bias' that seems particularly poorly understood and problematic** - the idea that people prefer to read information that confirms what they already believe. I argue, based on the existing literature and several of my own experiments, that the case for selective exposure is weak - and that this has important implications for how we think about and measure confirmation bias more broadly. In particular, I argue that the reason selective exposure effects are so mixed is that the measure of 'selective exposure' does not capture enough of what we're interested in to draw useful conclusions - the extent to which someone displays selective exposure does not actually capture how 'open-minded' or 'rational' a person is, though this conclusion is often implicitly drawn.

3. **An examination of the broader notion of 'rationality' as it has been treated in the psychological literature, and the implications for confirmation bias.** I argue that a lot of disagreement and confusion arises from people using terms like 'biased' and 'irrational' in different ways, and attempt to clarify some of these disagreements by distinguishing some different ways these terms have been used.

4. **A discussion of the concept of 'open-mindedness' as it has been treated in the psychological and philosophical literature.** I challenge the broad claim that people 'should be more open-minded', linking this to earlier discussion of normative issues surrounding confirmation bias. I argue that research on these related issues has not done enough to acknowledge certain tradeoffs that we face as cognitively constrained reasoners with multiple, often competing, goals - resulting in overly-simplistic claims about confirmation bias and open-mindedness.

I will conclude by turning to some more practical questions: how psychological research might continue to advance understanding confirmation bias going forwards, and the implications of our current understanding of confirmation bias for related real-world problems.

Chapter 2

When is confirmation a bias?

2.1 Introduction

The meaning of 'confirmation bias' seems at first glance fairly obvious: a bias towards confirming one's existing beliefs. But what exactly does it mean to confirm existing beliefs, and when is doing so a bias?

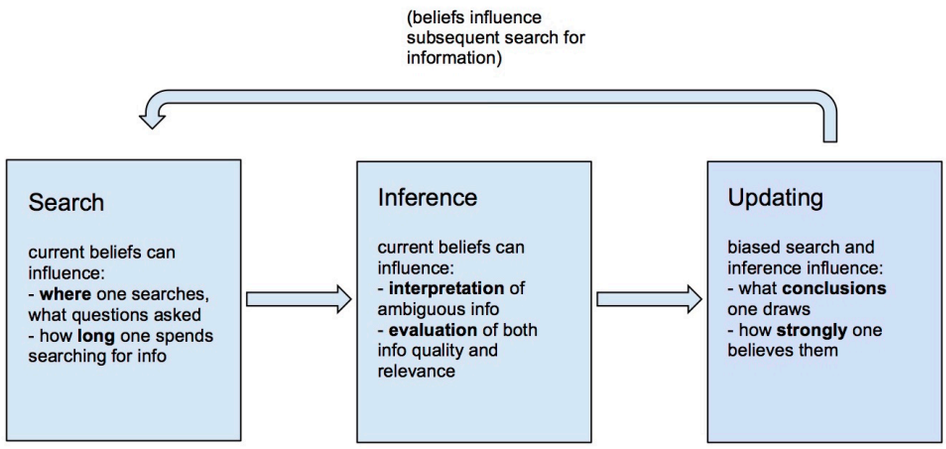
First, it is worth clarifying that there are various different stages of reasoning at which confirmation might occur: including how one searches for information and how one then evaluates or interprets that information. Figure 2.1 shows a simple model of different stages of reasoning, and how biases can arise at each stage. Distinguishing these different stages of reasoning at which confirmation bias can occur makes clear that there are really many different 'types' of confirmation bias. As Nickerson (1998) suggests, confirmation bias should really be considered a broad umbrella term, unifying a number of related biased processes of reasoning.

What unifies these biased processes - what exactly does it mean to 'confirm' a belief? In this thesis, I mean by this any processes that lead to *strengthening confidence* in the current belief. Confirmation bias therefore means reasoning in ways that systematically lead one to strengthen confidence in the favoured hypothesis over the alternative(s), even when doing so is not justified (when the available evidence for alternative hypotheses is equally strong, for example.) It's worth noting here that confirmation isn't inherently irrational - there are plenty of cases where it makes sense for me to strengthen my beliefs in face of the evidence - and so to show a confirmation *bias* exists we need to show that

people reason in ways leading to stronger confirmation than is rational, as judged by some normative standard. This is an issue I will keep coming back to over the course of this thesis.

The term 'confirming' also suggests viewing beliefs in a binary way, that there is a threshold beyond which we 'accept' a given proposition as true. However, most of our beliefs might better be viewed as probabilistic, held with degrees of (un)certainty - and under this framing, it's not clear that any evidence can ever 'confirm' our beliefs. This is why I have suggested we think instead in terms of strengthening confidence in a belief. If we think of beliefs as probabilistic, the issue of what really counts as the 'current' belief is also more complex - since it is not necessarily as simple as saying there are certain things I believe and certain things I don't. Instead of talking about the 'current belief', then, it may be more appropriate to talk about the *favoured* or *focal* hypothesis - i.e. the hypothesis which I think is most probable among available alternatives, and/or which is currently the focus of my attention.

FIGURE 2.1: How confirmation bias can arise at different stages of reasoning



This chapter reviews and discusses the main sources of evidence commonly cited for confirmation bias, with two main aims: (1) to clearly distinguish different types of confirmation bias, and how they relate to one another, and (2) to clarify the extent to which existing research provides evidence for a genuine confirmation bias - reasoning in ways that systematically lead to irrational strengthening confidence in the focal hypothesis. I will do this in two parts: first, reviewing and organising the psychological evidence for

different types of confirmation bias, and second, discussing a number of challenges to whether or not these demonstrate a genuine confirmation bias.

2.2 Different types of confirmation bias: a review

I noted above that confirmation bias occurs whenever one reasons in a way that systematically leads to overconfidence or irrational belief persistence. Unpacking this a little more, when we say a confirmation bias exists, we are making something like the following claims:

1. People tend to reason in ways that lead them to strengthen confidence in/reinforce whatever hypothesis they already favour.

2. In particular, the processes by which people (a) search for new information and test hypotheses, and/or (b) draw inferences from that information, tend to reinforce favoured beliefs more than is rational.

3. As a result, people are more confident in the truth of their beliefs than they should be, and tend to persist in believing the same thing for longer than is reasonable.

4. This tendency is systematic and occurs on average - that is, given a population of people with different prior beliefs who have the same opportunities to obtain information, but some freedom over how they search for information and how they draw inferences from that information - confirmation bias means that people will, on average, tend to strengthen their confidence in whatever it was they originally believed.

Statement (1) broadly summarises what we mean by confirmation bias. (2) and (3) elaborate on this a little - stating that this bias is the result of biased processes of search and inference, and results in certain biases in the beliefs people hold. (4) elaborates on what is required for this to really constitute a 'bias' - not merely occasional tendencies towards confirmation, but ways of reasoning that hold systematically, across people and situations.

The focus of this section is on the evidence for claims (2) and (3) - that people search for and draw inferences from information in biased ways, and that this results in overconfidence and belief persistence. Though in this section I will briefly mention some problems that different types of confirmation bias might face, a more detailed discussion of these issues will follow in the next section, focusing on claim (4) - whether a confirmation bias is systematic and occurs on average.

This review is split into three parts: bias in search, bias in inference, and the resulting biased beliefs, summarised in table 2.1.

As I will discuss later in this thesis, there seem to be multiple interpretations of what it means to attribute bias, and failing to distinguish between these different meanings can create confusion and disagreement. Particularly in this chapter, whenever I ask whether something actually provides evidence for a confirmation bias, I will try to spell out what precisely I mean by this. In general, when I talk about whether something constitutes a genuine confirmation bias, I mean this as outlined in points 1-4 above: is there evidence for a systematic tendency to reason in ways that lead to confirming one's prior beliefs more than is rational? As we will see, this is a much more complex claim to defend than it might seem, and certainly harder than showing that people's prior beliefs sometimes - or even frequently - influence subsequent reasoning.

Bias in search	Bias in inference	Biased beliefs
Seeking out information expected to confirm the current hypothesis • **Bias in hypothesis testing**: asking questions whose answers skew in favour of evidence for focal hypothesis (Snyder and Swann, 1978, Wason, 1960, 1968). • **Selective exposure**: looking for information in places expected to support current hypothesis (Hart et al., 2009) • **Myside bias in argument production**: selectively searching memory for confirmatory info (Toplak and Stanovich, 2003)	Interpreting ambiguous evidence as supportive of currently favoured hypothesis • **Pseudo-diagnosticity**: failing to choose information needed to compute likelihood ratios (Fischoff and Beyth-Marom, 1983) • Interpretation of pseudodiagnostic evidence (Feeney et al., 2000)	Failing to update/change ones beliefs when one should • **Persistence in the debriefing paradigm**: continuing to believe something even when original evidence discredited (Ross et al., 1975) • **Persistence of misinformation**/false beliefs (Lewandowsky et al., 2012) • **Conservatism bias**: updating beliefs conservatively with respect to Bayes rule (Edwards, 1965)
Stopping search for information when evidence points in favour of current/favoured hypothesis, continuing search for information when it does not (Ditto and Lopez, 1992)	Applying different standards of scrutiny/evaluation to supporting and conflicting evidence • **Overweighting confirmatory evidence** and/or underweighting disconfirmatory evidence (Pyszczynski and Greenberg, 1987) • Rating pro-attitudinal attitudes as stronger than counter-attitudinal ones (Taber and Lodge, 2006) • **Biased assimilation**: drawing undue support from mixed findings (Lord et al., 1979)	**Overconfidence**: holding beliefs more strongly or with more precision than is rational • **Overprecision** in judgement (Moore et al., 2015) - in the 2AFC paradigm (Griffin and Brenner, 2004), and the confidence-interval paradigm (Bazerman and Moore, 2013)

Table 2.1: Phenomena that have been associated with confirmation bias in the psychological literature

2.2.1 Bias in search

Before we even get to evaluating, interpreting, and drawing conclusions from information, we have to obtain information in the first place. If we aren't even exposed to information that might expose our views to be wrong, then we have no hope of ever changing our minds about anything. Understanding potential biases in the *search* for information is therefore a crucial part of understanding confirmation bias more generally. There are various different ways we seek out information related to things we believe - some much more active, like explicitly asking questions and trying to test hypotheses, and some much more passive, such as choosing which links or article headlines to actually click on and read. There are, accordingly, various different ways our search for information might be biased towards confirming or reinforcing what we already believe.

Here I discuss the evidence for several different ways bias might arise in the search for information: in what hypotheses we choose to test, and in where, how, and to what extent we search for new information. There are two main areas of research which are generally cited as examples of confirmation bias in search: the literature on bias in hypothesis testing, and the selective exposure literature. I also discuss two other areas of research which are more indirectly related to bias in search: whether people spend more time searching for information if it does not initially confirm their beliefs, and whether people find it easier to produce confirming arguments (which might indicate a bias in 'searching' memory.) Table 2.2 below summarises the different kinds of 'confirmation bias' in search reviewed.

Name/description	Key references	Challenges
Bias in hypothesis testing: asking questions/testing hypotheses in a way more likely to yield positive responses	Wason's (1960, 1968) rule-discovery task and selection task, finding people use a positive test strategy and fail to seek potentially falsifying information. Social hypothesis testing, with particular interest in stereotypes. Snyder and Campbell (1980), Snyder and Swann (1978) find people significantly more likely to ask questions expected to confirm the hypothesis given.	A positive test strategy does not lead to confirmation under all circumstances, so is not necessarily a confirmation bias (Klayman, 1995, Klayman and Ha, 1987, Oaksford and Chater, 1994) People are more likely to seek falsifying information in more familiar environments (Johnson-Laird et al., 1972, Wason and Shapiro, 1971) Later studies of social hypothesis testing find people do not always ask confirmatory questions, depending on the framing of the task or type of hypothesis (Snyder and White, 1981, Trope and Bassok, 1983, Trope and Mackie, 1987)
Selective exposure: seeking out information expected to confirm, rather than disconfirm, existing beliefs	Key reviews of the literature that support a selective exposure effect include Cotton (1985), Frey (1986), and Hart et al. (2009).	A review by Freedman (1965) challenges that selective exposure is a robust phenomenon, citing studies finding an opposite effect. Selective exposure measure does not necessarily capture bias/irrationality, since this depends on ones motivation and how one interprets info.
Bias in amount of time spent searching: deciding when to terminate search process based on how much support one currently has for position.	Mostly in the motivated reasoning literature - e.g. Ditto and Lopez (1992) find people spend more time questioning the results of a test and are more likely to re-take the test if it indicates a negative diagnosis.	Requires a leap from evidence for motivated reasoning - people are biased towards what they want to believe - to confirmation bias - bias towards current beliefs.
Bias in memory search: selectively remembering/searching memory for confirmatory information	Evidence of 'myside bias' in the production of arguments - i.e. finding it easier to produce arguments confirming ones position (Perkins et al., 1986, Toplak and Stanovich, 2003)	Whether or not this is a 'bias' depends what people think the purpose of the argument-production task is, which is often not clear. Wolfe and Britt (2008) find people find it easier to generate supportive arguments for a given position, regardless of whether they themselves believe it.

TABLE 2.2: Different types of confirmation bias in search

2.2.1.1 Bias in hypothesis testing - Wason's original experiments

A series of experiments by Wason (most notably Wason, 1960, 1968) looked at how people test hypotheses in abstract rule-discovery tasks. The overall conclusion from these experiments is that people exhibit a 'positive test strategy' - that is, tending to ask questions and test hypotheses in ways that are likely to yield positive evidence for the current hypothesis, rather than evidence that might support alternative hypotheses. This 'positive test strategy' has often been cited as one of the simplest forms of confirmation bias.

Evidence for a positive test strategy comes from two simple experimental paradigms. In the first, people are asked to determine a rule governing triplets of numbers by testing different possibilities, initially just being told that the triplet 2-4-6 fits the rule (Wason, 1960). People seem to reason by forming a hypothesis about what the rule is on the basis of this early information - for example, that the rule is numbers ascending by two each time - and then testing examples that fit this hypothesis (8 10 12; 100 102 104; 55 57 59, and so on). This strategy leads to an apparent confirmation bias because the test yields more and more evidence in favour of the hypothesis, but does not allow people to see evidence that might falsify it (triplets that do not fit the hypothesis but do conform to the rule). The rule in Wason's experiment, as it turns out, was much more general than many participants hypothesized - simply that triplets must be ascending, positive numbers - something which most participants failed to discover due to using a positive test strategy.

The second experimental paradigm Wason used is known as the 'selection task' (Wason, 1968), which looks at which tests people believe they need to conduct in order to determine whether a rule is true or not. Participants see four cards, labelled A, B, 2 and 5 - and are told that each card has a letter on one side and a number on the other. Their task is to determine whether the following rule is true: "if a card has a vowel on one side, then it has an even number on the other side", and are asked which cards need to be turned over in order to test this rule. Here also people seem to fall prey to a kind of 'confirmation bias' - choosing to turn over those cards that would confirm the hypothesis if true, but not those that would disconfirm it if false. That is, most people say that two cards need to be turned over - the A card, and the 2 card, to see if they have an even number and a vowel on the other side respectively. But in fact, turning

over the 2 card is less helpful than turning over the 5 card. If the 2 card has a vowel on the other side, this supports the rule - but if it has a consonant on the other side, this does not provide any information either way - the rule says that any card with a vowel on one side must have an even number on the other side, but *not* that any card with an even number on one side must have a vowel on the other side. Regardless of what is on the other side of the 2 card, the rule could still be true - so turning it over is not all that informative. By contrast, if there is a vowel on the other side of the 5 card, this disconfirms the rule - and so it is much more useful to turn this card over. The fact that many people fail to see this intuitively has been interpreted as evidence that we find it easier to seek evidence that confirms, rather than disconfirms, our existing beliefs.

However, as I will discuss towards the end of this section, whether or not a 'positive test strategy' actually constitutes a confirmation bias has been contested (Klayman, 1995, Klayman and Ha, 1987, Oaksford and Chater, 1994). Though it might seem like asking questions expected to yield a 'positive' answer is the same thing as seeking to confirm one's hypothesis, the two are in fact subtly different, and a positive test strategy can yield disconfirming information in certain circumstances. It has been argued (Oaksford and Chater, 1994) that the 'falsification' strategy Wason believes to be the correct response, is not necessarily the best way for people to maximise information gain in more realistic contexts - as I will explain in more detail later.

2.2.1.2 Subsequent research on hypothesis testing

A body of subsequent research has built on Wason's basic paradigm for studying bias in hypothesis testing, specifically looking at how these results generalise to less abstract tasks.

Some studies suggest that people are more likely to select disconfirming evidence if the selection task is presented in a more familiar context: asking people to test a hypothesis like, "every time I go to Manchester I travel by car", for example, rather than "every card which has a D on one side has a 3 on the other side" (Wason and Shapiro, 1971). Wason and Shapiro find that when asked to test the former rule, 10 of 16 subjects (62.5%) choose the correct cards to turn over to test the rule, whereas when asked to test the latter, more abstract, rule, only 2 of 16 (12.5%) choose the correct answer. (Though the

sample sizes are small, these differences seem substantial, and are statistically significant at $p < 0.01$).

Johnson-Laird et al. (1972) similarly find that when asked to test a rule concerning a "realistic relation" ("if a letter is sealed, then it has a 50 lire stamp on it") people find it easier to select the envelope needed to turn over to test the rule, than when asked to test a rule concerning "an arbitrary relation between symbols" ("if a letter has an A on one side, then it has a 3 on the other side.") They found that in the 'realistic' condition, 17 of 24 (71%) selected the correct answer over two different trials, whereas in the 'symbolic' condition, not a single person did. Johnson-Laird et al. (1972) discuss these findings, suggesting that it is the realistic relation between the contingencies in the rules which makes the task easier for participants: they are used to thinking about relationships between things such as destinations and modes of transport, or postal rates and envelopes, and so find it easier to consider different hypothetical possibilities.

A later wave of research also looked at how people test hypotheses in social contexts, with a particular interest in stereotypes, and whether people look for information to confirm whatever assumptions they already have about other people. Early research on 'lay interviewing' (where the participant's task is to formulate questions that will help them to decide whether the interviewee belongs to some trait category) found that people use something similar to a positive-test strategy (Snyder and Campbell, 1980, Snyder and Swann, 1978). Snyder and Swann randomly assigned participants to test the hypothesis that a person was either extroverted or introverted, and gave them the relevant personality profile to read beforehand. Participants were then asked to test their hypothesis by selecting 12 from a list of 26 questions asking about a wide range of beliefs, feelings, and actions. These questions had previously been categorised as either extroverted questions (questions which are likely to be answered in a way that would confirm the 'extrovert' hypothesis, e.g. "In what situations are you most talkative?"), introverted questions (e.g. "What things do you dislike about loud parties?") and neutral questions (e.g. "What are your career goals?") They found that participants were significantly more likely to ask questions expected to confirm the hypothesis they had been given. Snyder and Campbell found a similar result, with those assigned the 'extrovert' hypothesis choosing an average of 7 extroverted questions and 2.67 introverted questions, while those given the 'introvert' hypothesis chose an average of 4.87 extroverted questions and 5.33 introverted questions.

However, these findings were challenged by later studies of social hypothesis-testing, which suggested people do not always ask such 'confirmatory' questions (Snyder and White, 1981, Trope and Bassok, 1983, Trope and Mackie, 1987). Snyder and White found that when a task was explicitly framed as a falsification task - when people were asked to determine the extent to which a target did *not* fit a described personality type - people did ask questions which sought out disconfirming evidence. Snyder and White suggested that the problem is not that people are unable to seek disconfirming evidence - but rather that, as tasks are classically framed, people are focused on building a case, rather than testing a hypothesis.

Trope and Bassok (1983) found that subjects displayed a preference for hypothesis-confirming questions when the hypothesis was extreme (e.g. when they were asked to test whether a person is extremely polite) but not when it was less extreme (e.g. when asked to test whether the person is somewhat polite.) They also found that when some questions are clearly more diagnostic than others (i.e. some questions better distinguish between two hypotheses), diagnosticity is the most influential factor in how people choose what questions to ask. While people did choose more hypothesis-confirming questions, when consistency and diagnosticity conflicted, the latter won out. Trope and Bassok suggest that a preference for hypothesis-consistent information when testing extreme hypotheses may actually be explained in terms of diagnosticity: when discriminating at more extreme points of a distribution, hypothesis-confirming questions are perceived as more diagnostic. In a followup experiment, they have subjects rate the diagnosticity of different questions and find that when testing extreme hypotheses, subjects do indeed consider hypothesis-confirming questions more diagnostic.

2.2.1.3 Selective exposure

While the literature on hypothesis-testing looks at how people form hypotheses and test them with respect to abstract, simple, rules, the literature on selective exposure asks similar questions but with more complex beliefs. There are advantages and disadvantages to both approaches - abstract experiments make it easier to control for various factors and to identify general tendencies, but we have to make a bigger leap to draw conclusions about how people reason about the kinds of things they believe in less formalized hypothesis-testing scenarios. The selective exposure literature , by contrast,

gives more insights into how people seek out information related to things they actually believe, but the results are harder to neatly interpret because experiments are less easily controlled.

Selective exposure is the tendency to selectively seek out information that supports rather than conflicts with one's prior views. In the classic selective exposure paradigm, participants answer questions about their attitudes on a specific topic, and are then given the opportunity to choose to read different sources of information on that topic - some which they expect to support their position, and some expected to conflict with it (see Fischer et al., 2005). 'Selective exposure' is then measured as the average difference between the number of supporting and conflicting information sources chosen - with higher positive values indicating greater selectivity. However, selective exposure has been defined and measured in different ways in different studies: some observing people's actual choices/behaviour, and others instead asking people to indicate, hypothetically, how interested they would be in reading different sources of information on a scale.

In one of the earliest selective exposure studies (Adams and Stacy, 1961), for example, mothers were asked whether they believed child development was predominantly influenced by genetic or environmental factors. When then given the choice to hear a speech advocating either position, mothers overwhelmingly chose the speech that favoured their view on the issue. However, Sears et al. (1967) cast doubt upon the selective exposure hypothesis. They reported a number of experiments which failed to find a selective exposure effect and even some whose results actively opposed it, finding people chose more conflicting than supporting information. In one experiment, participants were presented with a tape of a mock interview portraying an interviewee in either a positive or a negative light (Freedman, 1965). After judging the interviewees themselves, a majority of participants preferred to read another evaluation that *disagreed* with their judgement.

In the mid-1980s, two new review papers were published (Cotton, 1985, Frey, 1986), both supporting the existence of a selective exposure effect. Cotton (1985) concludes that "dissonance-motivated selective exposure does appear to exist... Although the phenomenon has often been elusive and its support questioned, the research overall suggests that something is there", while Frey (1986) suggests, "shortcomings in experimental designs of previous research have largely been responsible for the lack of conclusive results in the earlier studies." Over twenty years later, in their meta-analysis of 91 studies across

67 papers, Hart et al. (2009) find "a moderate preference for congenial over uncongenial information (d=0.36)."

What these reviews - and subsequent attempts to make sense of the selective exposure literature - all acknowledge, is that confirmatory information search is by no means universal, and depends on a number of moderating factors. Subsequent research has therefore focused more on understanding these moderating factors. For example, Jonas et al. (2003) found that increasing mortality salience increased the tendency towards confirmatory information processing, as did increasing the relevance of the issue to the person's worldview. Fischer et al. (2005) found that restricting the opportunity people had to select information increased selectivity. In a meta-analysis, Hart et al. found that selective exposure was more likely to occur when challenging information was expected to be high quality, when prior attitude commitment was higher, and for individuals who score high on the personality trait of closed-mindedness (Hart et al., 2009).

2.2.1.4 Bias in amount of time spent searching

As well as controlling where and how we search for information, we can choose how *long we spend* looking for new evidence related to a given issue. Another way that confirmation bias might manifest in search, therefore, is if people spend differing amounts of time seeking out new information depending on how much support they presently have for their current belief. Confirmation bias would predict that people will spend more time searching for information if the initial search is not favourable towards their current belief, than if they quickly obtain supportive information.

Though little research has looked at this directly, there is some more indirect evidence that people apply different standards of evaluation to information that confirms than disconfirms their current beliefs (as I will discuss in more detail in the inference section.) This then suggests that people will spend more time searching for information if it suggests an opinion-inconsistent conclusion than if it initially supports an opinion-consistent conclusion. Ditto and Lopez (1992) find that subjects spend more time questioning the results of a test, and are more likely to re-take the test to gain more information, if they believe the test indicates a negative diagnosis. This also highlights the fact the bias in search and bias in inference are more closely related than is sometimes acknowledged - since the inferences we draw affect our beliefs, which in turn influence our motivation

and assumptions in seeking out new information (as indicated by the arrow from the end back to the start of the process in figure 2.1, and as I will discuss in more detail in 2.2.3).

To see why this might result in a confirmation bias, imagine that of all scientific studies conducted on gun control, there are in fact only three supporting the claim that gun control reduces homicides, ten that find no evidence of such a relationship, and five that find evidence of the opposite relationship. If I already believe that gun control is effective, the evidence I end up being exposed to - and therefore whether I change my mind or not - may end up being highly contingent on whatever I happen to read first. If I read a pro gun-control study first, I may be so convinced of the conclusion that I decide to stop there - whereas if I read one of the studies suggesting the opposite relationship first, my confusion may lead me to continue reading more, meaning I ultimately sample more of the available information and am better informed about the topic.

The evidence for this tendency to spend more time seeking out information if the initial results are not confirmatory comes primarily from the literature on motivated reasoning (Kunda, 1990), which subtly differs from confirmation bias. The claim of motivated reasoning is that people will reason in ways biased towards whatever they want to believe, (as opposed to confirmation bias which claims a bias towards whatever people currently believe.) Is it reasonable to expect this tendency would generalise to confirmation bias? Assuming that people want to continue believing whatever it is they currently believe - a fairly reasonable-seeming assumption - we might then expect findings from motivated reasoning to generalise to confirmation bias. However, it's worth recognising that there is a leap being made here, and the evidence does not directly support the claim for the case of confirmation bias.

2.2.1.5 'Myside bias' in producing arguments

A final way in which confirmation bias has been suggested to occur in the search for information is in the *production* of arguments: people supposedly find it easier to generate arguments for their preferred position than for the alternative. This seems appropriately categorised as 'bias in search' since we might think of producing arguments as searching memory for reasons - an asymmetry in the arguments one produces therefore might be thought of as biased memory search.

Toplak and Stanovich (2003) asked undergraduate participants to generate arguments on both sides of three issues, and found that people consistently generated more 'myside' arguments than other side arguments. Perkins et al. (1986) found a similar effect. However, Wolfe and Britt point out that "with respect to written argumentation, it is not at all clear that the tendency to generate more arguments for myside is an adequate definition of the myside bias, or even if it constitutes a bias of any kind." (Wolfe and Britt, 2008, p.3) Whether or not this can reasonably be considered a bias presumably depends on what people believe the purpose of generating arguments is: if trying to put a case forwards for one side is even a plausible interpretation of the task, then an asymmetry in arguments presented may be a very reasonable response. To understand bias in the generation of arguments, it's helpful to distinguish different settings: generating arguments in order to come to a conclusion about an important question is very different from doing so in the context of defending a specific position, for example.

Wolfe and Britt (2008) define myside bias in written argumentation more strictly, as a failure to make *any* reference to other side arguments or positions. In two experiments, they find that many participants only discuss one side of a given issue when asked to put forward an argument, but that this arises *independently* of their personal opinions. If asked to write an essay in favour of a proposal, many participants fail to discuss any arguments against the proposal, regardless of what their personal beliefs are on the issue. The authors suggest that the 'myside bias' is therefore not rooted in a 'personal opinion bias' or confirmation bias - but rather stems from misperceptions about the nature of argumentation. Questionable beliefs about thinking (Baron, 1991, 1995) - such as the impression one-sided arguments are stronger than two-sided arguments - may better account for what is going on here than a confirmation bias *per se*.

2.2.1.6 Bias in search: discussion

Two main strands of research are generally considered evidence of a confirmation bias in how people search for information: evidence for a 'positive test strategy' in hypothesis testing, and evidence for selective exposure. Research on motivated reasoning also provides somewhat more indirect evidence that the amount of time people spend searching for information is affected by the amount of support there is for the current hypothesis so

far, and there is also some evidence that people find it it easier to produce confirmatory arguments than opposing ones, suggesting a bias in memory search.

However, there are a few reasons we might hesitate to draw the conclusion that search strategies exhibit a clear confirmation bias, which we mention briefly here before discussing in more detail in the next section. Klayman (1995) points out that a positive test strategy does not always lead to confirming the focal hypothesis - it can enable one to discover false positives (but not false negatives), so whether a positive test strategy leads to confirmation bias depends on the nature of the hypothesis. Oaksford and Chater (1994) argue that under certain conditions, a positive test strategy may be the optimal way to maximise information gain. Though a positive test strategy certainly looks 'confirmatory' in nature, whether or not it actually leads to confirmation, and whether that is non-normative, depends on the features of the situation.

Second, it's not clear that a positive test strategy is systematic in the sense of occurring, on average, across a wide range of scenarios. In particular, it doesn't seem to extrapolate to less-abstract scenarios. When the rule-discovery task that involves choosing which cards to turn over is made less abstract - when the rule subjects are testing is something like, "People with brown hair also have brown eyes", people select the correct cards much more frequently than when the rule is abstract, e.g. "Cards with a vowel on one side have an even number on the other side." (Johnson-Laird et al., 1972, Wason and Shapiro, 1971).

Similarly, research on hypothesis-testing strategies in social perception initially suggested that people seek evidence in confirmatory ways (Snyder and Campbell, 1980, Snyder and Swann, 1978), but whether they actually demonstrate irrational confirmation is not clear. Later studies suggest that the questions people ask may be more determined by the diagnosticity of questions (i.e. how well they distinguish between alternative hypotheses) than has been thought - and what look like confirmatory strategies may sometimes be attempts to estimate diagnosticity under uncertainty. Skov and Sherman (1986) also argue that these studies fail to distinguish between two different types of hypothesis confirmation: (1) seeking information that is primarily *relevant* to the hypothesis under consideration, and (2) seeking information that *will actually make it seem more likely that the hypothesis is true* than it should. They argue that Snyder and Swann (1978) and Snyder and Campbell (1980) talk about hypothesis-confirmation in the second sense, but

in fact do not distinguish between the two senses - and the former type of confirmation is not necessarily suboptimal, nor does it necessarily lead to strengthening the focal hypothesis.

Overall, it's far from clear that people systematically test hypotheses in ways that lead them to strengthen confidence in those hypotheses more than they should. Whether or not people use apparently confirming search strategies seems to depend on various factors: including the abstractness of the task, the extremity of the hypothesis being tested, and whether or not information about diagnosticity is available. Furthermore, most of these studies fail to establish that testing strategies are genuinely confirmatory, in the sense of generating answers that disproportionately support the focal hypothesis.

The selective exposure literature also faces its own challenges. The evidence for selective exposure has been much more mixed than is often appreciated - with many studies failing to find a selective exposure effect, and some even finding the opposite (that people actively seek out more information that challenges their beliefs) (Sears et al., 1967). One interpretation of these mixed selective exposure findings is that selective exposure, like a positive test strategy, does not actually quite map onto a confirmation bias. Whether seeking out more confirmatory than disconfirmatory evidence leads to systematic overconfidence in the focal hypothesis depends in turn with what one does with that information - seeking out supportive information could lead one to reduce confidence in the focal hypothesis if one scrutinises that information carefully, for example. This suggests the importance of understanding not just bias in the search for information but also bias in the inferences people draw from that information, which we will cover in the next section.

We also briefly covered two other phenomena seemingly related to confirmation bias in search: bias in the amount of time spent searching, and bias in the production of arguments by search in memory. There is some limited evidence for the former (Ditto and Lopez, 1992), but it suggests at best that decisions to terminate search processes are sometimes influenced by whether or not one *likes* the conclusion - which is subtly different from a confirmation bias. Evidence of myside bias in the production of arguments suggests that people may find it easier to produce arguments that are supportive of their position (Perkins et al., 1986, Toplak and Stanovich, 2003) - but whether or not this is considered a bias depends on what people believe the purpose of the task is, which

is often not clear. Furthermore, Wolfe and Britt (2008) find that people find it easier to generate arguments that are supportive of any given position, regardless of whether they themselves believe it or not - suggesting this tendency may lie more in beliefs about argumentation, or greater ease thinking about 'positive' information generally, than a genuine bias towards confirming one's beliefs.

Overall, none of this evidence for confirmation bias in search seems to withstand scrutiny, and faces problems on multiple levels: both with the robustness of the results and whether these tendencies actually lead to a confirmation bias. I will now consider the evidence for bias in the inferences people draw from information, to see if this is any more convincing.

2.2.2 Bias in inference

In addition to seeking out information or testing hypotheses in ways that seem biased towards confirming one's beliefs, we can also be biased in how we draw inferences from whatever evidence we happen to come across. I discuss two main tendencies that seem to contribute to confirmation bias in inference: interpreting ambiguous information as more supportive than it actually is; and applying different standards of evaluation to supportive and conflicting information (summarised below in table 2.3).

Name/description	Key references	Challenges
Biased interpretation of evidence: interpreting information as more diagnostic of ones hypothesis than it in fact is (sometimes called pseudodiagnosticity)	Fischoff and Beyth-Marom (1983) - who argue that a lot of whats called confirmation bias can be interpreted in terms of pseudodiagnosticity Classic studies of pseudodiagnosticity that find people seek out diagnostically worthless information and then alter their beliefs based on it (Doherty et al., 1979, Kern and Doherty, 1982)	In some scenarios, people do seem better at judging diagnosticity. Mynatt et al. (1993), for example find people are more likely to correctly choose diagnostic information when the focal hypothesis is unlikely. Feeney et al. (2000) find people over-interpret the diagnosticity of information when features are rare, and that this may sometimes be adaptive. Nelson (2005) argues that maximising diagnosticity is sometimes inferior to other sampling norms. Crupi et al. (2009) also argue that normative standards typically used in pseudodiagnosticity experiments are flawed. There seems to be some confusion between pseudodiagnosticity as interpreting information as more diagnostic than it is, and as seeking out non-diagnostic information - and more evidence for the latter (so a bias in search) than the former.
Biased evaluation of evidence: applying harsher standards to conflicting than supporting evidence	Pyszczynski and Greenberg (1987) - people tend to overweight confirmatory positive evidence and/or underweight negative disconfirmatory evidence Lord et al. (1979) - people accept confirming evidence at face value but scrutinise disconfirming evidence, resulting in biased assimilation - drawing support for ones position from mixed/neutral findings. Several subsequent studies find similar results (Liberman and Chaiken, 1992, Taber et al., 2009, Taber and Lodge, 2006)	Kuhn and Lao (1996) fail to replicate Lord et als polarization results, suggesting biased assimilation/polarization not as robust as it first seems. Its unclear to what extent applying different levels of scrutiny to supporting/conflicting evidence is irrational - Lord et al. (1979) acknowledge that sometimes this may be adaptive or unavoidable.

Table 2.3: Different types of confirmation bias in inference

2.2.2.1 Biased interpretation of evidence

Fischoff and Beyth-Marom (1983) argue that a great deal of confirmation bias can be explained in terms of people misunderstanding how *diagnostic* information is. In particular, they argue, people interpret information as providing stronger evidence for their current hypothesis than it actually does, a tendency sometimes referred to as *pseudodiagnosticity.* For example, suppose I am trying to determine which of several different diseases a patient has. I have a hunch that the correct diagnosis is disease A, and so test to see if the patient has any of the symptoms listed for this disease: including a cough, temperature, and dizziness. When I find the patient does have one of these symptoms - the cough - I increase my confidence that she has disease A. The problem here is that a cough might also be a symptom of a wide range of other diseases, and so this information is not particularly diagnostic: it does not necessarily give me good reason to update in favour of my initial guess that the patient had disease A.

Doherty et al. (1979) find that subjects display a surprising and strong tendency to seek diagnostically worthless information - and to then alter their beliefs on the basis of that information. One plausible explanation for pseudodiagnosticity is that people simply fail to consider alternative hypotheses - asking, "is this evidence consistent with what I believe?" but not, "is this evidence equally consistent with some other explanation?" If I only consider whether evidence supports my current hypothesis, and not whether it might also support alternative hypotheses, I am likely to end up interpreting evidence as more supportive than it actually is.

However, it's worth highlighting that there are two subtly different ways that the claim of 'pseudodiagnosticity' might be interpreted:

(P1) When presented with data/information supportive of a hypothesis, people tend to interpret it as more diagnostic of that hypothesis than it in fact is.

(P2) When given the choice between different types of information related to a hypothesis, people often fail to choose the most diagnostic information.

(P1) essentially amounts to saying that people overestimate $\Pr(H \mid D)$ - the strength of their hypothesis, given the data D - because they misinterpret D as being more diagnostic of H than it fact is (more precisely, they overestimate the likelihood ratio $\Pr(D \mid H)/\Pr(D \mid \neg H)$. (P2), by contrast, says that if people already know $Pr(D_1 \mid H)$,

they will often choose to learn about $\Pr(D_2 \mid H)$ (the likelihood of some other data D_2 under the same hypothesis) even though learning about the same data under the alternative hypothesis - $\Pr(D_1 \mid \neg H)$ would be more diagnostic. That is, people choose to obtain information that's more obviously relevant to the focal hypothesis, and fail to appreciate that considering an alternative hypothesis would be more useful. Of course, these two tendencies are closely related: the choice of $\Pr(D_2 \mid H)$ seems likely to be based on mistakenly judging it as more diagnostic than it is. And if I only seek out information that is likely under my favored hypothesis (and not whether the same information is also likely under alternative hypotheses), then I'm likely to overestimate how diagnostic that information is, and therefore strengthen my belief in my hypothesis more than I should. This highlights again how inseparable bias in search and inference are - our choices of information to seek out depends on what kinds of inferences we think we'll end up drawing from that information, and the inferences we draw depend on how that information was obtained in the first place.

While research on pseudodiagnosticity is often cited as evidence for (P1), most studies in fact instead demonstrate something closer to (P2), using the following general paradigm (Crupi et al., 2009). Participants are presented with two mutually exclusive, and jointly exhaustive hypotheses - H and $\neg H$ (e.g. the patient has disease A, or the patient does not have disease A.) They are told that there are two types of data available - D_1 and D_2 (information on different symptoms, say) - and that it's possible to represent the probabilistic relationships between the available data and the hypotheses - the likelihood of observing each piece of data under each of the hypotheses: $\Pr(D_1 \mid H)$, $\Pr(D_1 \mid \neg H)$, $\Pr(D_2 \mid H)$, $\Pr(D_2 \mid \neg H)$. Participants are given one such relationship - told the likelihood of observing one of the symptoms if the disease is present: $\Pr(D_1 \mid H)$ - and told they can ask for one more (the likelihood of a second symptom if the patient has or does not have the disease, the likelihood of the first symptom if the patient does not have the disease). Studies find that subjects frequently ask to learn $\Pr(D_2 \mid H)$ - the likelihood of observing a second symptom if the patient has the disease - even though asking for $\Pr(D_1 \mid \neg H)$ - the likelihood of observing the known symptom if the patient does not have the disease - is in fact more highly diagnostic (Doherty et al., 1979, Kern and Doherty, 1982).

The evidence for pseudodiagnosticity in this sense is still more nuanced than it first seems. Mynatt et al. (1993) find that when the original value people are given for

($\Pr(D_1 \mid H)$) is low (the value they use is 0.35), people are more likely to ask to learn about $\Pr(D_1 \mid \neg H)$ - i.e. to ask to learn about the alternative hypothesis. If the probability given for H is low, they suggest, this shifts focus to the alternative hypothesis, as the most likely hypothesis. Mynatt et al. (1993) explain classic pseudodiagnosticity findings (where people choose to find out more about the focal hypothesis, rather than learning about an alternative which would be more diagnostic), as biasing processes of focalization - people essentially want to find out more about whichever they think is the most likely hypothesis. The authors go so far as to suggest that people can only think about one hypothesis at a time - but that, if a low initial value is given for $\Pr(D_1 \mid H)$, this shifts the attention onto $\neg H$. They also find that people are considerably more likely to seek alternative information ($\Pr(D_1 \mid \neg H)$ rather than $\Pr(D_2 \mid H)$) when the problem is an action problem - when people have to decide between different courses of action based on information - as opposed to a pure inference problem. This may be because for action problems (but not inference problems) it is easier to see why seeking out alternative information might actually be useful.

Feeney et al. (2008) argue that the usefulness of seeking out 'alternative' evidence and computing the likelihood ratio depends on how common or rare the features are that one is getting evidence about. They consider a scenario where you are trying to decide whether your sister's car is an X or a Y, and you have been told two pieces of information about it - its speed (D_1), and whether or not it possesses a radio (D_2). You are told how many X cars can go at the same speed: $\Pr(D_1 \mid H_X)$, and can learn about the speed of Y cars: $\Pr(D_1 \mid H_Y)$, or radios in X or Y cars: $\Pr(D_2 \mid H_x)$. If you're told that the car can go above 60mph, for example, and you have background information that most cars can drive above this speed (the feature of going above 60mph is a common one), then asking about the proportion of Y cars that drive at this speed might not be that informative - you would likely learn more by asking about the car radio. By contrast, if you're told that the car can go above 150mph (a more rare feature), and told what proportion of X cars can reach this speed, then asking about the speed of Y cars is likely to be much more informative. In two experiments, Feeney et al. (2008) show that participants are more likely to ask for information about the alternative hypothesis when the initial information concerns features that are rare - but that this effect only holds for familiar materials (where participants presumably have the background information needed to recognise the rarity of features.)

Feeney et al. (2000) also find that people's background beliefs about the rarity of features in the environment affects their interpretation of 'pseudodiagnostic' evidence: that people update their beliefs more when evidence concerns a feature they believe to be rare. They argue that, though from a strictly normative point of view people should not update their beliefs, it may often be adaptive to use one's background information to interpret incomplete information.

The general point here is that if one can ask about multiple different features under multiple different hypotheses, what questions are most informative (and what implications it carries for the focal hypothesis) depends not just on diagnosticity but on background beliefs about the different features one could possibly learn about. Nelson (2005) argues that maximising diagnosticity is sometimes inferior to other sampling norms. He compares diagnosticity to several other norms that have been proposed for assessing a question's usefulness: information gain (expected reduction in uncertainty from new information), probability gain (how much the information improves the expected probability of making a correct guess), and expected belief change. Though these all seem very closely related, Nelson shows using computational simulations that diagnosticity conflicts with the other normative standards in some situations, and that in these situations, the other normative standards perform better. Diagnosticity seems to lack useful properties compared to other norms, including sensitivity to priors, being finite, and being equally applicable in situations with many different hypotheses (Nelson, 2005).

Confusion about the normative status of diagnosticity may be a result of confusion between the two types of pseudodiagnosticity I distinguished above: interpreting information as more diagnostic than it is, and seeking out nondiagnostic information. Diagnosticity provides a normative standard, according to Bayes' rule, for how one should *interpret* and update one's beliefs based on evidence. Whether or not it provides a normative standard for what information one should *seek out* is less clear.

Interpreting evidence as more supportive of the focal hypothesis than it actually is - due to failing to consider that it might be equally consistent with alternative hypotheses - is one way that inference processes might be biased in favour of the focal hypothesis. Another related way that inference might be biased is if people apply different standards of evaluation to supportive and conflicting evidence: accepting the former at face value, while applying much more scrutiny and scepticism to the latter.

2.2.2.2 Biased evaluation of evidence

Pyszczynski and Greenberg (1987) argue that, in general, people tend to overweight confirmatory positive evidence and/or underweight negative disconfirmatory evidence. One specific case of this is the tendency for gamblers to persist irrationally in the belief that they are winning or on a lucky streak, because they accept their wins at face value as evidence of their competence, but explain away their losses (Gilovich and Thomas, 1983). In an experimental context, Gilovich and Thomas found that people spent more time explaining their losses than their wins when asked. Content analysis of the explanations given supported the idea that such explanations served to 'discount' losses and 'bolster' wins. Similarly, Kuhn found that when young adults were shown evidence inconsistent with a theory they favoured, they "either failed to acknowledge discrepant evidence, or attended to it in a selective, distorting, manner." (Kuhn, 1989, p.677) The exact same evidence was interpreted in one way in relation to a favoured theory, and a completely different way in relation to a theory that was not favoured. This provides evidence for a link between the tendency to weight confirmatory and disconfirmatory evidence differently, and the phenomenon of belief persistence discussed earlier.

Lord et al. suggest similarly that people are "apt to accept 'confirming' evidence at face value while subjecting 'disconfirming' evidence to critical evaluation, and as a result they draw undue support for their initial positions from mixed or random empirical findings." (Lord et al., 1979, p.2098) To test this hypothesis of biased assimilation of evidence, they exposed subjects with opinions on either side of the capital punishment debate to two fake studies: one seemingly confirming and one seemingly disconfirming the deterrent effects of the death penalty. As predicted, subjects with differing prior opinions seemed to differentially evaluate the quality and convincingness of the same empirical studies and findings. For example, for a study that claimed to show a deterrent effect of the death penalty, the mean rating given by participants for how well the study was conducted was 1.5 for those who initially supported the death penalty, and -2.1 for those who initially opposed it (on a scale of -8 to 8, from very poor quality to very good quality.) An analysis of variance on the differences between the ratings of convincingness of prodeterrence and antideterrence studies found a significant main effect of initial attitude ($p < 0.001$). They also found evidence that this differential evaluation led to attitude polarization - when asked for their final attitudes, those who were originally proponents reported being more

in favor of capital punishment ($p < 0.001$), whereas those who were originally opponents reported being less in favor ($p < 0.01$).

Several other studies have used experimental paradigms similar to Lord et al. (1979) and find similar results. Liberman and Chaiken (1992) asked non-coffee drinkers and heavy coffee drinks to read summaries of one study supporting, and one not supporting, a link between coffee drinking and disease - finding non-coffee drinkers strengthened their belief that coffee drinking caused the disease, whereas heavy coffee drinkers did not. Taber et al. (2009) found that, on eight different issues including marijuana legalization and tuition increases, attitude polarization occurred for participants with strong prior beliefs (people with strong prior beliefs in either direction strengthened their opinions and thus moved further apart - but the same did not occur for those with weak priors.)

Taber and Lodge (2006) found that participants rated pro-attitudinal arguments as stronger than counterattitudinal ones: a regression of argument strength ratings on prior attitudes found significant, positive coefficients for both topics (affirmative action and gun control), across two studies. They also found that people spent longer reading and processing attitudinally challenging arguments - though the average difference is fairly small (1-2 seconds), it is greater for the more politically sophisticated (4-7 seconds) - which they interpret as evidence that people actively counter-argue incongruent evidence in a way they do not with attitude-consistent information.

2.2.2.3 Bias in inference: discussion

I have discussed findings which suggest that people are biased towards the current hypothesis in how they draw inferences from information, in two related ways: interpreting information as providing stronger support for the current hypothesis than it does, and applying different standards of evaluation to supportive versus conflicting information.

There is some evidence for the phenomenon of pseudodiagnosticity, suggesting that people tend to interpret information as more diagnostic of the favoured hypothesis than it in fact is. However, research on pseudodiagnosticity seems to have confused two related tendencies - the tendency to interpret evidence as more diagnostic than it in fact is, and the tendency to seek out information that is not particularly diagnostic, and

ignore alternative information that would be more so. Most studies related to pseudodiagnosticity show the latter tendency - suggesting that diagnosticity does not guide the search for information as much as it perhaps should - rather than directly showing evidence that people misjudge diagnosticity when drawing inferences.In addition, some of our earlier discussion of hypothesis-testing strategies suggested that in fact, people do often appropriately consider the diagnosticity of information when choosing what to pay attention to. Later studies also suggest that people are more likely to consider diagnostic information needed to compute likelihood ratios when the features they are asking about are rare, or when the initially focal hypothesis is perceived as unlikely. It's therefore far from clear that the literature on 'pseudodiagnosticity' shows that people draw inferences in a way that's biased towards the focal hypothesis.

In addition, it's not clear whether diagnosticity is actually the appropriate norm against which to compare *selections* of evidence (as opposed to how to interpret evidence once obtained). How useful it is to learn information about a feature under alternative hypotheses depends on background assumptions about that feature's rarity: and sometimes it might be more informative to learn about a more rare feature under the focal hypothesis, than to learn about a very common feature under an alternative hypothesis. Nelson (2005) argues that diagnosticity is sometimes normatively inferior to other standards based more directly on expected information gain, or expected increase in the probability of selecting the correct option. Crupi et al. (2009) also argue that the normative standards typically used in pseudodiagnosticity experiments are flawed.

There is also some evidence that the phenomenon that Lord et al. (1979) report, applying different standards to conflicting and supporting evidence resulting in biased assimilation - is not as robust as it might seem. Kuhn and Lao (1996) fail to replicate their polarization results, and conclude that genuine polarization is a real but infrequent outcome of exposure to mixed evidence. If polarization does not occur as frequently as it has been claimed to, this casts doubt on whether the biased evaluation processes said to cause it are a genuine problem.

Regardless, it is not clear that the tendency to apply different standards to different types of evidence, or to interpret information in support of the focal hypothesis, are necessarily always irrational - or avoidable. Lord and Taylor acknowledge this in their discussion of biased assimilation, concluding that "Biased assimilation most likely occurs because

it is an adaptive cognitive strategy... Having preexisting assumptions and expectations can be advantageous even when the assumptions and expectations are wrong." (Lord and Taylor, 2009, p.831)

In the face of ambiguity, my prior beliefs about the world cannot *no*t guide how I interpret new information - for me to completely disregard everything I think I already know would make every encounter with new information impossibly cognitively demanding. A relevant analogy here is the role of theories in the philosophy of science - if a scientific theory is well-established and a new observation appears to conflict with the theory, it's generally accepted that one should try to accommodate the new observation within the theory, rather than throwing out the entire theory. To some degree it is also rational for me to judge an argument that conflicts with my prior beliefs as less trustworthy than one that aligns with them - that is, to explain the conflict in a way that allows me to maintain my prior beliefs. The normative issues here are complex, but we certainly cannot straightforwardly say that allowing one's prior beliefs to influence the treatment of new evidence is always irrational.

Finally, the fact that people rate pro-attitudinal arguments as stronger than counter-attitudinal arguments cannot necessarily be taken as evidence of bias on their part, unless we know that the arguments are unfamiliar to them. If people have already come across the arguments they are shown (or similar ones), then it may be that the arrow of causation goes the other way - they thought the arguments for one side were more convincing than the other side, which is why they formed the opinion they did initially. It's hardly surprising - and certainly not indicative of bias in the normative sense - that, when then shown similar arguments, people rate those supporting their position as more convincing. None of the studies we have covered on differential evaluation seem to adequately acknowledge or address this potential issue.

Overall, though there is some evidence that people's prior beliefs influence how they draw influences from new information, none of this evidence is enough to demonstrate that they do this in a non-normative way: that prior beliefs influence interpretation more than they rationally should.

2.2.3 The relationship between bias in search and inference

A final important issue in understanding confirmation bias is how bias in search and inference relate to one another. Earlier, we mentioned that it can be difficult to say whether a search process is truly biased without also understanding how the information is later evaluated and processed. A person who seeks out largely opinion-conflicting arguments might still be guilty of confirmation bias if they then do everything in their power to rebut them; conversely, a person who seems biased towards seeking opinion-reinforcing information might not be so biased if she is aware of this bias and accounts for it by not drawing strong conclusions from it.

Klayman (1995) discusses the importance of understanding both search and inference processes together for attributing a confirmation bias. Bias in search alone, he argues, can only produce inefficiency, not necessarily biased belief, and a genuine confirmation bias only arises if one also fails to appreciate the consequences of one's search strategy. If I'm aware that I gravitate more towards opinion-reinforcing information than the opposite, then I should be much less surprised by encountering opinion-supporting evidence than conflicting evidence, and so update my beliefs less as a result of encountering it. (More technically, recognising a bias in my search strategy means that confirming evidence is more likely to arise and therefore less diagnostic.) Similarly, Klayman argues, evaluation processes do not lead to confirmation bias in isolation: "a tendency to resolve ambiguity in favour of the focal hypothesis, for example, seems like a proximal cause of confirmation bias. Even here, however, there must be the additional assumption that people do not anticipate this aspect of their cognitive processes, and thus do not take it into account." (Klayman, 1995, p.398)[1]

[1]However, Le Mens and Denrell (2011) provide evidence against Klayman's claim that a biased sample of evidence can only lead to biased judgements if one is unaware of the bias. Le Mens and Denrell argue that even if we assume that decision makers are rational and process information according to Bayes' rule, and even if they are aware of able to correct for biases in their sample of information, judgements can still be systematically biased. Le Mens and Denrell show that if the goal of the decision maker is to maximise their payoff, rather than simply to develop knowledge, even under the conditions above, they will end up systematically biased in favour of options for which information is more readily available. This suggests that a bias towards the focal hypothesis may arise even for perfectly rational Bayesian agents who are aware of biases in their sample of information, if information about their favoured hypothesis is more systematically available. In particular, if they systematically receive feedback about that hypothesis regardless of whether they seek it out or not, whereas they only receive information about the alternative hypotheses if they actively seek it out. This is an interesting finding which implies a form of confirmation bias may arise in specific conditions even under assumptions of rationality. However, this does not undermine the more general claim that Klayman makes: that understanding confirmation bias requires understanding how bias in search and bias in inference interact, rather than studying them in isolation as they often have been.

Klayman suggests that, because these interactions between search and inference processes have not been adequately addressed, the situations under which a genuine confirmation bias arises are therefore much more limited than has been supposed. He highlights the importance of thinking about the process of belief development and revision as a system, not in isolation - the study of confirmation bias needs to understand the effects of certain search and inference strategies within the context of other components of the hypothesis development process.

2.2.4 Biased beliefs: belief persistence and overconfidence

If people search for and draw inferences from information in biased ways, we should expect this to result in certain errors: in particular, a tendency to persist in believing things that are either demonstrably false or one does not have reasonable evidence for, and a tendency to be overconfident in the things one does believe. Also relevant to understanding confirmation bias, therefore, is an understanding of how and to what extent these judgement errors arise.

It seems important, however, to distinguish clearly between biases that occur in the *formation* of beliefs and associated processes - biases in search and inference - and biases in the *outputs* of these processes, i.e. biased judgement. In general, biased processes and biased judgement have been studied separately, but studying them together could help us to better understand what processes actually lead to biased outputs, and where biases in judgement come from.

Table 2.4 summarises the main kinds of 'biased beliefs' that have been associated with confirmation bias, and some of the challenges they face in the psychological literature.

Name/description	Key references	Challenges
Belief persistence: continuing to believe something even when original evidence has been discredited	Experiments in the debriefing paradigm (e.g. Ross et al., 1975) find participants continue believing something even when they are debriefed and told the evidence they were given was fake. Research on the persistence of misinformation in society, despite attempted corrections - see Lewandowsky et al. (2012) for a review, and Nyhan and Reifler (2010) for research on misinformation in a more controlled context.	No clear normative standard in most of these experiments for what people should believe and when they should change - and this is challenging given we dont know every reason a person has for believing something.
Conservatism bias: updating beliefs conservatively with respect to Bayes rule	Experiments in the bookbags and pokerchips paradigm, finding people update probabilistic estimates conservatively with respect to Bayes rule (Edwards, 1965, 1982, Peterson and Miller, 1965)	Does not necessarily result in a confirmation bias - given people seem to update conservatively in either direction, i.e. whether evidence supports or challenges the currently favoured hypothesis.
Overconfidence: holding beliefs more strongly or with more precision than is rational.	Moore et al. (2015) review different types of overconfidence and suggest overprecision in beliefs is the most durable and ubiquitous. Studies in the 2-alternative forced choice (2AFC) paradigm (Griffin and Brenner, 2004) and the confidence-level paradigm (Bazerman and Moore, 2013) find robust overprecision effects. Ecological evidence of overprecision (Arkes et al., 1981), and Ben-David et al. (2010), for example.)	Studies of overconfidence generally focus on very different types of beliefs from confirmation bias - more objective things like estimating quantities - so unclear whether this constitutes evidence for overconfidence in other domains. Though wed expect overconfidence to result from a confirmation bias, not clear that the result holds - people may be overconfident for other reasons, and so evidence for overconfidence is not necessarily evidence for a confirmation bias.

TABLE 2.4: Different types of 'biased beliefs' associated with confirmation bias

2.2.4.1 Belief persistence

The literature on belief persistence, and related literature on the persistence of misinformation, make the case that people tend to persist in believing things which are either demonstrably false, or for which the original evidence has been discredited.

Experiments in the 'debriefing paradigm' focus on the latter tendency: participants are initially given information that leads them to form a certain belief, and later 'debriefed'

and told the information they originally received was false. For example, Ross et al. (1975) gave participants false feedback indicating that they had performed either much better or worse than the average student in a novel task. Subjects are later debriefed and told the information they received was false - in the above case, told that their performance feedback was in fact unrelated to their actual performance. Despite this debriefing, subjects consistently persist in whatever belief was created by the false information - Ross et al. find that even after debriefing procedures that led subjects to say that they understood the decisive invalidation of initial test results, the subjects continued to assess their performances and abilities as if these test results still possessed some validity. (Ross et al., 1975, p.884) This effect has been replicated in a range of different scenarios (Anderson et al., 1980, Davies, 1985, 1993, McFarland et al., 2007).

A body of literature in psychology and political science documents how misinformation can persist even after repeated attempts to discredit it (see Lewandowsky et al., 2012, for a review). This differs from research on belief persistence since it focuses on topics for which there is a scientific consensus, but for which many hold misperceptions (such as climate change or the relationship between vaccines and autism). Lewandowsky et al. (2012) make a number of observations about public opinion which suggest misinformation persists despite attempts to correct for it. For example, they remark that despite the Department of Health and several other health organisations pointing to the lack of evidence for a link between vaccines and autism, and urging parents not to reject the vaccine, in 2002 between 20 and 25% of the public continued to believe in the vaccine-autism link, and 39% to 53% continued to believe there was equal evidence on both sides of the debate. Although no weapons of mass destruction were ever found in Iraq and the grounds for believing Saddam Hussein had them turned out to be unsubstantiated, 20% to 30% of Americans believed that WMDs had actually been discovered in Iraq years after the invasion (Kull et al., 2003).

It seems likely that a key factor here is source reliability: misinformation persists because people do not trust those who are communicating corrections. Someone who believes in certain conspiracy theories, for example, may well also believe that the government or media are conspiring to cover them up or have malicious intentions. Given this lack of trust in these information sources, it is perfectly rational not to adjust one's beliefs as a result of attempted 'corrections.' (Though of course, whether the mistrust itself is

reasonable is another question.) This question of how judgements of source reliability influence beliefs is one we will return to later in this thesis.

Looking at misinformation persistence in a more controlled context, Nyhan and Reifler (2010) had subjects read mock news articles including either a misleading claim from a politician, or a misleading claim and a correction, before being asked a series of factual and opinion questions related to the political issues discussed. They found that corrections failed to reduce misperceptions, especially for those who held the strongest initial views - and that corrections even had a 'backfire effect' in some cases, strengthening the original views. It has been suggested elsewhere that this 'backfire effect' may occur because people make an active effort to argue against corrections, which, if successful, leads them to feel more confident in their initial views (Lodge and Taber, 2000, Redlawsk, 2002). Other experiments using a similar design to Nyhan and Reifler have found results consistent with this: that subjects continue to be influenced by misinformation even after it has been discredited (Gilbert et al., 1990, Johnson and Seifert, 1998);.

Also relevant to belief persistence is the phenomenon of *conservatism bias* - studied in much more abstract experimental paradigms in the 1960s. In the 'bookbags and pokerchips' paradigm (Edwards, 1965, 1982, Peterson and Miller, 1965), participants are shown two bags which they are told are filled with different distributions of red and blue pokerchips. They are then told that one of the two bags will be selected at random, and they are to guess which of the two bags it is by sequentially drawing chips from the chosen bag. Participants give probabilities for how likely they believe it is each of the two bags was chosen, and update those probabilities as they draw more chips. These experiments consistently find that people update their beliefs conservatively with respect to the normative standards of Bayes' rule: "opinion change is very orderly, and usually proportional to numbers calculated from Bayes's theorem - but it is insufficient in amount." (Edwards, 1982, p.359)

However, it is worth noting that the finding of a conservatism bias is subtly different from belief persistence. In the bookbags and pokerchips experiments, people started from a neutral perspective - believing it equally likely that chips were being drawn from either of the two bags. There is not initially any focal hypothesis - and so a tendency to update conservatively does not necessarily reflect a confirmation bias. Participants seem to update conservatively on all evidence, regardless of whether it supports or conflicts with

whichever hypothesis they currently believe to be more likely. Conservatism bias might still be a general tendency underlying a broader confirmation bias, however - assuming that I start out favoring a specific hypothesis, a tendency to update conservatively on new evidence might lead me to persist in my initial belief more than is rational under certain conditions - particularly if I'm encountering more conflicting than supportive evidence.

2.2.4.2 Overconfidence

Overconfidence and confirmation bias are clearly closely related: we would expect a confirmation bias to result in overconfidence, since we defined confirmation bias as reasoning in ways that lead to strengthening confidence in the focal hypothesis more than is rational. It's therefore surprising that confirmation bias and overconfidence have not been linked more in the psychological literature - most discussion of overconfidence does not mention confirmation bias as a possible cause, and discussion of confirmation bias has, if anything, tended to refer to overconfidence simply as another form of confirmation bias (as in Nickerson, 1998, for example).

One reason for this may be that the two phenomena have typically been studied with quite different methods. Studies of overconfidence have generally focused on looking at people's estimates of known quantities, or verifiable predictions. There is a reason for this - looking at beliefs about which there are objective answers makes it much easier to determine how confident people *should* be. The confirmation bias literature, by contrast, has largely dealt with beliefs whose objective truth is difficult to assess - questions of politics or religion, for example. As we have begun to see, this causes problems for confirmation bias, as it makes assessing when reasoning is genuinely biased challenging.

Though evidence of overconfidence is not necessarily evidence for confirmation bias (since overconfidence may arise for other reasons) it is certainly worth reviewing in the context of confirmation bias, given the close relationship between the two. The term overconfidence has been used ambiguously in the psychological literature - as Moore and Healy (2008), point out, to mean (1) *overestimation* of one's performance; (2) *overplacement* of oneself relative to others; (3) *overprecision* in one's beliefs. Here, we are interested in (3) - having greater confidence in one's beliefs than is warranted given the evidence.

Moore et al. (2015) review the evidence for overprecision in judgement, arguing that overprecision is the least well understood of the three kinds of overconfidence. They point out that overprecision is the most robust form of overconfidence - though there are numerous documented reversals of overestimation and overplacement (Erev et al., 1994), there are few, if any, documented studies that find the reverse effect of overprecision. "It is exceedingly rare for people to be less sure that they are right than they deserve to be." (Moore et al., 2015, p.185)

Overprecision in judgement has been studied using several different paradigms. The first is the '2-alternative forced choice approach (2AFC)' paradigm (Griffin and Brenner, 2004). Participants see a question, choose between two possible answers, and indicate how confident they are that they have chosen correctly. We can then compare confidence with actual outcomes over a number of questions to ask whether people are, on average, overconfident[2]. Research repeatedly finds that when people are most confident, their confidence is not justified by accuracy (Lichtenstein and Fischhoff, 1980).

A second way to study overprecision is the confidence-interval paradigm - participants are asked to provide a lower and upper estimate for a quantity (the number of cows in the United States, say) such that they believe it is 90% likely the actual value falls in that interval. People are consistently too narrow - overprecise - in their intervals: Alpert and Raiffa (1982) found that 98% confidence intervals included the right answer, on average, only 60% of the time. This effect has been replicated hundreds of times (Bazerman and Moore, 2013).

Finally, there is ecological evidence of overprecision: studies have documented physicians' tendency to be overconfident in a favoured diagnosis (Arkes et al., 1981); that scientists' estimates of physical constants are excessively confident (Henrion and Fischhoff, 1986); that investors are overconfident that they know what an asset is worth and too willing to trade on that knowledge (Daniel et al., 1998); and that organisational forecasts tend to be over-precise (Ben-David et al., 2010).)

It's not clear whether evidence for this kind of overprecision in judgement should be considered evidence for a confirmation bias at all. First, we said that a confirmation bias, if it exists, will lead to overconfidence - but overconfidence might arise by other

[2]If someone says they are 90% confident, for example, we would then expect them to be wrong for one in ten questions - if they are wrong more often than this, then we say they are overconfident.

means, so evidence for overconfidence is not necessarily evidence for confirmation bias. Second, the overprecision literature studies different kinds of beliefs and uses different measured from the confirmation bias literature. Giving too-narrow confidence intervals for one's estimate of a quantity seems like a very different tendency from holding one's political beliefs too strongly. It might be that ultimately these amount to the same thing - to believe adamantly in the death penalty suggests that I am overestimating the probability it is beneficial relative to alternative hypotheses, which seems similar to the kind of overconfidence measured in the 2AFC paradigm. However, research on measures of overconfidence suggests that different ways of measuring people's confidence in their beliefs do not necessarily correlate well with one another, and are perhaps tapping different types of confidence, or different interpretations of confidence by the participant (Langnickel and Zeisberger, 2016).[3]

All of this contributes to the concern that assessing just how strong someone's opinion *should* be, especially for questions that are not easily objectively verifiable, is far from straightforward. The overprecision literature suggests that at least for some ways of measuring attitude confidence and for certain kinds of beliefs, people tend to be over-confident. But whether this generalises to the kinds of beliefs that the confirmation bias literature tends to focus on, and whether this has any implications for the status of confirmation bias, is unclear.

2.2.4.3 Belief persistence and overconfidence: discussion

I discussed several different source of evidence that people tend to persist irrationally, or be overconfident in, their beliefs: that beliefs persist even when the original reasons for them have been retracted, even when such beliefs are demonstrably false; that people tend to update their beliefs conservatively even in very abstract tasks; that judgements of how likely it is one's beliefs are correct tend to be over-precise, and that people display more confidence in their beliefs than is warranted.

[3]This seems related to Krosnick et al. (1993) claim that attitude strength is not a single construct, but should rather be considered multiple related constructs - and different ways of asking people how strong their belief is may elicit different responses by highlighting different aspects.

However, whether any of these findings actually demonstrate a 'confirmation bias' is still unclear. Claiming that belief persistence is irrational, or that confidence is inappropriately high, requires some objective standard: for what people should believe and how strongly.

Using confidence intervals and asking people to assign probabilities to beliefs helps us to do this. However, asking people to assign probabilities to beliefs is problematic in itself, since there are a number of ways in which people may find it difficult to intuitively work with or understand probabilities (Moore et al., 2015). It is not clear whether cases of overconfidence are due to a confirmation bias, or whether they might be partially explained by the fact that people have difficult reasoning with probabilities. Studies of overconfidence that do not make use of probabilities, on the other hand, lack any normative standard against which we can say that people are really irrationally confident.

Normative interpretations of belief persistence studies are also problematic, given that we do not know every reason or piece of evidence a person has for believing what they do. Even if the original reason for a person's belief has been discredited, they might in the meantime have acquired other, independent evidence that supports their belief. In the debriefing paradigm, participants might also reasonably 'downgrade' their assessment of how trustworthy the information they are being given is - the experimenter did admit to giving them false information initially, after all, so why should their 'corrected' information be trusted? Accounting for when it is reasonable for one to persist in believing something is complex and depends a lot on what other relevant beliefs the subjects have, which we often do not have access to.

2.2.4.4 Biased beliefs independent of search and inference?

Belief persistence and overconfidence are naturally thought of as the outcomes of confirmation-biased processes: beliefs persist and are held too strongly because people tend to seek out and interpret information in ways that reinforce them, regardless of reality or the available evidence. However, this might not be the only way to think about belief persistence and overconfidence. It seems possible that these biased 'outputs' might arise for reasons other than biases in search and inference.

One reason to suspect this might be the case is that there are cases of both conservatism bias and overprecision even when people appear to have had no opportunity to search for or draw inferences from information in biased ways. In studies of overprecision, for example, people are often asked to give probability estimates or confidence intervals in response to questions they might never have thought about before, and to do so immediately. That is, people express overconfidence in their estimates immediately, rather than forming a judgement and then ending up overconfident in it because they seek out and interpret evidence in biased ways. Similarly, in bookbags and pokerchips experiments, participants do not have any control over what new information they encounter or how they interpret it - they are simply shown draws from the 'mystery' bag, which are unambiguously one colour or another. This suggests a more fundamental tendency to under-update from the focal hypothesis, which can't be explained in terms of updating from a biased sample of evidence.[4]

It seems plausible that a more basic form of 'bias' in judgement exists, rooted more in how we think about probabilities and represent beliefs than how we seek out and interpret information. Since we have limited cognitive capacities and imperfect information, we cannot simply do precise calculations using Bayes' rule every time we encounter new information, and it is often impossible for us to know how confident we are justified in being probabilistically. Our estimates of how confident we should be, and how much we should update our beliefs, are therefore just that - estimates, made using certain rules of thumb or heuristics. It therefore seems plausible that a tendency to overestimate how confident we should be in our beliefs, and to persist in believing things we should not, might arise from heuristics used to estimate confidence and revise beliefs directly - rather than being the output of search and inference processes.

The literature on overconfidence also discusses several explanations for such a tendency, none of them confirmation bias. This may be an oversight or a result of the strange gulf between these two similar literatures - but also suggests the possibility that overprecision arises for reasons other than a confirmation bias. For example, Moore et al. (2015) suggest the overconfidence in judgement may be at least partially explained by conversational norms - people want to come across as credible and persuasive more than they care about expressing their beliefs accurately. They also consider several other

[4]However, an alternative interpretation here might be that when people are posed these questions, they search their memories of past experience for relevant information - which then could explain biases in output in terms of biased search/inference processes.

possibilities: that overconfidence is a result of the fact we only have access to a small sample of information and do not appreciate/adjust for this; that overconfidence may be a compensatory mechanism that offsets other biases; or that it may simply be a consequence of people's failure to understand probability distributions.

If overconfidence is independent of confirmation bias, it's also possible that overconfidence creates and illusion of confirmation bias where it does not actually exist. If someone is overconfident in their beliefs but we believe they are appropriately confident, then it will look like they are seeking out and interpreting information in biased ways (relative to the level of confidence we think they have.) The problem here is not how they reason given their prior beliefs, but rather that their prior beliefs are overconfident in the first place. This is currently just speculation, but it's possible that overconfident judgements arise for some other reason than confirmation bias, and the fact people seem unwilling to change their minds stems more from overconfidence than from reasoning processes biased towards confirmation.

2.2.5 Summary: the evidence for confirmation bias

I have discussed the various ways in which a confirmation bias might arise in reasoning. The term 'confirmation bias' has been used broadly, to refer to a variety of different phenomena, without a clear explanation of how these phenomena link together. One intention of this review was therefore to provide a clearer picture of the different things 'confirmation bias' might refer to, and how these relate to one another at different stages of information processing.

Table 2.1 summarises the main different types of confirmation bias discussed in the literature, categorised by different stages of reasoning. As discussed, the third stage (biases in beliefs - i.e. the output of the process), might either be thought of as errors that result from the earlier stages - ways that biased search and inference can lead to irrationally persistent or overconfident beliefs - or plausibly to a third category of biases that occur independently of biased search and inference, due to more fundamental aspects of how we assign confidence to and revise beliefs.

Despite the number of ways in which confirmation bias might arise, and research purporting to investigate the bias at each of these stages, the evidence reviewed faces numerous

problems. A positive test strategy in hypothesis testing has generally been cited as evidence of biased search, but it is now generally agreed that a positive test strategy does not necessarily correspond to a confirmation bias (Klayman, 1995). The other literature cited as evidence for bias in search is the selective exposure literature, which has a mixed history, though Hart et al. (2009) do conclude that overall there seems to be a moderate effect in this direction. There is some evidence that time spent seeking information reduces as one becomes more confident in the current position - but this evidence is somewhat indirect, coming from studies of motivated reasoning rather than confirmation bias itself.

When it comes to bias in inference, the picture is also mixed. Perhaps the most convincing evidence comes from research on pseudodiagnosticity, which finds that people tend to interpret evidence as lending more support to the focal hypothesis than is reasonable, likely due to a failure to consider how the evidence might be interpreted under alternative hypotheses. However, we cannot avoid prior beliefs influencing the interpretation of new information *at all* - and it is unclear to what extent it is reasonable to expect people to consider alternative interpretations of evidence for any given piece of new information, especially under time and processing constraints. A number of studies also document how people seem to apply different evaluative standards to evidence that supports what they believe than evidence that conflicts with it. However, similarly, the normative issues here are not straightforward - many of the relevant experiments do not allow one to distinguish between the rational influence of prior beliefs on the evaluation of new evidence (as Bayes' theorem would prescribe), and a 'bias' that goes beyond that.

I next discussed evidence for belief persistence and overconfidence, arising either as a result of biased processing or independently. Again, the interpretation of evidence here is complex, and many of the relevant experiments do not give us enough information for us to determine whether the strength or persistence of people's beliefs is genuinely irrational. We also discussed whether a bias towards overconfidence and/or persistence of beliefs might exist independent of other biased reasoning processes, arising from the use of imperfect heuristics to estimate and revise confidence in beliefs.

Finally, I discussed the case made by Klayman (1995) that bias in search and inference should not be studied in isolation, since neither on its own necessarily leads to a systematic bias in favour of the focal hypothesis. It is particularly surprising that none of the

issues Klayman raises are discussed in Nickerson (1998) (and that Nickerson does not even cite Klayman), despite the fact that Klayman's paper was published three years earlier and raises genuine objections to the view of confirmation bias as the persistent, ubiquitous phenomenon Nickerson claims it to be.

On more detailed investigation, the case for confirmation bias seems much weaker than is generally appreciated in the psychological literature. I discuss the main challenges to confirmation bias in more detail in the next section.

2.3 The challenges to confirmation bias

In this section, I will discuss in more detail several reasons why confirmatory reasoning may not necessarily be evidence of any systematic irrationality:

1. Strategies that look confirmatory **may not be systematic** - they may only result in confirmation in specific scenarios, and research may have focused just on these non-representative cases;

2. Strategies that look biased in specific experimental scenarios may in fact **be accurate on average** across the kinds of real-world situations people encounter;

3. **Problems with experimental design** may mean we lack certain information that factors into participants' judgements, meaning we misinterpret them as irrational (when in fact they would seem rational given more information);

4. **A lack of clear normative standards**, or disagreement about what the appropriate normative standard is, means that many attributions of 'bias' and 'irrationality' are made unreflectively, without clarifying what unbiased or rational behaviour would be in contrast;

5. Finally, if we want to say that reasoning processes lead to biased beliefs, **we cannot study different parts of the reasoning process in isolation** - we need to understand how, for example, bias in search and inference interact - and most research on confirmation bias has little to say about this.

What this demonstrates is that establishing that a systematically irrational tendency towards confirmation exists is much harder than it first seems: requiring much more

information and clearer normative standards than most experiments typically provide, and requiring an understanding of both different stages of the reasoning process, and how reasoning processes interact with the features of different environments.

2.3.1 Strategies that look 'confirmatory' do not necessarily lead to a confirmation bias in all circumstances

One error that studies of confirmation bias have made is to assume that it is enough to show people use a strategy that looks 'confirmatory' under only certain circumstances. This is particularly the case in the hypothesis-testing literature, where it has been assumed that a positive test strategy always leads to confirmation.

Remember that a positive test strategy means a tendency to ask questions for which the answer would be 'yes' if one's hypothesis were true.[5] So, for example, when trying to discover a rule by asking whether given cases fit the rule, people tend to test cases they expect will fit the rule rather than those they expect not to (Wason, 1968). However, Klayman and Ha (1987) point out that testing cases expected to fit the current hypothesis is not the same thing as testing those that have the best chance of verifying the hypothesis. In some situations these two will be equivalent, and it just so happens that the task used by Wason (1968) was such a situation - but this does not always hold.

To see why, consider that the error most people made in Wason's 2-4-6 task was to focus on hypotheses that were too narrow. The actual rule was 'three ascending numbers, but many people started off with a narrower hypothesis, such as 'numbers ascending by two'. If, however, this had been the other way round - if the general rule was narrow, and people started off with the hypothesis that it was simply ascending numbers - then testing cases they expected to fit the rule would quickly falsify their hypothesis. In general, a positive test strategy *can* reveal errors in a hypothesis, but only false positives (cases one expects to fit the rule that do not) - not false negatives (cases one does not expect to fit the rule but do.) A tendency towards positive hypothesis testing therefore means that errors will primarily be in the direction of holding overly-narrow hypotheses, but does not necessarily imply a confirmation bias. Even when subjects use a positive test strategy, they do sometimes seem to be trying to falsify their hypotheses

[5] Another way of phrasing this is that a positive test strategy involves testing cases expected to have the property of interest, rather than those that might lack that property.

- by, for example, testing extreme or unusual instances (such as -2, 0 2 if the hypothesis is ascending numbers by two) (Klayman, 1995).

The problem here, then, is that experimenters have found that people reason in ways that leads them to obtain more confirmatory evidence under certain circumstances, and assumed that such reasoning strategies therefore constitute a confirmation bias - without considering how these strategies play out under different circumstances.

2.3.2 Strategies that look like they produce 'bias' may in fact be accurate on average

Building on the previous point, strategies that look like they produce bias under certain circumstances may actually be highly useful and accurate across a range of real-life scenarios. For example, if false positives are more prevalent and more important to identify than false negatives in most situations we encounter, then a positive test strategy may be a good general-purpose heuristic.[6]

Klayman and Ha argue along these lines that a positive test strategy "is actually a good all-purpose heuristic across a range of hypothesis-testing situations... Under commonly occurring conditions, this strategy can be well suited to the basic goal of determining whether or not a hypothesis is correct." (Klayman and Ha, 1987, p.212) They distinguish disconfirmation as a *goal* from disconfirmation as a *search strategy* - and show that sometimes, a positive search strategy can be a good or even the only way to discover falsifying circumstances. Beyond this, in many real-life environments, which are probabilistic (as opposed to deterministic laboratory settings), it is not even necessarily clear that seeking falsification is the way to get the most information.

The point being made here is not that the average response to Wason's rule-discovery task is the normative response to that task specifically. However, given the cognitive demands of figuring out the optimal strategy for every different situation, we may have to develop all-purpose heuristics which work well across a range of scenarios - and even if they produce errors under certain circumstances, they may be unbiased on average across all those scenarios. It may also be the case that the conditions in most important

[6]This alludes to some issues around how we think about rationality, given that we accept that we have cognitive limitations and therefore have to use shortcuts or heuristics a great deal of the time - which I discuss in more detail elsewhere.

hypothesis-testing situations we encounter in real life are very different from those in Wason's abstract experiments.

Oaksford and Chater (1994) apply a similar approach to provide a rational analysis of Wason's 'selection task' (Wason, 1960). The classical interpretation of this task is similar to the rule-discovery task, that people are irrational because they only seek to test positive instances, and do not seek to falsify the hypothesis. Oaksford and Chater argue, however, that under certain assumptions, falsification may not be the normative strategy for such a task.

In the selection task, subjects must choose which card-turning experiments they expect to give them the most information about which of the two hypotheses is true (whether or not a dependency - 'if p, then q' is true or not.) Oaksford and Chater (1994) formalize this notion of expected information gain using the theory of optimal data selection from statistics, and show that participants' choices on the selection task follow the theory of optimal data selection if it is assumed that the two features they are looking for (p and q) are rare in the environment.

To see why, recall that to test the rule 'if p, then q', subjects can choose from the following four cards: *p*, *q*, *not-p*, and *not-q*. The p card is clearly informative regardless of what is on the other side - if it has a q on the other side, this supports the rule, and if it does not, this falsifies the rule. The not-p card is clearly *not* informative - one could find either q or not-q on the other side and this would make no difference. The not-q card is informative *if* we find a p on the other side (which would falsify the rule), but not informative if we find a not-p on the other side (since this could be consistent with either hypothesis). Similarly, the q card is informative if we find a p on the other side, but not if we find a not-p. Whether the not-q or the q card is more informative therefore depends on how likely we think these different outcomes are - how likely each of them is to falsify the rule vs. provide no information. If we have reason to think that the feature p is rare, then the chances of the not-q card falsifying the rule are low. The q card is more informative the more rare both p and q are. Therefore, if we have reason to think that both the features p and q are rare, the q card will in fact be more informative in expectation than the not-q card (the supposedly 'irrational' choice many people made in these experiments).

Therefore, though Wason's experiments have often been considered the first evidence of a confirmation bias in reasoning, this link seems to be on shaky ground. Not only does a positive test strategy not necessarily lead to a confirmation bias under all conditions (Klayman, 1995, Klayman and Ha, 1987), but it may be argued further that a positive test strategy may actually be the best way to maximise information gain in normal conditions (Klayman and Ha, 1987, Oaksford and Chater, 1994).

2.3.3 Problems with experimental design

Corner et al. (2010) argue that 'conservatism bias' - the tendency for belief revision to be conservative with respect to Bayes' rule - may be an experimental artefact rather than indicating a systematic bias. They highlight a challenge for all experimental research, that "in order to be able to accurately understand behaviour in an experiment, it is vitally important to have a complete understanding of what the *participants* in the experiment think they are doing, in case it differs from what the *experimenters* think they are doing." (Corner et al., 2010, p.1627, emphasis in original).

In belief revision experiments, Corner et al. suggest that participants may not trust that the evidence they receive comes from a fully reliable source. This seems likely if they have participated in previous experiments, and especially if such experiments involved a deception manipulation (Kelman, 1967). A less reliable source should lead to more conservative belief revision - and so if participants treat the evidence they receive as coming from a somewhat unreliable source, they should update conservatively with respect to a normative standard. In the bookbags and pokerchips paradigm, for example, participants might be skeptical of the experimenter's claim that he is drawing chips from the bags randomly (which is very reasonable, given that often such draws are not random!) Given this assumption - that participants treat experimenters as only partially reliable sources in conservatism experiments - their responses are entirely rational, and do not support any claim of systematic bias (Corner et al., 2010).

More broadly, many experiments simply do not provide enough information - about what the participants' aims are, for example, or what prior information they have that may be relevant - in order to conclude that people are behaving rationally or irrationally. This is also part of the problem we noted with studies of selective exposure: without knowing what people already know, without knowing their motivations behind seeking

certain information, and without knowing how people interpret the information, it is very difficult to conclude anything about what information they *should* select.

2.3.4 Lack of normative standards

The phenomenon of belief polarization - when two groups with opposing initial views both strengthen their beliefs based on reading the same evidence - has often been cited as evidence of confirmation bias. It seems that the most plausible explanation for this polarization is that people apply different standards of evaluation to evidence that supports what they believe than that which conflicts with it, resulting in each side weighing supportive evidence more heavily.[7]

Jern et al. (2014) show that belief polarization can be consistent with a normative account of belief revision - that in some cases, rational agents with opposing beliefs *should* both strengthen their positions as a result of reading the same information. Typically, studies of belief polarization have not explicitly included normative models of how people should interpret information and update their beliefs, simply relying on the common-sense assumption that belief polarization is irrational. Jern et al. show that this assumption is not as reasonable as it might seem by presenting a normative probabilistic analysis within which belief polarization can arise. They then apply this model to previous studies of belief polarization to show how their results may be consistent with a normative theory of belief updating.

Consider the situation in which two people observe data D which bears on some hypothesis H. Contrary updating occurs whenever one person's belief in H increases after observing D, and the other person's belief in H decreases after observing D. This can be contrasted with parallel updating, where both people update their beliefs in the same direction. The conventional wisdom is that parallel updating is always the normative outcome (Lord et al., 1979).

However, in Bayesian terms, whether or not two people increase or decrease their belief in H after observing D depends on their likelihood ratios - which in turn may depend on the assumptions they each make about factors influencing the relationship between the hypothesis H and the data D. Jern et al. (2014) consider a number of different

[7]Belief polarization may better be thought of as an illustration of a broader tendency for people to interpret information in asymmetric ways, depending on whether or not it fits with their preconceptions.

relationships between H and D that might give rise to contrary updating, represented using Bayesian networks. As a simple example, suppose two doctors are given a patient's test result (D), which we assume has only two possible outcomes (positive/negative), and there are two hypotheses for what disease the patient has. If the patient has disease 1, the test is likely to produce a positive result, and it the patient has disease 2, the test is likely to produce a negative result. However, if factor V represents whether the patient has low or high blood sugar, and this factor affects the meaning of test result D, and two doctors disagree about the value of V, then two doctors could agree on everything else, behave as normative Bayesian agents, but end up updating in different directions based on data D.

More generally, in the real-world, hypotheses and data are rarely considered in isolation, and inferences about one hypothesis typically depend on other hypotheses and beliefs. Jern et al. (2014) take this approach to explain how the results of Lord et al.'s classic (1979) study may arise under normative probabilistic inference. To recap, in this study supporters and opponents of the death penalty were asked to read about two fictional studies, one supporting and another opposing the idea that the death penalty is an effective crime deterrent. Jern et al. suggest that if participants make two simple assumptions: (1) that studies are influenced by research bias, and (2) that one's own beliefs about the effectiveness of the death penalty differ from the consensus opinion among researchers, then belief divergence can arise through normative probabilistic inference. Given these assumptions, Alice's prior belief that the death penalty is an effective deterrent gives her reason to be sceptical of the study showing the opposite conclusion - she expects the researchers believed the opposite conclusion, and so researcher bias may have influenced the results. If Bob had the opposite prior belief and the same assumptions, he would be sceptical of the other study, and so each would put less weight on the study opposing their initial viewpoint, therefore leading them to update in opposite directions.

Of course, no claim is being made here about whether it is reasonable for participants to make such assumptions, or even that it is likely they were making such assumptions. This simply illustrates how, *given* certain assumptions, putting more weight on evidence that supports your prior beliefs may not be entirely irrational, and result in two people with different prior beliefs drawing opposite conclusions from the same information.

Jern et al. (2014) also discuss some other conditions under which findings of belief polarization may be normative. First, they suggest that polarization may emerge as a consequence of mapping an ordinal variable (the strength of the effect the death penalty has on crime deterrence) onto a binary variable (whether or not the death penalty is an effective deterrent.) This seems similar to the suggestion we made that overconfidence may arise from how people map their beliefs onto probabilities rather than biases in reasoning. Second, they consider the case where participants with strong and weak Christian beliefs read a story describing how church leaders had conspired to cover up new evidence undermining the idea that Jesus is the son of God (Batson, 1975). They suppose that participants have other beliefs which influenced both their initial judgements about whether Jesus is the son of God or not, and which influence their expectations about what the information would mean if he were - characterised as a certain worldview. For instance, someone with a Christian worldview believes that Jesus is probably the son of God, and that followers of Jesus are likely to have their faith challenged by others. Someone with a secular worldview believes that Jesus is probably not the son of God, but that if he were, his followers would be unlikely to encounter challenges. These worldviews affect their interpretation of the data that seems to challenge faith in Jesus - and so two people with differing prior views will disagree about whether this provides support for or against the hypothesis that Jesus is the son of God, and diverge as a result.

Again, the authors are not claiming that these interpretations necessarily explain what is going on in the experiments discussed. However, they are suggesting that these interpretations or similar ones are *possible*, and that therefore it is not straightforward to simply claim that belief polarization is irrational. More broadly, Jern et al. (2014) illustrate how, given certain assumptions, it is rational to give more weight to confirmatory evidence, and therefore interpret apparently 'balanced' or neutral data as supportive of one's current hypothesis. As Klayman puts it, "from a Bayesian point of view, the fact that a study gives a surprising result does constitute valid probabilistic evidence that the study was done incorrectly... how much distrust of disconfirming results is appropriate and how much is too much? The normative issues here are complex and remain unresolved." (Klayman, 1995, p.395) We will look at some of these complex normative issues, and their impact on understanding confirmation bias, in more detail in chapter 4.

2.3.5 Bias in search and inference cannot be studied in isolation

Though bias in search and inference have generally been studied as separate phenomena, perhaps they should not be - it is arguably only with certain combinations of search and inference that the real problems arise. Being biased in how one searches for information isn't so problematic if one interprets and updates on that information rationally. Similarly, a bias in how one interprets information is at least *less* of a problem if one starts with balanced and unbiased information. It is therefore difficult to draw any conclusions about confirmation bias as a broad phenomenon without understanding both bias in search and inference, and how they interact. As Klayman (1995) points out, the tendency to study bias at different stages of reasoning independently, and then to claim each demonstrates a confirmation bias, is perhaps one of the biggest problems with this literature.

This problem is particularly apparent in the selective exposure literature. I argued that mixed findings may fundamentally be because 'selective exposure', as typically defined, is a poor measure of confirmation bias - whether someone seeks out more confirmatory evidence or not tells us little about whether they are reasoning in ways biased towards the current hypothesis, since this in turn depends on their motivations and how they draw inferences from that information. A person might seek out balanced evidence and yet still be guilty of confirmation bias if they evaluate confirmatory and disconfirmatory evidence by unreasonably different standards. Conversely, a person might 'selectively expose' themselves to a great deal more supportive evidence, but if they are aware of this tendency and accordingly hesitant to draw any strong conclusions from it, then they do not necessarily exhibit a confirmation bias.

Mckenzie (2004) makes this point more formally - arguing in line Klayman (1995) that neither bias in testing hypotheses nor in the evaluation of information, in themselves, necessarily lead to confirmation bias - but certain combinations do. McKenzie discusses one such combination - a positivity bias in how one tests hypotheses, plus insensitivity to differences in the diagnosticity of different answers to questions - explaining how neither tendency on its own creates a combination bias, but together they do.[8]

[8]McKenzie also goes on to argue that even this combination leads to bias less often than might be supposed, since the bias seems to decrease when the materials used are familiar (rather than abstract.)

A positivity bias in testing hypotheses is essentially the same as the positive test strategy discussed previously - preferring to ask questions for which a 'yes' answer under the current hypothesis is more likely. Consider 2.5 below - where one's task is to determine whether a patient has disease A or B, and one can choose to do tests from 1 to 4. The probabilities in the table indicate the probability that the test will yield a positive result if either disease A or B is present. Positivity bias suggests that, if a doctor currently thinks disease A is more likely, he will choose to do test 4 - since this is most likely to yield a 'yes' answer under hypothesis A - and that he will prefer test 1 if he favors disease B.

Test	Disease A	Disease B
1	50%	90%
2	50%	60%
3	50%	10%
4	90%	50%

TABLE 2.5: Probabilities of observing positive test results given diseases A and B

Mckenzie (2004) explains how this positivity bias - though it looks like a form of confirmation bias - will not necessarily lead one to irrationally strengthen confidence in the focal hypothesis, so long as one updates rationally (i.e. in accordance with Bayes' rule) based on the evidence obtained. Assume that the doctor currently believes the patient has disease B, and chooses to do test 1. Although a positive result to test 1 is more likely than a negative result, a positive test result is also less *diagnostic* on account of being more likely - and so should cause the doctor to update his belief less. On average, then, choosing test 1 should not cause one to update in favour of diagnosis B.

In general, even if people display some 'bias' towards information expected to support the focal hypothesis, if they are sensitive to the diagnosticity of information, they should not end up overconfident in that focal hypothesis. This is because likely outcomes are less diagnostic than likely ones - so even if supportive evidence is more likely, unsupportive evidence should cause one to update more, which on average balances out.

Being insensitive to how diagnostic information is does not, in itself, lead to confirmation bias either - since whether one obtains more supportive information depends on

the questions asked. Insensitivity to differential diagnosticity should then affect both hypothesis-confirming and disconfirming evidence equally. What *does* result in confirmation bias is the combination of 'positive tests' and insensitivity to the diagnosticity of information. Asking questions more likely to 'support' the focal hypothesis means that a result supporting the focal hypothesis should be less diagnostic, since that result is more common. Insensitivity to diagnosticity means that people will, in addition, overestimate how diagnostic this supportive information is - and so people are both more likely to encounter supportive information, and more likely to overweight it. More intuitively: asking questions in ways that make you more likely to encounter supportive information does not lead to bias if you account for the fact that a supportive answer was more likely in how you weigh that evidence - and failing to discriminate between the diagnosticity of evidence does not lead to bias if this failure to discriminate affects supportive and conflicting evidence equally. It is only when these two 'biases' are combined, that genuine confirmation bias, and overconfidence in the focal hypothesis, results. Almost all research on confirmation bias discussed in the literature (including Nickerson, 1998, , which is often cited as conclusive evidence for confirmation bias) fails to appreciate this important point.

2.4 What remains of confirmation bias?

Having identified some key challenges faced by much of the research on confirmation bias, we can take a more systematic look at how each of the findings discussed fare against these challenges. We pass each of the findings discussed through three stages of scrutiny:

1. How **robust** is the finding? That is, does the tendency that is claimed to exist even clearly exist - before even we ask whether it's evidence of a confirmation bias?

2. Does the finding show a **systematic tendency** to confirm the focal hypothesis?

3. Is the finding said to be irrational relative to some explicit **normative standard**? How much agreement is there that this is the appropriate normative standard?

TABLE 2.6: How strong is each piece of evidence for confirmation bias?

Finding	**Robust?**	**Confirmatory?**	**Non-normative?**
Bias in search			
Bias in hypothesis testing: particularly the use of positive test strategies, asking questions expected to yield positive answers if the hypothesis is correct (Snyder and Swann, 1978, Wason, 1960, 1968)	Not very - positive test strategy seems robust in Wasons basic paradigm, but less clear when extended to more familiar contexts (where people seem more likely to ask diagnostic questions)	Not necessarily - a positive test strategy can sometimes lead to disconfirmation (by identifying false positives), so isnt the same as confirmatory reasoning.	Contested - some have argued that falsification is not the appropriate normative standard for hypothesis testing (Oaksford and Chater, 1994), and that a positive test strategy may be accurate across most real-life scenarios.
Selective exposure: looking for information in places expected to support current hypothesis (Hart et al., 2009)	No - several decades of research have produced very mixed findings, with some studies finding the opposite effect, and effects seeming highly dependent on subtle moderating factors.	Not necessarily - whether or not selective exposure leads to confirmation depends largely on how information is interpreted, and the motivations/intentions behind seeking out different types of info.	No - there are no explicit normative standards in selective exposure studies, and it is simply assumed that unbiased means reading equal arguments on either side of an issue.
Myside bias in argument production: selectively searching memory for confirmatory info (Toplak and Stanovich, 2003)	Fairly robust - no contrary findings, but we did not find particularly thorough or extensive research on this tendency. Not necessarily.	There is some evidence that people tend to find it easier to produce one-sided arguments, even if the side they are being asked to argue for is not their own position - suggesting that the tendency is not so much to confirm ones existing beliefs, but a difficulty splitting attention between two sides of an issue.	No explicit normative standards are used, and its acknowledged (e.g. by Wolfe and Britt (2008) that this tendency is not necessarily a bias - whether or not it is considered one depends on what people believe the goal/purpose of generating arguments is.

Finding	Robust?	Confirmatory?	Non-normative?
Bias in amount of time spent searching: stopping search for information when evidence points in favour of current/favoured hypothesis, continuing search for information when it does not (Ditto and Lopez, 1992)	Not very - most of the evidence for this tendency comes from studies of motivated reasoning, not confirmation bias per se.	Not necessarily - since much of the evidence here comes from the motivated reasoning literature, it may be that people terminate information search depending on whether they have reached a conclusion they like - which is not necessarily the same as confirming their current belief.	No - the way bias is generally measured in these studies is somewhat intuitive, but this intuition is not justified any further.
Bias in inference			
Interpreting ambiguous evidence as supportive of currently favoured hypothesis (Feeney et al., 2000, Fischoff and Beyth-Marom, 1983)	Unclear - much of the evidence commonly cited for this tendency is actually evidence for the related but subtly different tendency of pseudodiagnosticity - seeking diagnostically useless information. Evidence that people actually interpret evidence as more diagnostic than it should be is less clear.	Probably, but not always - if such a tendency did exist, it would seem likely to result in a tendency to confirm the focal hypothesis. This is not totally a given, however - as Klayman (1995) points out, this does also require that people do not anticipate and adjust for this tendency in how they search for information.	Not necessarily - the normative standards for assessing the diagnosticity of evidence are clearer than those used in many studies, but it has been challenged whether or not maximising diagnosticity is the best norm for choosing what information to sample (Nelson, 2005).

Finding	**Robust?**	**Confirmatory?**	**Non-normative?**
Applying different standards of scrutiny/evaluation to supporting and conflicting evidence - **biased assimilation and polarization** (Lord et al., 1979, Pyszczynski and Greenberg, 1987, Taber and Lodge, 2006)	Fairly - since Lord et al. (1979), several other studies have found similar results - that people with differing prior beliefs interpret the same information differently and therefore diverge in their resultant opinions. However, there have also been some failed replication attempts (Kuhn and Lao, 1996).	Yes.	Not necessarily - studies of biased assimilation generally do not use explicit normative standards, instead assuming that it must be irrational for people to draw different conclusions from the same information. However, Jern et al. (2014) point out that this can be rational, if people genuinely have different prior information that influences their interpretation of information - and use more explicit normative models to show how this can occur.
Biased beliefs			

Finding	Robust?	Confirmatory?	Non-normative?
Belief persistence in the debriefing paradigm (Ross et al., 1975)	Yes, fairly - the main effect has been replicated across a range of scenarios.	Yes, basically by definition - persisting in the current belief is a form of confirmation - however, whether belief persistence actually results from specific confirmatory reasoning processes (e.g. biased assimilation), is less clear.	No, not necessarily - studies of belief persistence generally do not invoke any explicit normative standards (other than the intuitive people should not persist in believing things when the evidence is discredited.) This fails to account for the possibility that people may have subsequently found or remembered additional evidence for their belief, or that they may not entirely trust the retraction of evidence, for example.
Persistence of misinformation in society despite corrections (Lewandowsky et al., 2012)	Fairly robust - this has been observed in public opinion for a range of topics (Lewandowsky et al., 2012), and in some more controlled contexts (Nyhan and Reifler, 2010)	Yes, with the same caveat as belief persistence above (its not clear exactly what kinds of processes lead to persistence.)	No, not necessarily. Particularly when looking at public opinion in naturalistic contexts, it is very difficult to draw normative conclusions about what people should believe, without knowing what information they have access to. In a narrow sense we can say that believing false things is non-normative, of course. But we cannot necessarily say that people are irrational to believe these things.

Finding	Robust?	Confirmatory?	Non-normative?
Conservatism bias: updating beliefs conservatively with respect to Bayes rule (Edwards, 1982)	Fairly robust - shown across multiple experiments in a highly controlled paradigm.	No, not necessarily - the finding is that participants update conservatively on all evidence, no matter which direction it points in.	Not necessarily - Corner et al. (2010) argue that conservatism bias may be an experimental artefact resulting from the fact that participants do not entirely trust the evidence given to them, rather than demonstrating irrationality.
Overconfidence: holding beliefs more strongly or with more precision than is rational (Moore et al., 2015)	Fairly robust - the overprecision form of overconfidence seems more robust than other types, but has also been less studied, so is harder to draw very strong conclusions about.	Again, by definition overconfidence is a kind of confirmatory reasoning - but its also unclear whether overconfidence is necessarily the result of certain confirmatory reasoning processes or not (since overconfidence has generally been studied independently of the processes leading to it.)	Not necessarily - saying how strong peoples beliefs should be requires using very specific experimental paradigms where normative standards are made explicit. A problem here is that this often means it is difficult to assess whether people are overconfident about the kinds of things we are typically most interested in (political beliefs say, as opposed to numerical estimates of quantities.)

Going through each of the tendencies we've discussed in this systematic way in table 2.6, we can see much more clearly just how tenuous the case for confirmation bias is. None of the findings discussed pass all three hurdles - demonstrating a robust tendency, that actually leads to confirmatory reasoning, and which can be shown to fall short of a clear normative standard. Perhaps those findings that fare best here are those related

to biased beliefs - findings of belief persistence and overconfidence seem to be the most robust of those we've covered. However, it's not clear whether (a) these tendencies are actually non-normative, particularly when it comes to issues where there is no 'correct' answer against which judgements can be compared, or whether (b) these tendencies are actually the result of the kinds of biased processes generally referred to as 'confirmation bias.'

The one dimension on which all of these findings fall short is the normative one - in every single case, normative standards are either not made explicit or have been contested for one reason or another. These normative issues - around how people should reason, what it means to be rational, and what constitutes a bias - are much more complex than they first seem, and I will discuss some of the disagreements that arise in more detail in chapter 4. For now, perhaps we might actually get a clearer picture of what is going on if we set these complex normative issues aside, at least temporarily - stop asking how people *should* reason, and instead just ask what we know descriptively about how people *do* reason. For example, we might consider the following research conclusions in a more descriptive sense, independent of any claims of 'confirmation bias':

1. People do seem to often use a positive-test strategy in testing hypotheses, asking questions for which the answer would be 'yes' if their current hypothesis is true (Wason, 1960, 1968). This means people err towards holding overly-narrow hypotheses, and it is easier to identify false positives than false negatives.

2. People do not necessarily always seek out whatever information would be most diagnostic, and often misjudge how diagnostic different pieces of information are, or are insufficiently sensitive to differences between diagnosticities of different pieces of information (Slowiaczek et al., 1992).

3. There is some evidence that people have a weak preference for opinion-supportive information in general (Hart et al., 2009), though this is highly dependent on and easily outweighed by other factors, such as how useful the information is for a given task.

4. People do seem to have difficulty considering alternative hypotheses, and certainly considering multiple alternative hypotheses at once, unless explicitly instructed to do so (Klayman, 1995, Toplak and Stanovich, 2003)

A few themes begin to arise here, and looking at these phenomena purely descriptively, we can ask why these tendencies might arise, what heuristics might underlie them. For example, difficulty considering more than one hypothesis at once might explain quite a lot: why people tend to interpret information as more supportive of the focal hypothesis than it in fact is (they only consider $\Pr(D \mid H)$ and not $\Pr(D \mid \neg H)$; an inclination to seek out more supportive information (such information will be more salient and easier to search for); and more readily accepting supportive information (with only one hypothesis in mind, supportive information is easier to interpret and make sense of, whereas information that conflicts with the focal hypothesis will take a lot more effort.)

It may also be true that people generally find it easier to reason with 'positive' information than negative information. As well as a positive test strategy in hypothesis testing, there is also evidence that both people and animals find it easier to learn when learning is based on the presence of features than their absence - supporting this idea that positive and negative evidence are treated differently. Hearst and Wolff (1989) found that pigeons learned twice as quickly when food would be available if this was indicated by the *presence* of a light than by its absence, and Newman et al. (1980) find similar results for human learning.

Rather than continuing to focus on the general tendency of confirmation bias, research might be better off focusing first instead on simply improving our descriptive picture of how people form and revise beliefs - what general principles and heuristics guide the search for information, the testing of hypotheses, and the inferences people draw from information. If we're able to get a clearer picture of how these processes work, uncomplicated by terms like 'bias' and 'rationality', then perhaps we can begin to ask normative questions - looking at the costs and benefits of different tendencies in different scenarios. This is not to suggest that it is not important and useful to ask normative questions; to ask where reasoning processes might perform better or worse - but that it might be helpful to more clearly separate out descriptive and normative questions. This would help to clarify some complex normative issues and ensure that our descriptive understanding is not confused by being too quick to draw unclear normative conclusions.

2.5 Conclusion

Having reviewed and discussed the evidence for confirmation bias of various types, it seems that the case is less convincing than it might seem. This is particularly due to the fact that (a) biases at different stages of information processing have largely been studied in isolation, whereas understanding confirmation bias requires understanding how processes of search and inference interact; and (b) most research on confirmation bias makes normative claims without engaging with or even acknowledging the complexities involved.

I suggested, therefore, that we might make more progress if we first simply tried to understand the kinds of processes and heuristics that underlie how people search for and draw inferences from information, in a purely descriptive sense, temporarily setting aside questions of how people *should* reason. Separately, we might try to clarify some of the complex normative issues related to confirmation bias - and then, armed with a better descriptive account of belief formation, we can begin to ask questions about bias and irrationality.

In this chapter, we have taken a broad overview of the literature related to confirmation bias. This raises questions both on a narrower level - what is really going on with many of the specific phenomena discussed as evidence of confirmation bias? - and on a much broader level - what does it really mean to be rational, or biased? The next two chapters will therefore attempt to look at some of the issues raised from both a narrower, and then a much more zoomed-out, perspective. In chapter 4, I will discuss some of the normative issues raised more broadly - different ways in which terms like 'rationality' and 'bias' have been understood, and different normative models in psychology - as well as the implications for the confirmation bias literature. Before zooming out, however, we will first take a narrower look at a specific aspect of confirmation bias where the literature seems particularly confused and problematic: selective exposure.

Chapter 3

The mixed evidence for selective exposure

3.1 Introduction

In the previous chapter, I discussed how findings that have commonly been cited as evidence for 'confirmation bias' face a variety of problems, suggesting the case for confirmation bias is much weaker than it might first seem. In this chapter, I will take a closer look at one of these findings in particular: the idea of 'selective exposure', that people prefer to seek out and read information that supports their existing views, rather than engage with alternative viewpoints.

The selective exposure hypothesis seems particularly worthy of further examination, because there is such a stark difference between its initial plausibility, and how difficult it has been to demonstrate experimentally. We still do not really understand the conditions under which selective exposure occurs, despite widespread acceptance of the idea that people prefer to read things they agree with. In particular, there has recently been more attention given to the idea that selective exposure online may be a driver of political polarization and conflict in politics more broadly (for example, Conover et al., 2011, Hsu, 2009, Iyengar and Hahn, 2009).

In this chapter, I will ask how we can square this impression that selective exposure is a driver of real-world problems, with the surprisingly mixed evidence for selective

exposure in the psychology literature. I will begin by reviewing the existing literature on selective exposure in more detail, building on the review from the previous section. I will look at the selective exposure literature from two more specific angles: looking at studies of selective exposure with political beliefs specifically, and those conducted in more 'naturalistic' environments (particularly online). In doing so, I will question whether there is any discrepancy between findings of 'selective exposure' across the board and findings of selective exposure in these more specific domains. Though we do find that selective exposure effects are slightly stronger when we look at research in these domains more specifically, overall findings are still very mixed.

I also conducted a series of online experiments using Amazon's *Mechanical Turk*, looking at selective exposure with respect to political beliefs. Results are very mixed: selective exposure effects do not always seem to occur, when they do they are relatively small, and seem to be very sensitive to subtle changes in the framing of the task.

Though mixed findings in the selective exposure literature are not new, what *is* new is that such mixed findings arise from such subtle changes to the same experimental paradigm (in the past, mixed findings arise as a result of looking at selective exposure in different contexts and for different topics.) Our findings are also novel insofar as they show little evidence of selective exposure even in a domain where selective exposure seems particularly likely to occur - with respect to political issues and beliefs. Taber and Lodge (2006) argue that selective exposure has been elusive so far in psychological research because experiments have failed to arouse enough of an 'affective response' in participants to motivate bias. They conduct two experiments using political materials designed to arouse a stronger partisan response, and *do* find evidence for a selective exposure effect. However, using the same materials with a larger sample and subtly changing how the information is presented to participants, we are unable to replicate this result - suggesting that even in this context, selective exposure is not particularly robust.

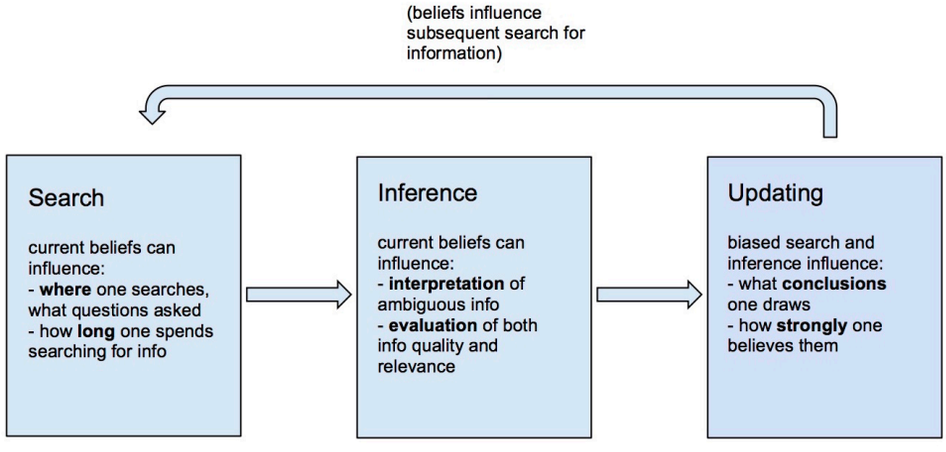
I discuss these results in the context of the broader 'replication crisis' in psychology (Open Science Collaboration, 2015), as well as discussing the implications for future study of selective exposure. I suggest that the root of the problem may be that: (a) if a 'selective exposure effect' exists, it is small and easily swamped by variations in

experimental design and in individuals, and (b) as a measure, selective exposure fails to capture what we are really interested in - the broader tendency of confirmation bias.

3.2 Background

The idea that people prefer information that confirms what they already believe has widespread acceptance both in psychology, and increasingly, in everyday thought. This idea goes back to Francis Bacon, who is often quoted as saying that, "the human understanding, when it has once adopted an opinion, draws all things else to support and agree with it." (Bacon, 1620)

The term 'selective exposure' refers to the psychological tendency to disproportionately seek out and pay attention to information supporting one's existing beliefs. As discussed in the previous chapter, selective exposure is generally thought of as a specific form of confirmation bias: the kind of confirmation bias that arises from how one searches for new information. Figure 2.1 (reproduced from the previous chapter) below shows a simple model of different stages of information processing, with biases possible at each stage - to clarify where selective exposure sits relative to other processes and the broader phenomenon. In what follows, we will repeat some of the earlier discussion of selective exposure for context, before exploring some specific areas in more detail.



Over the past 50 years, selective exposure has received a fair amount of attention in the psychological literature. In a meta-analysis of 91 studies across 67 papers, Hart et al.

find "a moderate preference for congenial over uncongenial information (d=0.36)." (Hart et al., 2009, p.555) However, the results of selective exposure research have been much more mixed than many realise, with many studies finding no effect or even the opposite effect; that people seek out more disconfirming than confirming evidence (e.g. Freedman (1965). This has led some to seriously challenge the idea that selective exposure is a robust phenomenon at all (particularly Sears et al., 1967). Though Hart et al's meta-analysis concludes that there is overall moderate evidence for selective exposure, the conditions under which the tendency does and does not arise are still not well understood. More recent research has mostly focused on identifying long lists of moderating factors, with little organization or theoretical understanding (see Hart et al., 2009, for an extensive summary of these moderating factors). Though understanding the conditions under which selective exposure does and does not occur is certainly useful, there is a risk here if the moderators are generated 'post hoc' and not theoretically justified. As Hahn and Harris point out, post-hoc justification of moderators can "make findings pitifully trivial" and may "prop up a bias that does not exist, thus obscuring the true underlying explanation."(Hahn et al., 2009, p.75) Though some of the moderators in the selective exposure literature do seem to have received prior theoretical justification, the long list of factors does not seem particularly unified or well justified theoretically. It's therefore not clear whether these moderating factors point to a genuine effect occurring under only specific circumstances, or whether they are merely propping up an effect that does not exist (as Hahn and Harris, 2014, caution).

The term 'selective exposure' is a broad one, and studies of selective exposure encompass a wide range of situations. In particular, the topic or issue covered varies widely: in some cases, people are asked to form an opinion about some matter that is not particularly important or relevant to them, such as whether a candidate is suitable for a job in a hypothetical scenario (e.g. Freedman, 1965); in others, participants are asked to report their pre-existing attitudes on issues in which they are likely to be much more invested, such as politics (Taber and Lodge, 2006). Another important difference is between studies that look at selective exposure for beliefs/attitudes, and selective exposure for *decisions* (i.e. whether 'supporting' information means 'information that supports a belief the person holds' or 'information that supports a decision the person has made.')

How people choose what information to pay attention to is likely to vary a great deal between these different scenarios. Selective exposure seems more likely to occur when

people are relatively invested in an issue, assuming that people engage in selective exposure in order to avoid challenges to their current positions. What motivates selective exposure might also depend on the situation. If I've already made a decision and can't go back on it, I might seek out validating information in order to avoid unproductive regret, whereas avoiding arguments that challenge my political beliefs might be motivated by a deeper-rooted fear of being wrong. Construing 'selective exposure' so broadly therefore may mask some important nuances - making it hardly surprising that we see such a long list of moderating factors. To understand selective exposure better, it might help to study selective exposure in these different contexts separately. While a meta-analysis that does not separate out these differences can give us insight into a very general tendency, it has limitations in what it can tell us about selective exposure in more specific scenarios.

The specific kind of selective exposure that is the focus here is selective exposure for well-established and fairly strongly held attitudes (as opposed to more transient opinions, or decisions), such as political opinions. In what follows I report the results of several online experiments exploring selective exposure with political beliefs, in order to better understand the extent to which selective exposure is a problem in this specific context. Before turning to the present research, however, I will give overview of the past research on selective exposure literature more generally.

3.3 Past research on selective exposure

I discuss past research on selective exposure in three main sections. First, I give a broad overview of the state of selective exposure research since it was first studied over 50 years ago, and more recent attempts to synthesise these findings, recapping and building on discussion in the previous chapter. I then consider two perspectives on selective exposure research which have not been explored in detail in past reviews: selective exposure specifically related to political attitudes, and more naturalistic studies of selective exposure. These later two sections attempt to supplement existing, broader, reviews by zooming in on two areas of selective exposure research with more practical relevance.

3.3.1 A broad overview

The notion of selective exposure first came to prominent attention as a result of Festinger's popular cognitive dissonance theory (Festinger, 1957). Festinger argued that, once people have committed to an attitude, belief, or decision, they tend to gather supportive information and neglect unsupportive information in order to avoid or eliminate the unpleasant state of 'cognitive dissonance'. A number of studies emerged in subsequent years supporting this idea. For example, in one of the earliest selective exposure studies (Adams and Stacy, 1961), mothers were asked whether they believed child development was predominantly influenced by genetic or environmental factors. When then given the choice to hear a speech advocating either position, mothers overwhelmingly chose the speech that favoured their view on the issue.

However, a review paper by Sears et al. (1967) cast doubt upon the selective exposure hypothesis. The review reported a number of experiments which failed to find a selective exposure effect - and even some whose results actively opposed it, finding people chose more conflicting than supporting information. In one experiment (Freedman, 1965), participants were presented with a tape of a mock interview portraying an interviewee in either a positive or a negative light. After judging the interviewees themselves, a majority of participants preferred to read another evaluation that disagreed with their judgement.

Sears et al. (1967) suggest that though there is ample evidence for *de facto* selective exposure - i.e. people tend to passively become exposed to information that supports their beliefs - there is little evidence for *active* selective exposure to attitude-congruent arguments when given the choice. The fact that people tend to get a lot of information from social media and recommendations from friends may make attitude-confirming information much easier to come by - but this doesn't mean people would necessarily prefer this when given an explicit choice between arguments on either side of an issue.

In the mid-1980s, two new review papers were published (Cotton, 1985, Frey, 1986), reopening the case for selective exposure. Cotton concludes that "dissonance-motivated selective exposure does appear to exist... Although the phenomenon has often been elusive and its support questioned, the research overall suggests that something is there", while Frey suggests, "shortcomings in experimental designs of previous research have

largely been responsible for the lack of conclusive results in the earlier studies." Supporting this perspective, in their meta-analysis of 91 studies across 67 papers, Hart et al (2009) find "a moderate preference for congenial over uncongenial information (d=0.36)"

What these reviews and subsequent attempts to make sense of the selective exposure literature all acknowledge is that confirmatory information search is by no means universal, and depends on a number of moderating factors. Subsequent research has therefore focused more on understanding these moderating factors. For example, Jonas et al. (2003) found that increasing awareness of death increased selective exposure, as did increasing the relevance of the issue to the person's worldview. Fischer et al. (2005), found that restricting the opportunity people had to select information increased selectivity. In their meta-analysis, Hart et al. (2009) found that selective exposure was more likely to occur when challenging information was expected to be high quality, when prior attitude commitment was higher, and for individuals who score high on the personality trait of closed-mindedness.

Review papers and meta-analyses attempt to understand 'selective exposure' as a very general tendency to seek 'confirming information' across a wide range of scenarios. Individual papers and studies take a much more focused approach and attempt to understand the extent to which selective exposure occurs in much more specific conditions - with these specific materials, this specific topic, this measure. There is a risk, however, that researchers (and the media and public) are too quick to draw very general conclusions from quite narrow studies. What may be helpful is something in between specific studies and broad conclusions: picking a narrower class of situations in which selective exposure might occur and be problematic (with respect to certain important attitudes, or a certain important type of decision most people make, for example), and then doing a thorough investigation into the extent to which selective exposure occurs here.

3.3.2 A narrower look at selective exposure - social and political attitudes

How someone seeks out information related to a decision they just made (whether to buy a car or not, say), seems like a very different phenomenon to how someone seeks out information related to an attitude they have held for a long time (certain political

or religious attitudes, for example.) Similarly, how someone seeks out information related to a long-held, well-established attitude also seems likely to be very different from how they seek out information related to a judgement they just made (about whether a hypothetical person should be hired for a hypothetical job, say.) Classing these together under the heading 'selective exposure' - as the Hart et al. (2009) meta-analysis does, allows us to look at very general information-seeking tendencies, but may mask differences in these different types of selective exposure.

Here we focus on selective exposure with respect to social or political attitudes that are well-established, reviewing studies that look at this specific type of selective exposure (as opposed to those focusing on decisions/behaviour, or relatively trivial attitudes formed within an experimental context.)

Of those 67 papers included in Hart et al. (2009), 13 fit this criterion: asking people about their attitudes on social, political or religious attitudes they already hold, and then looking at how people seek out information with respect to these attitudes. These studies generally do seem to find that selective exposure occurs to some degree - but with many finding that this tendency is easily influenced by subtle moderating factors. Many of the papers explicitly acknowledge that the selective exposure evidence has been mixed; propose a hypothesis for why evidence for selective exposure has been elusive; and then set out to test that hypothesis. Below I discuss some of the evidence for selective exposure of this narrower kind, and some of the factors proposed to moderate the effect. The studies cited are mostly selected from those included in Hart et al's meta-analysis, though I also discuss a few additional studies which look at selective exposure in relation to political attitudes, which do not feature (particularly Taber and Lodge, 2006). A table with more details of the 13 studies selected from Hart et al's meta-analysis is available in appendix B.

3.3.2.1 Strength of opinions and dissonance

Brannon et al. (2007) hypothesise that selective exposure may not occur because participants' views are not sufficiently strong to motivate defensive processing. They ask participants about their opinions on a range of social issues (including abortion, the death penalty, and international conflict), and then ask them to indicate on a nine-point scale how desirable it would be for them to read different articles about these topics.

Participants were presented with article titles which indicated what position the article would take on the issue without giving away any more of their content. Results found that participants indicated significantly greater interest in reading articles consistent with their views - and that this effect was greater the stronger the person's attitude was.

Cotton and Hieser (1980) similarly claim that "many of the discrepancies in the results of previous studies of the phenomenon of selective exposure can be traced to the use of inadequate design" - in particular, suggesting that many selective exposure experiments may not create enough dissonance to arouse any kind of biased information processing. They try to arouse a greater sense of dissonance by having subjects begin by writing an essay for a position they disagree with on the topic of nuclear power plants - and manipulating the extent to which they felt like they had a 'choice' in doing so. Having done this, subjects then indicated how interested they would be in reading different pamphlets, and how much they'd like to be in a discussion group exchanging information in favour of or against the issue. These were framed as real, not hypothetical choices - subjects actually expected to be able to receive the pamphlets and to participate in the discussion groups. Cotton and Hieser (1980) found that when subjects felt they were forced to write a counterattitudinal essay, they were significantly more interested in pro-attitudinal pamphlets and discussing with people who agreed with them - but the same effect did not occur when they felt they had chosen to write such an essay. So though, in some sense, this study did find evidence of selective exposure, it is only under very specific conditions - when people are first asked to write an essay they disagree with, and when they feel like they did not have a choice in doing so.

Taber and Lodge argue, similar to Cotton and Hieser, that evidence for selective exposure has been elusive because "the arguments and evidence used in many of these studies failed to arouse sufficient partisan motivation to induce much biased processing." (Taber and Lodge, 2006, p.756) They therefore use statements and arguments taken directly from political interest groups (on the topics of affirmative action and gun control), which are expected to be more contentious and therefore arouse more biased processing. The experiment finds a significant selective exposure effect - subjects choosing to read more supporting than opposing arguments across the board, with the effect being particularly pronounced for those who were more politically sophisticated/knowledgeable (as

measured by a political knowledge survey earlier in the experiment), and for those with stronger prior views.

3.3.2.2 Goals and motivations

Clarke and James (1967) looked at how different goals might affect selective exposure - speculating that if subjects expect to participate in a debate they may be more likely to seek out supportive information than if they expect to simply discuss the issue in a relaxed setting. Participants completed a 50-item attitude questionnaire containing Likert-type items for a variety of issues, including political, moral, religious, and career questions. They were then shown magazine articles with titles indicating arguments for either side of different issues, and asked how much they wanted to read each of the issues. Results found that participants were overall significantly more interested in reading articles consistent with their view, but that there was no significant difference between the debate and discussion groups.

Though Clarke and James (1967) do not find any effect of different goals on selective exposure, a later study by Lundgren and Prislin (1998) finds that people engage in selective exposure when given a 'defense' motive, but not with an 'accuracy' motive, an 'impression' motive, or when given no explicit motive. In the 'accuracy motivation' condition, participants were told that the purpose of the study was to examine their logic and reasoning abilities; in the 'impression motivation' condition subjects were told they would be evaluated on their agreeableness and other rapport skills; in the 'defense motivation' condition subjects were told their opinions were being surveyed to help a board decide about the issue; and in the final, control, condition, subjects were given no explicit goal. When given the opportunity to read arguments on either side of the issue, only the defense motivation group spent more time reading counterattitudinal articles (i.e. the opposite of selective exposure), with no significant difference found for any of the other groups.

Smith et al. (2007) also find that giving subjects an explicit goal affects selective exposure. People who were told they would later be asked to give a speech which would be recorded and shown to others who shared their views, were found to engage in significantly more selective exposure than those who were not expecting to give such a speech. They also found that when under time pressure, people engaged in significantly more

selective exposure - and that those who had *both* the speech goal and time pressure were the most likely to selectively expose themselves to supportive information.

3.3.2.3 Personality differences

Feather (1969) suggests that mixed selective exposure results may be due to the fact that studies don't allow for personality differences, particularly differences in tolerance for ambiguity/uncertainty. Feather also hypothesizes that people may be more likely to engage in selective exposure when information is expected to be novel, since novel information is more likely to be threatening. To test these two hypotheses, participants complete an attitude scale about the issue of American intervention in Vietnam, as well as a scale measuring intolerance of ambiguity and dogmatism. Subjects then list arguments they know for and against American intervention in Vietnam (so that arguments can later be identified as either novel or familiar, depending on whether it had already been listed.) Subjects are then asked to rate their interest in reading different arguments about the issue. Overall, participants indicated significantly higher interest in consistent than in inconsistent information, and this difference was greater under the hypothesized conditions - when dogmatism/intolerance of ambiguity was higher, and when the information was expected to be novel.

More recently, Lavine et al. (2005) looked at a similar hypothesis - that selective exposure depends both on (a) personality differences, particularly what they call 'authoritarianism', and (b) the extent to which someone feels threatened by new information. They have 145 undergraduates complete a questionnaire which measures their political party identification, views on a range of political issues, and a right-wing authoritarianism scale. Half the subjects are then asked to "describe the emotions that the thought of your own death arouses in you" - intended to create a feeling of threat by increasing mortality salience (Jonas et al., 2003). Subjects then rate their interest in reading three different articles on capital punishment on a scale of 1 to 7 - one containing arguments in favour of the policy, one containing arguments opposed to the policy, and one containing a mix. High authoritarians were found to express significantly greater interest in the attitude-congruent message relative to the incongruent message when threat was high than when it was low, but this effect did not occur for low authoritarians. Overall,

a selective exposure effect was only found for subjects who scored high on authoritarianism who were in the mortality salience condition, with no preference for supporting information found in any of the other conditions.

3.3.2.4 Other factors influencing information choice

Hillis and Crano (1973) attempt to control for the *perceived utility* of information when measuring selective exposure - noting that in past studies, people may have selected more supportive information expecting it to be more useful. The question asked should therefore be, "do people prefer supportive information to *equally useful* conflicting information?", rather than simply, "do people prefer supportive to conflicting information?" After completing a questionnaire about attitudes on several issues including abortion, subjects are told that they will be asked to present either a pro-choice or pro-life speech, and that a series of arguments favoring and opposing abortion would be available to help them prepare. This goal helps to establish which kind of information is likely to be most useful to participants, independently of their prior attitude - some subjects were asked to prepare a speech supporting the view they had expressed in the attitude questionnaire, and others to prepare a speech for the other side. Results indicated that information utility for the task was a stronger determinant of which articles participants viewed than selective exposure - overall, participants viewed more slides containing information necessary for the task irrespective of whether it supported their prior view or not.

Valentino et al. (2009) find that provoking anxiety (by asking subjects to recall an event related to the current political campaign that made them anxious) boosts more balanced information seeking, but only when subjects were expecting to have to explain and defend their opinion. Similar to Hillis and Crano (1973), this suggests that people are willing to seek out conflicting viewpoints when they expect doing so to be useful to them - and that this utility of conflicting information can outweigh a motive for defensive processing.

Messing and Westwood (2012) find that social endorsements can also outweigh any desire to read attitude-consistent information. Noting that a great deal of political news and information is now acquired through social media, they hypothesize that referrals and endorsements from people we know may play an important role in determining

information selection. In an experimental context, they find that social endorsements do increase the probability that people will select to read an article, and that the presence of such social endorsements reduces selective exposure tendencies.

Overall, the evidence for selective exposure for social and political topics seems slightly stronger than selective exposure more generally - with most studies finding some evidence of selectivity, at least under certain conditions. This is not necessarily surprising, since we would expect political attitudes - often long-held and emotionally charged - to be exactly kind that would motivate biased search for information.

However, even in this narrower sense, the case for selective exposure is not totally unproblematic. Whether people tend to prefer pro-attitudinal information seems to depend on quite a few moderating factors - in the absence of which we see no evidence of selective exposure. For example, Lavine et al. (2005) find no evidence of selective exposure when the 'mortality salience' exercise is not included - of four conditions (2x2 design - mortality salience vs control, high authoritarianism vs low authoritarianism), only the one where participants were both high in authoritarianism, and made to think about their death, did any evidence of selective exposure occur. Hillis and Crano (1973) found selective exposure only when people were explicitly given the task of presenting the case for the side they disagreed with. Feather (1969) found that the selective exposure effect was much weaker when participants were low in dogmatism, and that it virtually disappeared when information was not expected to be novel. Lundgren and Prislin (1998) found a small effect in the opposite direction (a preference for conflicting rather than supporting information) when other motives such as being accurate or engaging in agreeable discussion with others were made salient. Smith et al. (2007) found much weaker evidence for selective exposure when subjects were not given an explicit goal and were not under time pressure. All this suggests that although there is some evidence of selectivity in these studies, this tendency can easily be eliminated or even reversed with relatively subtle changes to the experimental context.

Existing research on selective exposure with social/political attitudes has a few other limitations. Many of the studies were conducted decades ago, and many with relatively small samples - Hillis and Crano had a sample of 123 college undergraduates across eight different conditions (so 15 per condition); Feather had 158 students across four conditions (40 per condition); Cotton and Hieser had only eight participants per condition;

and Lundgren and Prislin had only 16 subjects per condition.

There is also substantial variation in the design and measures used across these studies. There seems to be no standard way to assess selective exposure - with some studies simply asking people to indicate on a numerical scale how interested they would be in reading various different articles (including Brannon et al., 2007, Clarke and James, 1967, Feather, 1969, Lavine et al., 2005), and others having subjects actually make choices between and read information (Hillis and Crano, 1973, Lundgren and Prislin, 1998, Smith et al., 2007). As some authors have noted, we should perhaps be wary of drawing generalisations about how people actually seek out and select information based on the preferences they express in hypothetical scenarios.

Even when we actually observe subjects' decisions (rather than asking about preferences or hypothetical scenarios), these decisions are quite artificial: rarely in real life are we explicitly presented with a list of different articles on two sides of an issue and asked to choose between them. It therefore also seems worth looking at whether we can learn about about selective exposure in more 'naturalistic' settings - as some recent research has begun to do.

3.3.3 Naturalistic and field experiments

In recent years, new research has begun to emerge looking at selective exposure in more 'naturalistic' settings - making use of survey data and/or online behaviour. In combination with lab experiments of selective exposure, this can help to shed some light on when and to what extent people tend to actively seek out, and/or become passively exposed to, information that validates their already held beliefs.

Stroud (2007) argues that topics like politics are more likely to inspire selective exposure than others, and that research needs to look more at habitual media exposure patterns, rather than single decisions (as is often done in the lab.) Stroud uses data from the 2004 *National Annenberg Election Survey* which asked people about both their political leanings and their habitual media use (including newspapers, political talk radio, cable news, and political websites), to investigate whether views influence media use. The data suggests such an influence - 64% of conservative Republicans report consuming at least one conservative media outlet, compared to 26% of liberal Democrats (with

some survey respondents not consuming any political media at all), and 76% of liberal Democrats report consuming at least one liberal outlet while only 43% of conservative Republicans say the same. Similarly, Gil de Zúñiga et al. (2012) use US national survey data (collected between December 2008 and January 2009 by a research unit at the University of Texas Austin) and find that the more conservative a person is, the more inclined they will be to watch Fox news (r=.38, $p < .0001$), and the less likely they will be to watch CNN (r=-.18, $p < 0.001$) - and that correspondingly, the more liberal a person is, the more likely they will be to watch CNN and the less likely to watch Fox News.

A general limitation of using survey data like this is that it is often only correlational. This means that even if we find a relationship between political views and media consumption, we can't conclude that people's political views are definitely *causing* them to seek out certain kinds of media - it's also possible that instead exposure to certain media sources leads people to form the corresponding political views. Though Gil de Zúñiga et al. (2012) suffer from this problem, Stroud (2007) does use a strategy of panel analyses to attempt to determine causality. By including a lagged measure of the dependent variable (media exposure) it is possible to evaluate whether the independent variable (political views) has a causal effect on the dependent variable. The survey data used in this study contains measures of media use at two different times during the 2004 presidential campaign, which makes this kind of analysis possible. Stroud finds that people's political beliefs are significant predictors of what media outlets they suggest at the later time, even after controlling for their selections at the earlier time: providing more evidence for a causal effect of political attitudes on information selection.

Conover et al. (2011) look at networks of political communication on Twitter - amassing over 250,000 tweets from the six weeks leading up to the 2010 US congressional midterm elections. They show that the network of political retweets exhibits a "highly segregated partisan structure, with extremely limited connectivity between left- and right- leaning users." (Conover et al., 2011, p.89) This suggests both some degree of active selective exposure - people are more likely to engage with those who have similar political views to them - and passive selective exposure - due to the nature of these networks, people will end up more easily exposed to supportive viewpoints even without displaying an active preference for them. Himelboim et al. (2013) also look at connections on Twitter among users talking about the US president's state of the union speech in 2012.

They find that users participate in "fragmented interactions and form divided groups, in which people tune into a narrow segment of the wider range of politically oriented information sources." (Himelboim et al., 2013, p.195) Groups and networks seem to form in such a way that people generally expose themselves to sources and information that disproportionately support what they already believe.

Several studies also report quasi-experimental evidence looking at how people select different information, particularly news sources. Iyengar and Hahn (2009) find that, when given an explicit choice, conservatives tend to prefer to read news reports attributed to Fox News and to avoid news from CNN and NPR, while liberals exhibit the opposite preference. Here, it seems like the source of information is key - people aren't simply choosing to read arguments they expect to agree with, but rather are choosing to read arguments from sources they trust or like (perhaps indirectly because those sources tend to agree with them.) There is a complex question here of when it is rational for me to distrust a certain source of information, which we will return to later.

Garrett (2009) conducts an online study with subjects recruited from two partisan news websites, and tracks their choices of news items and time spent reading (with a sample of 727 subjects, substantially larger than many past selective exposure studies.) Results indicate weak evidence for selective exposure - the news articles that subjects selected were more likely to be opinion-reinforcing than those they did not select, but there was no evidence that people made any active effort to avoid opinion-challenging information. Building on this, Garrett et al. (2011) challenge the claim that avoidance of opinion-challenging information is becoming increasingly common over time. They show, using data from a series of national RDD surveys between 2004 and 2008, that Americans' use of attitude-consistent political sources is positively correlated with the use of more challenging sources. Though people may be seeking out more and more information that confirms their beliefs, but they are not necessarily driven to avoid attitude-challenging information they may simply be seeking out more information on the topic in general.

3.3.4 Summary: the state of selective exposure research

'Selective exposure' has generally been construed as a very broad phenomenon - with 'confirming information' referring to anything that reinforces something one believes, a decision one has made, or past behaviour - all ranging from the fairly trivial to the

much more consequential. Though it's certainly interesting and useful to ask whether a selective exposure tendency exists in this very broad sense, construing selective exposure so broadly masks a lot of variation and subtle moderating factors - and so it is perhaps not surprising that findings have been so mixed. I suggested that it might help, therefore, to ask whether selective exposure occurs when defined in slightly narrower ways - particularly focusing on cases where selective exposure is particularly likely to be important or problematic.

I looked specifically at cases of selective exposure with respect to political attitudes. This seems like a context where selective exposure may be particularly likely to occur - if we believe it is motivated by a desire to defend existing beliefs that are personally important - and also particularly likely to be problematic - given the importance of these issues. Looking just at those studies of selective exposure which look at political attitudes, there is some evidence for an effect, but the effect is sensitive to factors such as the strength of belief, personality differences, and the goals and motivations present. More research on selective exposure in political contexts, with larger samples and clearer outcome measures, could help further understanding here.

There is also a question of the best way to study selective exposure, if ultimately we are interested in how people form and change their opinions in 'real-life'. Selective exposure has generally been studied in abstract, experimental contexts, where people make selections between different sources of information in a fairly contrived way. This looks at 'active' selective exposure: whether people display an explicit preference for opinion-supporting information when given the choice as opposed to more 'passive' selective exposure: whether people end up indirectly exposed to more supportive information, whether through any explicit choice of their own or not. Recently, more research has arisen looking at selective exposure in this more passive sense, by studying the information people end up exposed to in their day-to-day lives. Surveys of media use patterns as well as data on online behaviour suggest that people do end up exposed to more information that supports rather than conflicts with their political views though whether or not this is because people have an active preference for supporting information, or due to more indirect factors (like the way information is structured in the environment or how people get information from those around them), is still unclear. The focus of the following studies, therefore, is to better understand the extent to which selective exposure arises with respect to political beliefs in this more 'active' sense.

3.4 The present research

The current research aims to shed more light on the extent to which active selective exposure occurs for political attitudes. Though many selective exposure studies have been conducted over the past decades, relatively few look explicitly at political attitudes. Of those that do, results have been mixed - finding that selective exposure occurs but only given certain moderating factors or manipulations. Sample sizes used have also generally been quite small, and many studies measure selective exposure by asking participants to indicate their interest in reading different articles, rather than measuring what information people actually choose to read when given a choice.

The following studies therefore aimed to investigate to what extent selective exposure occurs with respect to political beliefs in a relatively simple paradigm, before looking at a moderating factor which may explain some of the discrepancies in results. We begin by setting up and making some modifications to the basic paradigm, before attempting to replicate one of the most recent demonstrations of selective exposure with respect to political beliefs (Taber and Lodge, 2006).

3.4.1 Experimental paradigm

The general paradigm used as a basis for the experiments reported here, is as follows:

1. Participants begin by answering questions about their attitudes on a chosen issue. Typically this involves questions designed to measure attitude position (i.e. what they believe) and attitude strength (how strongly they believe it) independently.

2. Participants are told they will have the opportunity to learn more about the issues by reading some arguments arguing for the two different sides. They are told that they can read some limited number of the available arguments (say, 4 out of an available 8.) The available arguments are balanced equally across the two positions, so it is up to subjects to determine what proportion of supporting/conflicting arguments they want to read.

3. After reading each arguments, subjects are asked to rate how convincing/high quality they thought the argument was, on a sliding scale (1 to 10).

4. Often participants are then also asked to re-answer the same attitude questions that they answered at the beginning, to establish whether their beliefs on the issue changed at all - or sometimes simply asked to report subjectively whether they felt any of the arguments changed their beliefs at all.

For analysis, a participant's 'bias score' is calculated as the number of supporting arguments selected minus the number of conflicting arguments selected. Positive scores indicate selective exposure - i.e. a tendency towards selecting more supporting arguments. Secondary outcome measures often used in these studies include participants ratings of how convincing the arguments were - giving some measure of whether biased evaluation occurs alongside selective exposure - and whether or not participants actually change their mind on the basis of the reading they did.

Some studies of selective exposure have not asked subjects to actually read information, but instead have asked them to rate how interested they would be, hypothetically, in reading different arguments. We chose to measure selective exposure based on actual information selections, since we believe this gives a more reliable measure of actual behaviour, more likely to generalise to real-life information choices. Overall, we started with a very simple paradigm which was then improved in a few small ways as we learned from the first experiments. However, this measure and the general paradigm is not without its downsides, which I discuss later.

3.4.2 Experiment 1: setting up a selective exposure paradigm

3.4.2.1 Background

This first selective exposure experiment had two aims: (1) to test an experimental paradigm that could be used in future studies, and (2) to test a specific hypothesis about the conditions under which selective exposure might be more or less likely to occur, building on the idea discussed previously that goals/motives seem to be an important determinant of selective exposure (Clarke and James, 1967, Lundgren and Prislin, 1998, Smith et al., 2007).

I hypothesized that people would engage in less selective exposure if they felt they were being asked about their political *knowledge* than their political *opinions*. When

people are asked about their personal opinions, the assumption is that people will engage in selective exposure because they are motivated to defend those opinions, and seeking supportive information is a good way to do that. However, if people are asked about their knowledge on a topic - and particularly if they expect to later be tested on that knowledge in some way - they may be more motivated to seek out conflicting views in order to ensure their knowledge is accurate. This builds on the work of Lundgren and Prislin (1998) who find that when participants were told the purpose of the task was to express their opinions to help a board make a decision, they chose to read significantly more supportive arguments, but that when participants were told they were being assessed on their logic and reasoning abilities, no such selective exposure effect occurred. The authors suggest that in the first case, subjects were motivated to *defend* their position, leading to selective exposure - whereas in the latter case, subjects were motivated to be *accurate*, leading to more balanced information search.

This study therefore looks at whether the knowledge/attitude distinction described is sufficient to evoke different motives and therefore change information-seeking behaviour. In particular, does framing political issues in a more objective way, in terms of political knowledge, increase people's motivation to be accurate, and therefore reduce selective exposure? This hypothesis is also particularly interesting because, if supported, it could provide useful practical suggestions for how to frame political communications in order to minimize defensive processing.

3.4.2.2 Design and procedure

The experiment was conducted using Mechanical Turk, and we recruited 196 US subjects, roughly equal numbers of male and female (100 male, 96 female.)

The design followed the same general method outlined above. Participants were asked questions about four political issues: income inequality, minimum wage, death penalty, and gun control - before being given the opportunity to read arguments from opposing sides on these arguments. Each participant was able to choose 4 arguments out of a possible eight (four on each side of the issue), based on a simple sentence summary of each argument which made clear which side it was arguing for. Our main outcome measure was the average 'bias score' across all participants for each topic - the difference between the number of supporting and conflicting arguments chosen.

To test the hypothesis that framing questions in terms of facts/knowledge would induce less selective exposure than framing questions in terms of personal opinion, participants were randomly allocated into one of three conditions:

1. **'Opinion' (control) condition**: participants were told they were participating in a survey of their opinions on political issues, and asked whether they personally believed that certain statements about political issues were true, such as "Do you believe that abolishing the minimum wage would benefit society?"

2. **'Knowledge' condition**: Participants were told that they were going to be tested on their knowledge of different policy issues, and asked simply whether certain statements about political issues were true, such as "Would abolishing the minimum wage benefit society?" These were phrased so as to be as similar to the questions used in the control condition as possible, so that the only difference was whether the participants personal opinion on the topic was emphasised or not.

3. **'Knowledge and answers' condition**: This was identical to the knowledge condition, except that participants were also told that at the end they would be given a summary of the existing evidence on the topics, against which to check their understanding. Here it was hypothesized that expecting to have one's knowledge later 'tested' might create a further incentive for participants to seek out disconfirming viewpoints.

For each of the four policy issues, all participants answered one 'empirical' question, about the effects of a certain policy, and one broader 'value judgement' question, about whether they thought the policy was a good idea overall. In the first 'opinions' condition, participants were told at the beginning of the study that the aim was to find out their opinions on these policy issues. In addition, all questions were framed purely in terms of the participants' opinions - even the empirical questions framed as "do you believe that policy x has effect y?." In the second, 'knowledge' condition, participants were told that they were going to be asked both about their knowledge and opinions on various policy issues. Questions of fact and values were clearly distinguished, for example, "Your knowledge: what is the effect of policy x on y?", and "Your opinions: do you believe that policy x is a good idea?" In the third condition, the 'answers' condition, participants were given exactly the same instructions as the 'knowledge' condition, except that they

were also told at the beginning that they would be told 'answers' at the end - the current state of the evidence on each of the issues.

We hypothesise the following:

1. A selective exposure effect: participants will, on average, choose to read more supporting arguments (as measured in relation to their answers to the attitude questions at the beginning), than conflicting arguments.

2. This selective exposure effect will be lower in the 'knowledge' condition than in the 'opinions' condition, and lower again in the 'knowledge and answers' condition than in the knowledge condition.

3.4.2.3 Handling of moderates

A slight complexity in this analysis arises from the fact that, when participants were asked for their initial opinions on the different topics, their options were either 'pro', 'against' or 'not sure/no opinion.' For those participants who answered 'not sure/no opinion', we therefore cannot say that they chose arguments that either supported or conflicted with their initial opinion. In analysing the data, we have two options: (a) to exclude these participants from analysis, and look only at selective exposure for those participants who expressed an opinion in one direction or the other, or (b) to include those participants who expressed no opinion, with a bias score of zero by default.

Feldman et al. (2013) discuss the handling of 'moderates' - those without a clear ideological preference - as one issue selective exposure researchers have to deal with which can potentially affect results. They discuss two main approaches for dealing with moderates - either 'forcing' them into positions in how the questions are framed, or excluding them altogether - and find, unsurprisingly, that excluding moderates produces a larger estimate of the frequency of selective exposure. The authors note that excluding moderates provides a conceptually clear estimate of how many opinionated subjects prefer like-minded information sources, but does not provide a useful estimate of the likelihood with which selective exposure occurs among all people - since many citizens have moderate opinions.

In this case, we did not phrase the question in such a way that moderates could be 'forced' into a position so we opt to exclude those who indicated they were 'not sure' from the main analysis. Bias scores and analysis reported are therefore only for those subjects who indicated a view on one side or the other of the issue. However, we also include some analysis of moderates separately as a kind of 'control' - checking whether moderate participants show roughly the same pattern of argument choices as participants with stronger views.

3.4.2.4 Results

After removing any participants who did not complete the experiment or who failed 'attention checks', we are left with 195 subjects who completed all questions on income inequality (of which 28 were 'moderates' - expressing no initial opinion), 172 who completed all questions on the minimum wage (of which 32 were moderates), 168 who completed all questions on the death penalty (of which 50 were moderates), and 161 who completed all questions on gun control (of which 21 were moderates.) [1]

Recall that participants were randomly allocated to one of three conditions: a control ('opinions') condition, or one of two treatment conditions ('knowledge' or 'knowledge and answers'.) The below table shows the breakdown of participants in each condition and some key demographic data.[2]

	Sample size	Gender (%F)	Age (mean)	Income (mean $)
Opinion	66	48%	39.5	41,841
Knowledge	50	58%	43.4	45,260
Answers	80	50%	35.3	51,831

TABLE 3.1: Demographic characteristics, experiment 1

[1] For each topic, we include in the analysis all participants who completed all the questions for *that topic* - even if they did not complete the whole experiment - therefore essentially looking at four different datasets, one for each of the four topics.

[2] Note that the size of the three conditions is somewhat uneven - perhaps due to the relatively small sample size. This is worth bearing in mind when interpreting the results.

A one-way ANOVA finds no significant difference in the mean 'bias score' (no. of supporting arguments - no. of conflicting arguments selected) between the three conditions, for any of the four topics. (income inequality: F(2,164)=0.27, p=0.76; minimum wage: F(2,137)=0.25, p=0.78; gun control: F(2,137)=0.59, p=0.56; death penalty: F(2,115)=0.69, p=0.51). This suggests that the effect of our intervention was not as hypothesised: the framing of the experiment did not affect the arguments people chose to read.

We can also look at the average bias scores for each of the four topics for each group[3], and ask whether they differ significantly from zero (i.e. do people choose significantly more supporting/conflicting arguments?)

When we look at the average bias scores across all four topics for each group, we find that the average bias score for the opinions condition is significant (mean = 0.53, t=2.68, $p < 0.05$ - see table 3.2 below), but not for either of the other two conditions. We therefore find some evidence of a selective exposure effect when people are asked about their opinions on political issues, but no evidence of such an effect when people are asked about their knowledge on those same issues. This might be interpreted as support for our initial hypothesis, but given the non-significant ANOVA results, this seems weak evidence at best.

The pattern is also less clear when we look at the breakdown of bias scores by topics. We dont see the same pattern holding for peoples information choices for the separate topics, but only when we average choices across them all. Looking across all three conditions, we find that the bias score only significantly differs from zero for the topic of income inequality (mean bias score = 0.57, t=3.57, $p < 0.001$). Within the topic of income inequality, the bias score is significant for those in the opinions condition (mean = 0.63, t=2.34, $p < 0.05$), and the answers condition (mean = 0.66, t=2.56, $p < 0.05$), but not in the knowledge condition (mean = 0.37, t=1.14, $p = 0.2627$.) That is, we find evidence of selective exposure in the opinion and answers conditions, but not in the knowledge condition. This is somewhat surprising, given that the knowledge and answers conditions were identical except for the fact that those in the latter group expected to be given factual answers to some of the questions at the end, which we hypothesized would make

[3] Here, we take the 'bias score' of each participant for each topic, and average the four values to get an 'average bias score' for each participant. These average bias scores are then averaged across individuals in each condition, to get an overall bias score for each condition.

people more likely to seek balanced information, not less. For the other three topics, bias scores do not significantly differ from zero for any groups: there is no evidence that people display any consistent preference for supportive over conflicting information.

One way to interpret this might be that the manipulation does subtly influence selective exposure, ut the extent to which selective exposure occurs also depends on certain features of the topic/issue - so that when we look across several topics the influence of the framing becomes apparent, but for individual topics the influence of the framing may be outweighed by factors specific to the topic.

Finally, we note that when we look at average bias scores across all three conditions, we only find a significant bias score for one of the four topics, and when averaged across all four topics, the mean bias score (0.23) does not significantly differ from zero (t=1.68, p=0.095). This suggests that, though selective exposure seems to occur under certain conditions, this study provides little evidence that it holds as a broad tendency across a wide range of scenarios.

	Income Inequality	Minimum Wage	Gun Control	Death Penalty	Average - all topics
Opinion	0.63*	0.22	0.34	-0.12	0.53*
Knowledge	0.37	0.00	0.29	0.38	0.18
Answers	0.66*	-0.07	-0.03	-0.11	0.07
Average	0.57***	0.05	0.20	0.05	0.23

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.2: Bias scores by condition and topic, experiment 1

3.4.2.5 Analysis of moderates

We mentioned above that due to the phrasing of the initial questions, some participants get classed as 'moderates' - having no opinion either way on the topic. This meant we could not use them in the analysis above, since their choices could not be classified as either 'supporting' or 'opposing' their prior view.

However, we can still look at the argument choices of those with moderate opinions, and compare them to the choices made by those who *did* express opinions on either side of the issue. We look at the number of 'pro' arguments chosen for each topic by those who initially expressed an opinion for or against the issue, compared with those who expressed no opinion. If selective exposure occurs, we should expect those initially in favour to select more 'pro' arguments than those who were initially against. Those who did not express an opinion can act as a kind of 'control' group here.[4]

	Income Inequality	Minimum Wage	Gun Control	Death Penalty
In favour	2.23	1.94	2.11	1.91
Against	1.54	1.95	1.94	2.08
No opinion	2.14	2.09	2	2.04

TABLE 3.3: Average number of 'pro' articles selected by initial opinion, experiment 1

The difference between the number of 'pro' articles selected in the three groups is significant for income inequality (one-way ANOVA, F=7.004, p¡0.05), but not for minimum wage (F=0.26, p=0.77), gun control (F=0.59, p=0.56), or the death penalty (F=0.53, p=0.59.) In general, moderate participants seem to show roughly the same pattern of argument choices as those with expressed opinions.

3.4.3 Experiments 2 and 3: how robust is the evidence for selective exposure?

3.4.3.1 Background

In the previous experiment, participants were presented with sentence summaries of the different arguments, which clearly indicated which side they were arguing for - such as "Some inequality may actually be needed to promote economic growth, so income inequality is not a problem", or "Income inequality concentrates political influence in the hands of the elite, so income inequality is a problem." However, though this approach has standardly been used in many selective exposure experiments, it fails to control for

[4]By 'pro' arguments here we mean more specifically arguments that: 'income inequality is a problem', 'we should keep the minimum wage', 'gun control should be increased', or that 'we should use capital punishment.'

features of these descriptions other than which 'side' they are on, which might influence people's choices - some sentence summaries may simply sound more interesting or appealing in other ways, and some may be more interesting to some participants than others. We might have avoided this by having different participants independently rate how interesting each argument sounded to them based on the summary beforehand - if there are no systematic differences across these ratings on average, then we should not expect participants' choices between them, independent of bias, to lean in one direction, on average. However, even simpler would be to present participants with arguments in a way that gives them as little information as possible beyond the 'direction' of the argument, which is the approach we take in the next study.

3.4.3.2 Design and procedure

First, we ran the same experiment again but (a) with no intervention - all participants are asked about their opinions on the issues, as in the 'control' condition of the previous experiment; (b) participants were given a choice between reading "an argument in favour of/against gun control/minimum wage", rather than reading sentence summaries of the arguments, in order to control for other factors in those summaries that might influence choice. Finally, we used a slightly more complex measure of opinions (more details in the appendix) - asking two different questions about each topic, and allowing answers on a scale from -3 (strongly disagree/oppose), to 3 (strongly agree/support.) This allows us to capture slightly more variation in initial opinions, rather than simply categorising every participant as either for, against, or unsure. It also 'forces' more subjects into expressing an opinion in one direction or another [5] Otherwise, the design of the experiment is the same, and participants are again recruited using Amazon's Mechanical Turk. 120 US subjects took this survey.

We also run a second experiment, identical in design and procedure, recruiting participants using the UK market research platform Bilendi - to see if the same results hold for a UK population. In this experiment, we focused on just two topics: minimum wage and gun control, since these were the topics for which the original questions and statements could easily be transferred to a UK audience. 246 British subjects took this survey.

[5] Though it is still possible to express a neutral opinion, by choosing "neither agree nor disagree" for both questions. In practice, very few participants do this -fewer than 10 - but we again remove these participants from the final analysis of bias score. Since the number of 'moderate' subjects is so small, we do not include them in any further analysis.

In all other ways the procedure for these two experiments was the same as in the previous experiment. We report the results of both experiments together.

3.4.3.3 Results

Mechanical Turk - US participants

We first analyse the data in exactly the same way as the previous experiment - classifying all participants as either 'for' or 'against' on each topic, and then calculating their bias score based on how many of their choices support their position.

After removing those who failed to complete all questions for each topic, we are left with 118 participants for income inequality, 112 for the minimum wage, 107 for the death penalty, and 102 for gun control. Of those who completed the whole experiment (for all four topics), 51% were female, the average age was 35.9, and the average income $59,510.

Here we find that participants read more conflicting arguments than supporting arguments on average (see 3.4). The average bias score is negative for all four topics, though is only significantly different from zero for two topics - income inequality (mean= -0.42, t=-2.22, $p < 0.05$), and the death penalty (mean= -0.73, t= -3.17, $p < 0.01$), but not for minimum wage (mean= -0.32, t=-1.46, p=0.15) or gun control (mean = -0.31, t= -1.39, p=0.17).

	Income Inequality	Minimum Wage	Gun Control	Death Penalty	Average - all topics
Bias score	-0.42*	-0.32	-0.31	-0.73**	-0.50**

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.4: Bias scores by topic, experiment 2

However, as mentioned above, we actually have a more fine-grained measure of people's opinions in this study: responses on a 7-point scale for two questions, which we can average to get a single opinion measure for each subject. This allows us to ask about selective exposure in a slightly subtler way - rather than asking about choices that support/oppose one's opinions in a binary sense, we can ask to what extent people's

prior opinions correlate with their choices. That is, do people who have stronger 'pro' opinions (i.e. higher values) choose more 'pro' articles on a given topic? To answer this question, we look at the correlation between people's initial opinions, and the number of 'pro' articles selected on each topic. Since higher opinion measures correspond to being more 'pro' gun control/affirmative action, a positive correlation between these two values would indicate selective exposure. Table 3.5 below shows the results of this analysis. [6]

	Income Inequality	Minimum Wage	Gun Control	Death Penalty
Pearson's r	-0.039*	-0.14	-0.19	-0.28*

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.5: Correlation between initial views and choices, experiment 2

This broadly fits with our initial findings - there is a weak/moderate negative correlation between initial opinion and article choice for the death penalty which is statistically significant (r=0.28, t=-2.83, $p < 0.05$). We find no relationship with any of the other topics, including income inequality, despite the significant negative 'bias score' we found in the other analysis - demonstrating that these two different ways of looking at selective exposure do make some difference to the conclusions we draw.

[6]Note that we do not calculate 'average' values across all topics for Pearson's r in the way we do with bias scores. This is because, while it makes sense to ask what the average bias score is across topics (and we can calculate this fairly easily - by calculating an average bias score for each participant as the total difference between all supporting and conflicting arguments), it does not necessarily make sense to ask what the 'average' correlation is (we cannot calculate this simply by averaging correlation coefficients.)

Bilendi - UK participants

246 subjects were recruited for this experiment - after removing those who did not complete all questions, we are left with 240 for the topic of gun control, and 225 for the topic of minimum wage. Of those who completed the entire experiment (both topics), 47% were female, the mean age was 48, and the average income was 38,640.

This time we found no evidence of selective exposure for either topic when looking simply at bias scores (minimum wage: t=0.54, p=0.59; gun control: t=-1.21, p=0.23, see 3.6). Note that this is consistent with the findings of the previous study, since though we found a significant effect across all four topics, the two topics used in this study (min wage and gun control) were the two topics for which we did not find a significant effect in the previous study.

	Minimum Wage	Gun Control	Average
Bias score	0.044	-0.16	-0.041

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.6: Bias scores by topic, experiment 3 (Bilendi)

Again, we also look at the correlation between initial opinions and article choices (not that we expect to find anything given these results - but for the sake of completeness, and to double-check these results.)

	Minimum Wage	Gun Control
Pearson's r	0.045	-0.085

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.7: Correlation between initial views and choices, experiment 3

Unsurprisingly, we find no evidence of a correlation between initial opinions and argument choices 3.7.

3.4.4 Experiment 4: a replication of Taber and Lodge, 2006

3.4.4.1 Background

The fact that we found very mixed evidence for selective exposure in the last three experiments - weak evidence of selective exposure in the first, weak evidence against selective exposure in the second, and no evidence in either direction in the third - is not necessarily that surprising given the history of mixed findings in the selective exposure literature. What is perhaps surprising, though, is that these findings seem to conflict more directly with the results reported by Taber and Lodge (2006), who report the results of an experiment with a very similar design to the two I conducted, with quite different results - they found evidence of selective exposure where I found none. Though the findings of experiments 1-3 conflict with other studies in the selective exposure literature, we decide to focus on Taber and Lodge (2006) because (a) their paper was published more recently, with many other studies having been conducted before 2000; (b) their study design seems most similar in details.

To explore this discrepancy further, I began by attempting to replicate Taber and Lodge's experiment as closely as possible, the main difference being that my replication took place online rather than with subjects in a laboratory. Rather than being a problem for the replication, however, this subtle difference allowed me to investigate the robustness of their results - and to explore whether the different context (lab versus online) might explain the discrepancy of results.

3.4.4.2 Design and procedure

The procedure for the experiment followed the original design of Taber and Lodge (2006) as closely as possible. Participants began by answering a number of questions evaluating their attitude towards either affirmative action or gun control (the order of these two issues was counterbalanced by random assignment.) Items used to measure attitude position and strength were those used in the original study - four items to measure attitude position (100-point sliding response scales) and six to measure attitude strength (9 point agree/disagree Likert items.) Both these scales were tested for reliability by the original authors - see Taber and Lodge (2006) for details. Participants were then given the opportunity to read eight of sixteen total arguments, half on either side of the issue.

Arguments were labelled with a known source (either political parties - Republican or Democrat - or one of two interest groups for the specific issues - the National Rifle Association and Citizens Against Handguns for the topic of gun control, for example), and participants were explicitly told each group's position on the issue. After choosing and reading arguments, participants then completed the attitudinal questions a second time, followed by a set of demographic questions.

3.4.4.3 Results

We surveyed 188 US participants, again using Amazon's Mechanical Turk. 38% of subjects were female, with an average age of 35, and an average income of $69,150. We also collected data on participants' political views in this experiment - 18% of participants identified as 'conservative', 63% as 'liberal', 14% as 'moderate', and the remainder as 'other.'

Here, we find significant, positive, bias scores for both topics - an average bias score of 0.42 for affirmative action (t=2.17, $p < 0.05$), and an average of 0.47 for gun control (t=2.51, $p < 0.05$), and a significant, positive correlation between initial opinions and article choices - indicating that people do in general choose more supporting than conflicting arguments, consistent with Taber and Lodge's original findings (3.8).[7]

	Affirmative Action	Gun Control
Bias score	0.42*	0.47*
Pearson's r	0.22*	0.38***

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.8: Bias scores and Pearson's r, experiment 4

This suggests that the discrepancy in results between this and my experiments 1-3 cannot simply be explained by the sample used, or the fact my studies were conducted online:

[7]We cannot calculate an 'average' bias score across the two topics in this case, because in this experiment the topics were randomised so that each participant only chose articles for one of the two topics. This was in keeping with Taber and Lodge's original design, which was slightly more complex in order to allow them to test additional hypotheses. In particular, they wanted to investigate how people's prior attitudes influenced their *evaluations* of information, but needed to keep this separate from article choices so that all participants were evaluating the same information. Therefore, half the participants chose arguments for affirmative action and evaluated arguments (that they did not choose) for gun control, and the other half did the opposite.

it must be due to some aspect of the experimental design. Some possible hypotheses include:

- Different measures of initial attitude: perhaps the measures used by Taber and Lodge (2006) to measure attitudes were more reliable and so better able to detect the extent to which people were choosing supporting/conflicting arguments;
- Question framing: I gave people a choice between arguments for/against an issue, whereas in Taber and Lodge's design people chose between arguments from political interest groups, which could affect the choices people make;
- Topic: as I found in my first experiment, the extent of selective exposure seems to vary greatly by topic. Since the two studies used different topics, this might explain the discrepancy in results.

This third hypothesis - that the difference is down to the topics - is made less likely, however, by the fact that there was at least some overlap in topics across the experiments. Both used gun control, and so we would still need to explain this difference. We therefore next design an experiment to test the second hypothesis: that the difference is due to the question framing. If we find that this does not explain the difference - if changing the question framing still leads to the kind of selective exposure results we saw in the previous replication - then it may be worth investigating whether more reliable measures can explain the difference.

3.4.5 Experiments 5 and 6

Perhaps the most notable difference between experiments 2 and 3, and the successfully replicated Taber and Lodge study (experiment 4) lies in how the choice of arguments were presented to people. That is, the information people have about the available arguments, on the basis of which they can decide what they are most interested in reading, varies. In the first study I ran, people chose between arguments on the basis of reading sentence-summaries of those arguments, which explicitly stated the 'side' of the issue the argument lay on. However, I decided to change this in my second experiment, realising that these sentence summaries made it very difficult to control for 'interestingness' - some of the summaries may simply have sounded more interesting

to people than others, and so their choices may have been influenced more by this than an awareness of whether the arguments would support or conflict with their prior beliefs. In my second study, therefore, people were presented with a very abstract choice between reading "an argument in favour of [issue]" and "an argument against [issue]", so that it was easier to isolate the effect of the direction of the argument on people's choices. In making this small change, we found weaker evidence of selective exposure, and some evidence for the opposite effect - suggesting that selective exposure is slightly more likely to occur when people make choices between sentences summarising different articles (but which nonetheless indicate the direction of the argument), than when given a more explicit choice between arguments on two sides of an issue.

In Taber and Lodge (2006), as detailed above, participants instead choose between arguments from different groups with known positions on the issue. Rather than choosing between arguments for/against gun control, participants chose between arguments from the Republican party and the Democratic party, with the explicit awareness that the Republican party opposes and the Democratic party supports gun control. As with my sentence summaries, this does make isolating the effect of the argument direction a little trickier, since participants might have been choosing arguments based on whichever party they had more positive feelings towards, and paying less attention to the explicit direction of the argument.

I therefore hypothesized that it might be this difference - in how the choice between arguments was presented to people - that might explain the different results. Perhaps when given an explicit choice between arguments for and against gun control, people are less likely to exhibit selective exposure since they are made more aware of the fact that there are two sides to the issue, making them more aware of the need to make a balanced assessment. By contrast, when choosing between arguments from different interest groups or political parties, people's choices may be more influenced by those groups they identify with most strongly or feel most positively towards (which are also likely to be those groups that generally agree with them on issues.)

This hypothesis also seems somewhat similar to the distinction Sears et al. (1967) draw between active and 'de facto' selective exposure: recall their claim was that people do not have an explicit preference for supportive information, but do in reality end up exposed to more supporting than conflicting information in more indirect ways. One

such 'indirect' way that people might end up exposed to more supporting information is through social factors and group affiliation: people may seek out information from people and groups they feel positively towards, without any explicit intention of confirming their views. Of course, those we like and feel are similar to us are likely to also agree with us on important issues, so seeking out their viewpoints may well indirectly result in us seeking out supporting viewpoints.

3.4.6 Experiment 5: information source manipulation

To test this hypothesis, we ran the exact same experiment - an online replication of Taber and Lodge (2006) as before - but with a subtle manipulation of how information choices were presented to participants.

3.4.6.1 Design and procedure

In the 'control' condition (pro/con in table 3.9 below), participants chose arguments as in experiments 2 and 3 - selecting between the two options "I would like to read an argument from someone who believes that gun control should be increased", and "I would like to read an argument from someone who believes that gun control should be reduced." In the 'treatment' condition (political groups), participants chose arguments as in Taber and Lodge's original experiment - selecting between four options of the form "I would like to read an argument from X, a group who oppose/favour gun control" (with four different groups used - two political parties and two interest groups - two on either side of the issue.)

3.4.6.2 Results

After removing any participants who did not complete the survey or failed attention checks, we are left with 170 US subjects (97 in the 'pro/con' condition, and 73 in the 'political groups' condition.) In the 'political groups' condition, 42% of participants were female, with an average age of 35.75 and an average income of $48,940. 23% of participants identified as 'conservative', 55% as 'liberal', and 21% as 'moderate.' [8]

[8] Due to an unfortunate error in data collection, we did not manage to collect demographic data for those in the 'pro/con' condition.

We find no significant difference between the bias scores for the two conditions, using a one-way ANOVA (gun control: F(1,168)=0, p=0.998; affirmative action: F(1,168)=1.53, p=0.218).

Looking directly at the bias scores for the different groups, we find none of them differ significantly from zero:

	Affirmative Action	Gun Control	Average
pro/con	-0.16	0.25	0.04
political groups	0.33	0.25	0.29
both conditions	0.047	0.25	0.15

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.9: Bias scores by condition, experiment 5

	Affirmative Action	Gun Control
pro/con	-0.14	0.1
political groups	0.15	0.32***
both conditions	0.03	0.16*

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.10: Pearsons's r by condition, experiment 5

This suggests already that introducing the information source manipulation - adding a condition where some participants choose between arguments for/against an issue, without any source information - reduces/eliminates the selective exposure effect compared to Taber and Lodge's original study.

When we look at the different bias scores between the two conditions, we find that the average bias score in the control condition (pro/con), is lower than the average bias score for the treatment condition (political groups) for affirmative action, but that the difference is not statistically significant. For gun control, there is no difference in bias scores for the two conditions. When we average across the two topics, again the bias score is lower for the pro/con condition than for political groups, but the difference is not statistically significant.

However, when we look at the relationship between initial attitudes and choices (recalling that this allows us to take account of more variation in initial attitudes, rather than categorising all participants as either 'for' or 'against'), we do find a weak positive correlation between initial attitude and choices for gun control, driven largely by a moderate positive correlation in the political groups condition (see 3.10). This, along with the fact that bias scores were higher in the political groups condition for affirmative action (but equal for gun control) provides some relatively weak support for our hypothesis that people are more likely to select arguments that support their prior beliefs when those arguments are presented from specific groups. It's also worth noting that the correlation coefficient for gun control in the political groups condition is very similar to that we observed in experiment 4, the first recplication of Taber and Lodge (0.32 vs 0.38), but for affirmative action is slightly lower. As noted earlier, this also reinforces how analysing the data in slightly different ways can produce subtly different results - perhaps contributing to the literature's mixed findings, as I will discuss later.

A final point worth noting here is that if we look at the average bias score just in the condition where choices are presented as in Taber and Lodge (2006), we should presumably expect to find roughly the same results as in that original study (and in our replication.) However, the results here are not quite as expected: we find an average bias score for gun control of 0.25, and for affirmative action of 0.33, neither which differs significantly from zero. Neither are the correlations between initial opinion and article choices anywhere near as strong. Partly this may be because we are underpowered to detect the effect - since our sample was roughly the same size as in the previous experiment, but the sample was split into two conditions, we only have half the number of people in the treatment condition as were in the original replication.

3.4.6.3 Participant's reported reasons for choices

In experiment 5, we also included an open question asking people why they chose to read the arguments they did, to see if this could help shed any light on people's motivations . Of course, these responses have to be taken with the awareness that they may be vulnerable to social desirability bias - people answering in a way that they think sounds good, rather than with their actual motivations. The responses are interesting nonetheless.

For both gun control and affirmative action, over 50% of participants made some reference to being fair, unbiased, or balanced in their choice of arguments to read - 58% of participants for gun control, and 66% of participants for affirmative action. It's also interesting to compare this to the proportion of people who were *actually* entirely balanced in their selections - only 50% (gun control) and 60% (affirmative action) actually chose an equal number of arguments on either side, so there were at least a few participants who chose more arguments on one side or the other but explained their choices in terms of being unbiased. The fact that that so many people gave such an explanation suggests that, even if people aren't always very good at *being* unbiased, they are certainly well aware that being (perceived as) unbiased is desirable and/or important. A further 26% (gun control) and 17% (affirmative action) said that they were curious about what the other side had to say - though several of these explanations also include some comment indicating that they don't expect to be convinced or are even curious because they expect the arguments to be poor. Only a small proportion - 13% for gun control and 6% for affirmative action - said that they chose to read arguments that would support or reinforce their existing views.

Of course, what is difficult here is judging whether people are accurately reporting their motivations, or simply saying what they think sounds good (in a way, it is surprising that so many people were so blunt as to say they read things to reinforce what they already believed.)

One reading of these results - the lack of clear evidence for selective exposure, and the fact many people talk about the importance of being unbiased - might be to say that people simply are much less biased than we often give them credit for. A more sceptical reading might be that people are choosing mostly balanced arguments because they think it is important to appear unbiased - but are still highly biased in their interpretation/evaluation of those arguments. Pulling apart these two explanations is far from easy, without a lot more information about what people already know and how they are actually engaging with the information - as I will discuss in more detail later.

3.4.7 Experiment 6: information source manipulation 2

We next run the same experiment as previously, but with a larger sample size, to check whether the relatively small sample size explains the lack of a significant effect in the

previous study.

3.4.7.1 Design and procedure

With a sample this time of 379 (originally 400, minus any participants excluded for not completing the study properly or failing attention checks)[9], we run the exact same procedure as in the previous study - replicating Taber and Lodge (2006), with participants randomised into control and treatment conditions in which the choice between arguments is presented differently.

3.4.7.2 Results

The below tables show the breakdown of demographic characteristics, and of political views in the two conditions:

	Total size	Gender (% F)	Age (mean)	Income (mean $)
pro/con	199	50%	36.8	50180
political groups	198	49%	36.4	54810

Table 3.11: Demographic characteristics, experiment 6

	conservative	liberal	moderate
procon	31%	54%	15%
political groups	28%	50%	23%

Table 3.12: Political views, experiment 6

As before, a one-way ANOVA finds no significant difference between the bias scores for the two conditions, though we do get slightly closer to significance with the larger sample size (gun control: $F(1,394)=1.845$, $p=0.175$; affirmative action: $F(1,394) = 0.97$).

Looking at the bias scores across conditions themselves:

[9] 'Attention checks' are questions added to the survey with obvious answers to ensure the participant is paying attention.

	Affirmative Action	Gun Control	Average
pro/con	-0.12	-0.08	-0.1
political groups	0.1	0.15	0.13
both conditions	-0.01	0.035	0.013

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.13: Bias scores by condition, experiment 6

	Affirmative Action	Gun Control
pro/con	0.02	-0.16*
political groups	0.16*	0.12
both conditions	0.085	-0.017

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE 3.14: Pearsons's r by condition, experiment 6

This time, the bias scores are yet lower and the difference between conditions even smaller than previously (3.13). Across both conditions the average bias scores for gun control and affirmative action respectively are 0.035 and -0.01 - neither significantly different from zero, unsurprisingly. The pattern we see between the two conditions is similar to in the previous study, except that the bias scores are lower and the differences smaller: for gun control we find bias scores of -0.08 (control) and 0.15 (treatment), and for affirmative action -0.12 (control) and 0.1 (treatment). In neither case is the difference between the two groups significant. And when we again look at the bias score in just the treatment condition - where the procedure follows our original replication of Taber and Lodge (2006), we find that it does not differ significantly from zero for either topic - affirmative action: 0.1, and affirmative action: 0.15.

As for selective exposure measured as the correlation between initial attitudes and article selection (3.14), we find a roughly similar picture: the general pattern is that we see slightly stronger and more positive correlations for the political groups condition, but the correlations are weak and only sometimes statistically significant.

3.5 Summary and discussion

3.5.1 Summary

My first three experiments produced very mixed results: in one case, weak evidence for selective exposure, in another, evidence of the opposite tendency, and in the third, no evidence in either direction. These mixed results are not particularly surprising given the mixed status of the selective exposure more generally, but what is perhaps surprising is that we obtain such mixed results from relatively subtle tweaks to the same paradigm.

Furthermore, these results seem to conflict with Taber and Lodge (2006), who found a significant selective exposure effect using a very similar experimental design. We successfully replicate this result in an online experiment following Taber and Lodge's design as closely as possible - prompting the question of what might account for the difference in results between this and the previous studies.

The most notable difference between these experiments seems to be how the options of what to read are presented to participants: Taber and Lodge (2006) have participants choose between arguments from different political interest groups, whereas we simply present people with a choice between an argument 'for' or 'against' a given issue. Our initial findings provide weak support for the hypothesis that selective exposure is less likely to occur when choosing between arguments in this latter, more abstract, way - of the first three studies, the two in which arguments were presented more abstractly (experiments 2 and 3) found less selective exposure than the first experiment (where people chose between arguments from sentence summaries.) Of course, these differences may be explained by some other factor, but the argument presentation does stand out as the most obvious difference between these studies.

I therefore tried investigating the impact of how choices are presented, by re-running the Taber and Lodge replication with a manipulation that varied how arguments were presented to people. The primary hypothesis was that we would find more selective exposure when people were choosing between arguments from known sources (political interest groups, as in Taber and Lodge (2006), than when choosing between arguments abstractly labelled as either 'for' or 'against' a given issue. We ran two experiments to test this hypothesis. In the first (experiment 5), though bias scores are slightly higher in the 'political groups' condition, the difference is not significant. We also find that

the overall bias score across both conditions is lower than in Taber and Lodge's original experiment, and not significantly different from zero. However, this lack of significant findings may be because our sample is smaller and therefore underpowered (we had the same number of subjects as Taber and Lodge, but split into two conditions.) We therefore run the same experiment on a larger sample (experiment 6) - but this time find that effect sizes are even smaller, and again not significant. Though bias scores are still higher in the political groups condition, they are not significantly so - suggesting that if there is any meaningful effect here, it is small.

When we analyse the data using a slightly different method - looking at the correlation between initial attitudes and arguments chosen, and therefore capturing more of the variation in people's opinions - we find broadly similar results. However, especially in the later experiments, this measure of selective exposure sometimes produces statistically significant results where the simple 'bias score' measure does not. In particular, this indicates slightly more support for the hypothesis that reading arguments from specific sources makes people more likely to engage in selective exposure. We should be wary of drawing conclusions from this, however, since the correlations are weak (and not significant in all cases), and especially given the lack of significant findings with the bias score measure. What this does reveal, however, is that analysis choices can have subtle but significant effects on the conclusions drawn. If we had solely used the correlational analysis then we might have been more likely to conclude support for our hypothesis, than had we used just the bias score measure (or both measures.) I discuss the implications of this in more detail in the next sections.

3.5.2 Discussion: just another failed replication?

What should we make of all this? First, our results taken together suggest that selective exposure in the domain of political attitudes is not a particularly robust phenomenon, at least not measured in the way it has been here. Though mixed evidence for selective exposure is not a novel finding (Hart et al., 2009), we are not aware of previous research which finds such mixed results for selective exposure within the same basic paradigm.[10] Second, though the way information sources are presented may affect the extent to which

[10]Prior research has discussed how selective exposure effects seem to vary greatly depending on different topics, contexts, etc. - not how they vary across different studies of the same topics and context.

selective exposure occurs, we do not find convincing evidence for this in my experiments - if an effect like this exists, then it is relatively small.

What might explain such confusing and mixed findings? It might be that the effect Taber and Lodge (2006) found does in fact exist, but the later studies I conducted failed to replicate the right conditions in some subtle way. Alternatively, perhaps no such effect exists, and Taber and Lodge's original finding was some kind of statistical fluke. Both of these explanations seem implausible, however, given that (a) I did successfully replicate Taber and Lodge's result once; and (b) the series of experiments I conducted were practically identical in design and procedure. What is perhaps more likely is that some tendency to select more supporting arguments *does* exist, but there are also a large number of other motivations and factors influencing how people select information, and a 'selective exposure' effect is not strong enough to override them. This means that any 'selective exposure' effect is hard to detect and may easily be swamped by variations in the sample or experimental design.

It's also worth considering these mixed results in the context of the recent 'replication crisis' in psychology. In the last few years, the robustness of various findings in psychology has been challenged, as many have failed to replicate. Most notably, a large team of researchers as part of the 'Open Science Collaboration' recently attempted to replicate 100 studies published in three different well-established psychology journals. The final paper (Open Science Collaboration, 2015) reports that replication effect sizes were on average half the magnitude of originals, only 36% of replication studies found a significant effect (compared to 97% of the original studies), and in all, only 38% of the studies were 'subjectively rated' as successful replications.

The authors acknowledge that there is no single, agreed upon standard for evaluating the success of a replication - and use a combination of different measures, including the statistical significance of the replication effect, the difference between the original and replication effects, directly comparing effect sizes, and combining original and replication results in a meta-analysis. These different methods have their respective advantages and disadvantages - which has allowed others to challenge whether the failures of replication are really as severe as claimed. Gilbert et al. (2016) argue in a commentary paper that the replication project provides no grounds for drawing pessimistic conclusions about

reproducibility in social science. They suggest that due to sampling error and noise, we should not expect all effects to replicate even if they are true.

Statistician Andrew Gelman responds to Gilbert et al., suggesting that they are too quick to take the original result as solid evidence - and suggests that the burden of proof should be the other way around (Gelman, 2016a,b). Gelman suggests using a 'time-reversal heuristic': if the replication result had come first, if first someone had run a study with a large sample that found no effect, and then afterwards someone else came along with a smaller sample and did find an effect, would we then be confident that such an effect exists? It seems like we would not - and so we should not continue believing in the effect simply because the positive finding came first.[11]

The fact that psychologists and statisticians cannot even agree on what it means for a finding to be reproduced successfully should itself concern us about the state of evidence in psychology. If we can't agree on what it means for a finding to be successfully reproduced, can we really say we know what constitutes solid evidence for a psychological claim in the first place?

Beyond this specific discussion of reproducibility in psychological science, there are various deeper reasons to be concerned about the validity of findings in social science. Going as far back as the 1960s, Meehl (1967) argued that, due to noise and variation in statistical studies, it's possible to find statistically significant effects where no true effects exist if you look hard enough. Gelman and Loken (2013) explain how this can occur even without any deliberate attempt to 'fish' for a significant effect on the part of the researchers. This builds on the notion of 'researcher degrees of freedom' (Simmons et al., 2011): there are various different decisions that researchers have to make in collecting and analysing data - when to stop collecting more data, which observations to include, what kind of analysis to do, and so on - and often these decisions are not all made in advance. This means that even if researchers do not actually conduct multiple different analyses, there are multiple different analyses they *could* have done - meaning the chance of any one of these analyses yielding a statistically significant results is substantially larger than 0.05. Simmons et al. (2011) show that subtle manipulations to these researcher degrees of freedom can result in finding statistically significant results

[11] It's also interesting to note here that the idea that what result we hear first influences our interpretation of subsequent results is, of course, very close to a claim of confirmation bias!

supporting a hypothesis that is blatantly false - that listening to children's music makes people younger.

John Ioannadis argues independently, in a provocatively titled paper, that in all likelihood, "most published research findings are false" (Ioannidis, 2005). Ioannadis shows, using Bayesian reasoning, that due to publishing and analytic practices, more than half of published research results are likely false. Gelman (2016a) further argues (also using the provocative term "piss-poor social science") that it's basically impossible for most of social science's findings to be true, since so many claim large effects of small manipulations. If all these findings represented genuine effects, Gelman argues, we would live in a very strange world (not to mention the fact that many such results appear to directly contradict one another.)

Gelman argues that the solution to all of these issues is not as straightforward as simply improving statistical practices (Gelman, 2016a,b). More replications and better statistics will help of course, but if many of the purported findings do not exist, then better practices will simply result in lots of negative findings and no real knowledge. What we need really, he argues, is better measurement tools and research design: we need to learn to ask better questions, and to reward scientists for rigour and careful testing of clever hypotheses, not for findings that make good headlines. Better methodological practices will help indirectly, in that they will shift the cost-benefit ratio - so that it is harder, and so costlier, to produce 'sexy' results by using sloppy practices - but are not themselves the solution.

With this backdrop of issues - questions about the reproducibility of psychological studies and the robustness of scientific evidence even more broadly - the mixed results for selective exposure are perhaps not surprising. Perhaps the most straightforward explanation for what is going on is that the 'selective exposure' effect is relatively small, just one factor influencing how people approach information among many. And as Gelman (2016a) suggests, the solution to getting clearer on what is going on here may not simply be larger samples and better statistics. Instead, what may be needed is to step back a bit and ask: why are we interested in selective exposure, really - and how might we measure what we're interested in better?

It seems to me that it's not selective exposure *per se* that's important, but more the broader ways that people may be biased towards confirming their existing beliefs.

Selective exposure may be one way in which this occurs, or one contributing factor - but if it is confirmation bias more broadly that's actually important to understand, we need to consider bias in all stages of reasoning. Instead, then, of continuing to try to understand when and whether this specific kind of 'selective exposure' occurs, we might learn more by coming up with better ways of measuring how people form and update their beliefs more broadly.

3.6 General discussion

Given the very mixed results discussed here, it seems reasonable to conclude that there is no clear, strong selective exposure effect, even in the narrow domain of political beliefs. The simple claim that, "people prefer to read things that support their existing views to things that challenge them", while intuitively plausible, does not hold up to scrutiny, at least not using the standard methods of measuring selective exposure. This is not to say that the desire to validate or confirm one's current opinion is not an important factor influencing what information people tend to seek out - but rather that it may not be as strong a factor as has sometimes been supposed, since subtle manipulations seem able to eliminate or reverse the effect.

To conclude this section, I discuss some of the main issues related to selective exposure research, and their implications.

3.6.1 Design issues in selective exposure research

One possible reason for the lack of clear evidence for selective exposure, which seems relatively under-discussed, is that the experimental design of most selective exposure studies may produce strong demand effects. It seems likely that many participants will guess that their reading choices are being judged - and even if not, simply being in an experimental context where they are being 'watched' may make people more self-conscious about appearing biased than they would otherwise be.

Relatedly, most selective exposure tasks present a much more explicit, abstract choice between different sources of information than people would encounter in everyday life. We do not generally choose what to read from a list of options in such an explicit

way, and so the way people make choices in this setup may differ from how they make decisions in everyday life.

These two issues raise concerns about how far we can generalise from people's behaviour in selective exposure studies to their real-life behaviour. Some studies attempting to bridge this gap can help - experimental studies that attempt to present information more like a natural online browsing environment, for example. Some studies also attempt to reduce demand effects by framing the experiment in different terms - telling participants we are primarily interested in how they form arguments or discuss with a partner, for example. However, we should still be conscious that the results of most selective exposure studies show us primarily how people behave in such abstract scenarios - and the features of the much richer environments in which we actually obtain information need to be taken into account when attempting to generalise.

It's also worth briefly mentioning a couple of other experimental design issues which might partially account for mixed selective exposure findings. First, results are heavily dependent on the assumption that the study has accurately captured subject's attitudes - and therefore whether or not they are choosing to read 'opinion-confirming' information. However, attitude measurement is complex, and research has found that answers to attitude questions can be very context-dependent: varying as a function of the phrasing of the question asked and earlier items in a survey (Tourangeau et al., 1989). Second, something that is not often discussed in the selective exposure literature is how the framing of the task itself might affect participant's responses: what exactly do people think the purpose of reading different articles is? Presumably people's choices will be very different if their goal is genuinely to learn about a topic than if they think they might be expected to defend their view, and different again if all they are really doing is trying to finish the experiment as quickly as possible. Though different goals are sometimes discussed as moderating factors (e.g. Clarke and James, 1967), different interpretations of the 'goal' of the task might be worth exploring further.

3.6.2 Different ways of measuring 'selective exposure'

As I touched on briefly earlier, different design and analysis choices can also notably affect results - and whether or not a significant 'selective exposure' effect is found (Feldman et al., 2013). There is a fair amount of variation in how different studies have measured

'selective exposure', and a number of possible choices that researchers have to make: including whether people make real or hypothetical decisions; whether or not to include moderates in the final analysis; whether to ask opinion questions in a way that 'forces' people to express an opinion one way or another, and whether to look specifically at those who have more 'extreme' opinions. Feldman et al. (2013) show that a number of seemingly subtle methodological choices can affect the presence and strength of selective exposure effects.

In my analysis, I distinguished two main methods of analysis that have been used in selective exposure studies, which can yield subtly different results. In the first case, people's opinions are classified in a binary way as either 'for' or 'against' an issue, as are the different arguments - allowing us to look at the average difference between the number of 'supporting' and 'conflicting' arguments chosen. This is the method used in many selective exposure studies, but has the limitation that it fails to capture variation in the extremity of people's opinions: both someone who indicates weak support for gun control and someone who indicates very strong support are equally classified as 'supporting gun control'. Where we have people's opinions on a continuous scale, it makes sense to use this additional information - and instead to measure selective exposure as the extent to which people's prior views are correlated with their article choices. Although when we used these two different analysis strategies we did get broadly similar results, there were some subtle differences - cases where a correlation was statistically significant but the bias score was not, or vice versa.

A third way we could have chosen to analyse the data differently is by using what Feldman et al. (2013) call an "ideology-based selectivity measure" as opposed to an "attitude-based selectivity measure." Feldman et al. (2013) point out that there is a difference between claiming that people choose what to read on the basis of their specific attitudes on that issue, and that they choose what to read based on their broad political attitudes. Comparing analyses of selective exposure measured in these two different ways, they find more evidence of selective exposure when using an attitude-based selectivity measure than using an ideology-based one (however, they recognise this may be because more people are classified as 'moderates' on the ideology-based measure.) In the analyses discussed so far, selective exposure has always been defined on the basis of specific issue attitudes.

In Appendix C, I show how using an ideology-based selectivity measure to analyse the data in the last two studies produces again slightly different results. In contrast with Feldman et al. (2013), we actually find slightly more evidence of selective exposure when using an ideology-based measure - but only for the conditions where arguments are presented as coming from specific groups. On reflection, this is not particularly surprising - since two of the groups named were political parties, essentially what this tells us is that Republicans/Democrats are more likely to choose arguments from the Republican/Democratic party (and that this kind of selectivity is more common than people simply choosing arguments they expect to agree with their views on specific issues.)

The fact that there is variation in how selective exposure is measured across studies, and that different methodological choices can lead to different conclusions about the presence and strength of selective exposure effects, has two implications. First, it might well at least partially explain the mixed findings in the selective exposure literature: lack of consistency in methods and analysis unsurprisingly leads to inconsistent results. Second, it has implications for some of the problems related to the replication crisis discussed at the end of the last section. Gelman and Loken (2013) refer to these different choices researchers have to make as 'researcher degrees of freedom' and argue that they can lead to a multiple comparisons problem even when researchers only perform a single analysis of their data - since there are multiple potential comparisons that could have been done were some of these choices made differently. "A dataset can be analyzed in so many different ways that very little information is provided by the statement that a study came up with a $p < 0.05$ result." (Gelman and Loken, 2013, p.1) This suggests that we should perhaps be a little more hesitant to accept 'statistically significant' findings in the selective exposure literature at face value.

3.6.3 Selective exposure and bias

A larger problem for the selective exposure literature, beyond such mixed evidence, is that it is not entirely clear how one should interpret findings of selective exposure (or a lack thereof) in a wider sense. It is often implicitly assumed that selective exposure is a bias - Hart et al. (2009) refer to selective exposure as a 'congeniality bias', for example - and that reading a balance of arguments is the normative response in these tasks.

However, there isn't actually any clear normative standard in most selective exposure experiments - and in most studies, we do not have enough information (about subject's motivations, what they already know, and what they do with the information) to say whether their responses are really irrational.

A person who chooses to read more opinion-confirming information certainly looks more biased than someone who reads a balance, in the simplest sense, but it is not always this simple. Perhaps the first person reads opinion-confirming arguments genuinely intending to scrutinise them and ensure the basis for her position is solid. Perhaps the person who reads a balance of arguments, by contrast, is much more critical and dismissive of those that conflict with her viewpoint than those that support it. Perhaps a third person who actively seeks out counter-arguments to her view does so with the intention not of changing her mind, but of making sure she can develop convincing rebuttals to them when challenged.

Moreover, equating selective exposure with bias assumes that the only, or primary, goal one has is to obtain as accurate information as possible. But people might legitimately have other goals - such as saving time and energy, and staying happy - and sometimes, seeking out conflicting viewpoints might conflict with these goals. Whether it's appropriate to prioritise one's happiness or saving time over having accurate beliefs is another, very complex, question in itself - but putting this aside for now, selective exposure might often be said to be 'rational' if we assume certain goals. Therefore, though selective exposure has generally been considered a bias, whether it is genuinely irrational depends a lot on the context and the different goals at play.

Given this, findings within the selective exposure literature ideally need to be considered within a broader context: with a better understanding of subjects' motivations and goals, and how their information choices relate to other reasoning processes, such as how arguments are evaluated. At very least, we need to be very careful to maintain that a tendency towards selective exposure is, in itself, an interesting psychological phenomenon, but not necessarily a bias - acknowledging that the normative issues here are complex and require a great deal more information.

3.6.4 Selective exposure in the 'real world'

Even if selective exposure doesn't hold up as a strong effect in the simplest sense in abstract experimental settings, there is some evidence that a more passive form of selective exposure may be a real-world problem. Even if people do not have a strong active preference for opinion-confirming information, it may be that people often do end up exposed to more opinion-confirming information, or even seeking out more such information, more indirectly.

A number of different factors could explain this apparent discrepancy. It might be that it is simply easier for people to access opinion-consistent information - because the sources are more familiar, perhaps. If people mostly encounter information through social networks online and offline, and we are generally socially connected to people similar to us, then we will end up exposed to more information that we agree with and less we disagree with. If we prefer to read articles from websites we are familiar with and feel positively disposed towards, we will probably largely read articles from websites we agree with - even if that was never our explicit intention. We might also find sources that agree with us easier to understand, and find it easier to evaluate the quality of their claims - and so find it more informative, at least in a narrow sense. There are a variety of reasons why we might find it difficult to encounter, or even prefer to avoid, information that challenges what we believe, that don't require us to avoid or dislike conflicting information in itself.

More research has begun to look at information choices in naturalistic settings, especially using online behaviour, which can help to shed light on whether this more 'passive' form of selective exposure occurs, and if so, what factors are driving it. As suggested above, the research on selective exposure implies that the desire to validate one's current position is certainly one factor influencing information choices, but it is far from being an overwhelming one. Rather than focusing too much on the question, "are people really motivated to defend their beliefs by seeking out supportive information?" (to which the answer is, "yes, to some extent, but the relative strength and power of this motive varies"), we might be better to focus more on understanding the different basic factors influencing how people choose what information they pay attention to more generally, in both experimental and more naturalistic contexts. In Appendix A I outline some of these factors and how they may/may not provoke selective exposure under different

circumstances. In addition, rather than studying what information people choose to pay attention to in such a broad sense, we might learn more from picking out specific situations in which people seem particularly likely to make poor decisions (depending on what their goals are), and to try to understand why.

3.6.5 Final thoughts and implications

Why, ultimately, has the research community been so interested in selective exposure? I think it is because it is indicative of a broader tendency that seems problematic: the tendency to reason in ways that lead us to confirm what we already believe, and that prevent us from changing our minds even when doing so might be important. As I said at the beginning, selective exposure is just one form of a broader tendency known as confirmation bias. In more intuitive terms, selective exposure is interesting to us because it seems to indicate a lack of open-mindedness, and a lack of open-mindedness seems like something to be concerned about. In turn, the reason we are concerned about confirmation bias and/or closed-mindedness is not purely intellectual - they seem to have certain, concerning, practical implications - preventing us from figuring out the truth on certain important scientific and political issues, fueling conflict, and hindering progress. I will discuss these issues in more detail - particularly concerns around a lack of 'open-mindedness' - in chapter five.

The selective exposure literature, however, as it has got caught up in this question of whether and to what extent the effect really holds, has begun to lose sight of this wider question. Selective exposure in itself is not particularly interesting - it's interesting insofar as it contributes to confirmation bias and/or a lack of open-mindedness. But as discussed above, as it is currently studied, selective exposure isn't actually a very good measure of bias or closed-mindedness: it's possible for an entirely rational person to exhibit selective exposure under certain assumptions, and someone might seek out counter-evidence but still approach it with an entirely closed-minded attitude. Selective exposure also hasn't been studied in a way that easily allows us to draw conclusions about its practical implications - that is, it's hard to say from these very abstract lab studies whether a tendency towards selective exposure, if it exists, causes genuine problems for individuals or society.

I think the reason selective exposure effects have been so mixed and elusive is that there are many different factors influencing someone's choice of information: their motivations, what they already know and believe, their expectations about different sources of information, which are difficult if not impossible to capture experimentally. The reason that mixed effects for selective exposure have been viewed as so surprising is that it's been assumed that selective exposure broadly corresponds to 'open-mindedness' (and it's broadly assumed that people are not open-minded.) This stems from a failure to recognise that without understanding these motivations, beliefs, and expectations, it's very difficult to draw conclusions about open-mindedness or rationality from findings of selective exposure (or a lack thereof.)

In studying selective exposure, I suggest, we need to take a step back: and think about these findings with a clearer understanding of other reasoning processes, and more background of what it means to be biased, what it means to be open-minded. Rather than asking "is there evidence for selective exposure?", we instead need to ask questions like, "what does it really mean to be open-minded, why is this important, and how do we measure it?", and "to what extent is confirmation a problem, and how much of this comes down to how people select information versus bias in other stages of the reasoning process?" I'll now turn to these broader questions: first looking in more detail at what it really means to be biased or irrational, and the implications for confirmation bias, before turning to a closer look at the related concept of open-mindedness.

Chapter 4

Bias, rationality and improving human reasoning

4.1 Introduction

I've reviewed a variety of research claiming that a 'confirmation bias' arises in human reasoning - and discussed some reasons that the evidence for confirmation bias is less strong than it first seems. A theme we keep returning to is the difficulty of establishing standards by which to evaluate how people *should* reason - and how most of the research on confirmation bias fails to adequately address this. Ambiguous use of terms like 'biased' or 'rational' also seem to create unnecessary disagreement and confusion. This chapter therefore focuses in on these normative issues, clarifying some of the terminological confusion to see what substantive disagreements we are left with.

Many others have acknowledged the difficulty of establishing when and how it's rational to reason in ways that 'confirm' what one already believes. Nickerson says at the end of his review on confirmation bias that "the question of the conditions under which one should retain, reject or modify an existing belief is a controversial one", and that "it is natural to be biased in favour of one's established beliefs... whether it is rational is a complicated issue that can too easily be treated simplistically." (Nickerson, 1998, p.209) In many classic studies of confirmation bias, the normative standards against which performance is compared are vague, disputed or nonexistent, and several arguments have been made that tendencies classically interpreted as evidence of confirmation bias

may actually show no such systematic bias when these normative standards are made clearer (Austerweil and Griffiths, 2008, Jern et al., 2014, Klayman, 1995, Klayman and Ha, 1987, Oaksford and Chater, 1994, Perfors and Navarro, 2009). Klayman concludes that "it is quite clear that quite a few of the putative sources of confirmation bias do not directly imply any consistent bias towards the focal hypothesis." (Klayman, 1995, p.398) Even if a tendency to favour what one currently believes does exist, others still have argued this might be rational if we assume people have different goals than simply epistemic truth-seeking (Friedrich, 1993, Mercier and Sperber, 2011, Tooby and Cosmides, 1992).

Though it's widely acknowledged that these complex normative issues exist, it's still not clear what the implications for confirmation bias are. Many authors continue to talk past each other, using terms like 'biased' and 'irrational' to simply mean different things. Nowhere in the literature are the different positions on what these terms mean laid out clearly so that we can see how disagreements might arise in different places, and what their implications are. My aim in this section is therefore to lay out as clearly as possible the different ways in which people might disagree about the 'rationality' of confirmation bias.

Initially, this will involve taking a step back and looking at how normative issues have been discussed and debated in the psychological literature more broadly, though I will return to examples in the confirmation bias literature throughout. At the end of the section, I will look more specifically to the question of how this impacts confirmation bias.

Of course, getting clearer on what we mean by terms like 'bias', 'rationality', and how people 'should' reason is important not just for understanding confirmation bias, but for psychology as a whole. A great deal of the psychological research conducted over the past 50 years has painted a picture of human reasoning as prone to bias and error, of human beings as irrational. The natural response to this is to ask: how might we fix these biases, how might we *improve* human reasoning? Particularly as it begins to look like aspects of human irrationality may contribute to real-world problems, the project of improving human reasoning and reducing bias is potentially an incredibly important one (Larrick, 2004, Lilienfeld et al., 2009). However, it is also a much more complex task than it first seems - there still isn't really a clear consensus on what it even *means*

to be biased or irrational, let alone an answer to whether it is possible (or desirable) to improve things.

This chapter will be structured as follows. First, I will review some disagreements in the psychological literature about how to apply different terms and normative standards, and how these disagreements affect the confirmation bias literature. In particular, I will discuss three things: how the term 'bias' has been used in different areas of research; how the normative models that provide the standard against which bias is measured are justified; and further confusion arising from different notions of 'rationality'. I will attempt to clarify where there are really substantive disagreements about confirmation bias and rationality, as opposed to mere terminological confusion. I will then ask why all of this matters - arguing that these subtleties in what it means to be biased or rational are not mere pedantry, but have important implications for how we think about the possibility of improving human reasoning. Finally, I will summarise the implications for the literature on confirmation bias more broadly.

4.2 What does it mean to be biased?

4.2.1 'Bias' in different areas of research

The term 'bias' has been used to mean different things in different contexts (Hahn and Harris, 2014, Klayman, 1995). In its everyday use, to be 'biased' generally means a *lack of impartiality* - showing an undue preference for a particular alternative or perspective (as in racial bias or gender bias, for example). The *Cambridge English Dictionary* defines bias as, "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgement." Of course, what exactly makes a preference 'unfair' - and therefore, what makes it constitute a bias - is unclear. What would it mean, exactly, for one's personal opinions *not* to influence one's judgement, at all? What distinguishes a fair reason for supporting a particular person or alternative, from an unfair one? These issues suggest that even our intuitive notion of bias is not entirely straightforward.

In some areas of social psychology, researchers have tended to refer to 'biases' in a similar way: as a tendency to express an unfair preference for a particular group or

idea, such as the 'ingroup bias' - tending to evaluate one's ingroup more positively relative to an outgroup (Mullen et al., 1992). In other parts of psychology, 'bias' means failing to conform to *general principles of rationality* - things like the 'neutral evidence principle', which says that evidence which is neutral (equally supportive of a hypothesis and its negation) should not change one's beliefs in one direction or another. This understanding of bias underlies the claim that people evaluate and assimilate evidence in a 'biased' manner (Lord et al., 1979, for example) - because they strengthen their beliefs on the basis of apparently neutral evidence.

In cognitive psychology (particularly work on 'heuristics and biases'), 'bias' is defined more precisely, as a systematic deviation from some normative model, such as probability or decision theory. For example, 'base rate neglect' is said to be a bias because people give less weight to base rates in estimating probabilities than Bayes' theorem would prescribe (Tversky and Kahneman, 1974). The difference between this and the previous notion is that 'bias' is understood as a deviation from some *formal theory* - rather than 'general principles of rationality' which are based largely on intuition. One challenge for the more intuitively-based notion of bias is that different people sometimes have differing intuitions about what is rational, and these intuitions also sometimes conflict with normative models. This suggests that we should not blindly trust what 'sounds reasonable' intuitively; the standards against which bias is measured need more thorough, perhaps formal, justification.

It is worth clarifying several points about this more precise notion of bias, and how it differs from the intuitive or social psychological notion of bias. First, *systematic deviation* means that to constitute a bias, the same patterns of error must occur repeatedly, across individuals and some range of different scenarios. This means that bias is a property we attribute to a heuristic, or some kind of decision-making strategy - not to an individual judgement or decision. Second, bias should be distinguished here from *noise* - if judgements are noisy, they might deviate on average from some normative model, but in a random rather than systematic way. Finally, bias is defined relative to some *normative model* - so what that normative model is, and what justifies its normative status, is crucial (something I will discuss in more detail later).

This is much closer to how 'bias' is understood in statistics, as a property of an *estimator*: some formula or strategy for estimating an unknown quantity. An estimator is said to

be biased if, on average, it shows a systematic pattern of errors from the 'correct answer' - the value it is trying to estimate. Again, this means bias is not something that can be applied to a single estimate - bias is a property of a procedure for estimating something, so can only be identified when the output of that procedure is observed repeatedly.

If we think of heuristics used in reasoning as 'estimators' (shortcuts for making judgements and decisions given cognitive constraints), then to claim that a bias exists is to say that a given heuristic deviates, on average and systematically, from some normative model. So if we want to claim that a bias arises in human judgement or decision making, we need to be able to specify (a) what the heuristic is that produces the bias, and (b) what the heuristic is trying 'estimate' - what the optimal solution to the problem would be. While these features are present in some discussion of bias in the psychology literature - most notably the heuristics and biases program (Tversky and Kahneman, 1974) - this is far from the standard way of talking about bias. In the literature on confirmation bias in particular, there is often little discussion of what the normative standards are for different tasks, and what heuristics people might be using that result in the claimed biases. [1]

There are two more nuances in this discussion of what it means to be biased worth mentioning. First, bias does not always necessarily come at a cost to accuracy. Sometimes a more biased strategy will be more accurate than a less biased one, if the second strategy is very high variance. This is because of something known as the *bias-variance tradeoff* in statistics: for a given estimator, there's generally a tradeoff between how biased it is (how much it errs in one specific direction), and the variance of its estimate (how much it deviates from the actual value overall, regardless of direction.) If an estimator sometimes errs in one direction and sometimes equally far in the opposite direction, its overall 'bias' might be close to zero - but it still makes large errors. By contrast, an estimator might be more biased if it tends to err in the same direction systematically, but at the same time more accurate, if the errors aren't too far off the actual value.

A nice analogy for understanding this is to imagine two types of darts player: a high variance player might not display any bias, sending darts all over the board, whereas a

[1] Why use this definition of bias rather than a more intuitive notion? Being more precise about what it means to be biased allows us to draw clearer conclusions about what these biases mean - if we want to say things about what it means to reason well or poorly, we need some kind of clear standard against which to measure reasoning. Without any kind of precise definition of bias, discussion can become confusing - as different people have different understandings of exactly what the term means, and what its implications are.

highly biased player may well be more accurate if all their darts fall in the same place not too far from the bullseye (4.1). To see how this translates to bias in judgement: consider two people attempting to estimate the probabilities of different events. One person consistently underestimates the likelihood of these events, by a similar amount each time, and so we would say they were biased. A second person displays no bias on average but their estimates are very high variance - sometimes wildly underestimating, other times wildly overestimating. We would very likely say the first person's judgement is more accurate and prefer to trust their forecasts, despite the fact they are more biased.

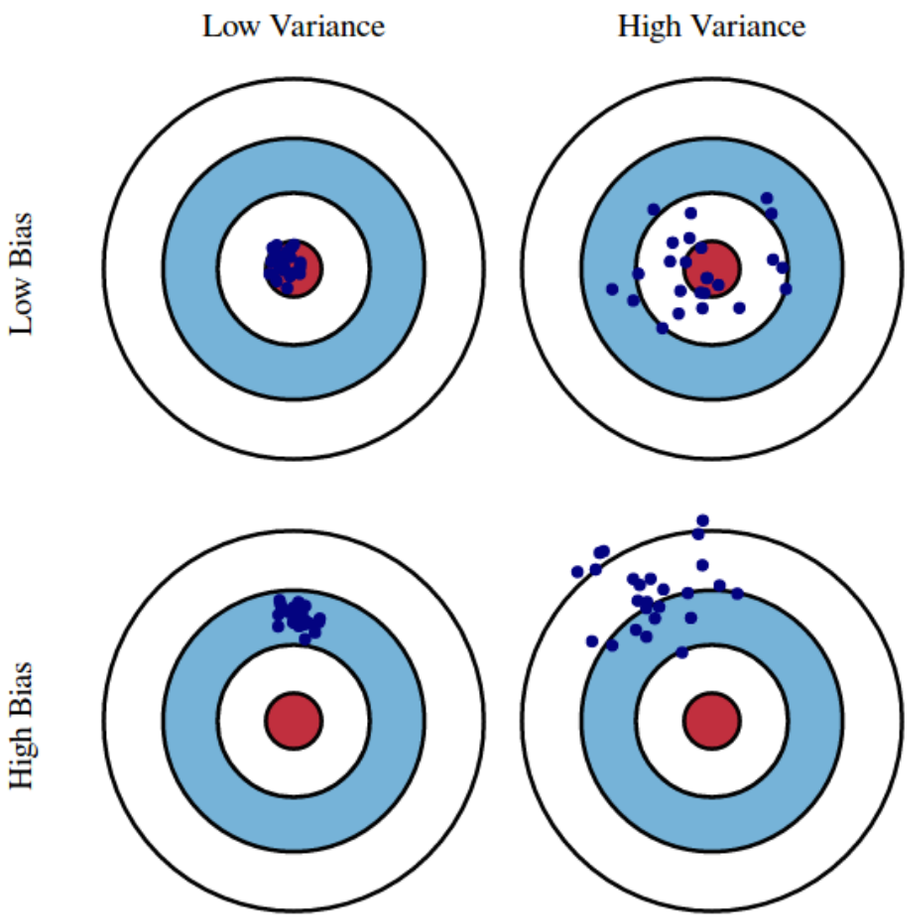


FIGURE 4.1: An illustration of the bias-variance tradeoff

A second nuance in our understanding of bias arises when we ask what kind of 'average' we are talking about in defining bias as 'deviation on average.' If we take this to be the *mean*, then it's possible for a strategy or estimator to follow a highly skewed pattern, but for the 'average' deviation to still be zero. For example, suppose the true quantity I am trying to estimate is zero, and the strategy I am using 'undershoots' 90% of the

time, but only slightly: falling around -0.1. The other 10% of the time, I overshoot: my estimates falling around 0.9. This strategy is technically 'unbiased' - but still, my estimate undershoots *far* more often than it overshoots, and so we might be tempted to say that it is 'biased' towards under-estimating. Le Mens and Denrell (2011) show that it's possible for even Bayesian rational agents to end up systematically favouring one of two hypotheses if there is an asymmetry in the information they receive about them (even if they are aware of this asymmetry in information.) This is because the distribution of judgements about the two hypotheses is highly skewed. When averaged across all individuals, the distribution of estimates has mean zero (i.e. there is technically no bias when averaging across judgements), but still a large proportion of individual judgements come out in favour of one hypothesis rather than the other. If what we care about is accuracy, calling this strategy unbiased seems a little strange, as does calling a high-variance strategy unbiased.

Klayman also distinguishes between bias as a "systematically flawed judgement process that is ultimately deleterious to the interests of the actor or society", and a "moral middle ground... people may deviate systematically from theoretical standards, but may still be behaving optimally when broader concerns are taken into account." (Klayman, 1995, p.386) This raises the question of not just whether judgements are biased in the sense of deviating from a normative model, but what the *consequences* of those biases are. I will return to this issue in our later discussion of rationality more broadly. For now, following Hahn and Harris (2014) I assume that the main consequence we are interested in is whether a bias comes at a cost to *accuracy* - a heuristic or strategy is biased if it deviates systematically from some normative standard, and does so at a cost to accuracy.

4.2.2 'Bias' in the confirmation bias literature

In the last section I distinguished three main meanings of the term 'biased':

1. A **tendency to favour** one response/choice over another (without normative or evaluative implications)
2. Failure to conform to **intuitively-based principles** of rationality
3. Systematic deviation from a **formal normative model**

These different meanings of 'bias' have been used in different places in the confirmation bias literature, underpinning some of the disagreement about whether confirmation bias 'really exists.' For example, Snyder and Swann (1978), when discussing confirmatory strategies in social hypothesis testing, equate bias with the tendency to ask more confirmatory than disconfirmatory questions - without explicitly justifying why this is irrational at all. The selective exposure literature (Hart et al., 2009) similarly seems to eschew any explicit normative standards, instead assuming that reading 'unbalanced' articles is non-normative. Here, 'bias' is being interpreted in the first, most simple, sense - as any tendency to favour one side over another.

Lord et al.'s (1979) study on biased assimilation and attitude polarization is a classic example of attributing the second kind of bias - they do not discuss formal normative models, but claim that subjects' behaviour deviates from the (intuitively compelling) rule that two people with prior beliefs should not strengthen those beliefs in different directions after reading the same evidence. We see more formal, explicit definitions of what it means to be biased and clearer normative models in the literature on pseudodiagnosticity (Crupi et al., 2009, Doherty et al., 1979), and in the literature on overconfidence, where it is more common to compare judgements to actual 'correct answers' (Moore et al., 2015).

We can also understand many of the disagreements about confirmation bias in light of these different understandings of bias. In the hypothesis-testing literature in particular, disagreement about whether a positive-test strategy should be interpreted as a bias seems to be rooted in disagreement about what the correct normative model for the situation is - with Wason's original standard being that of falsification, and others arguing for more complex normative models based in probability theory (Austerweil and Griffiths, 2008, Klayman and Ha, 1987, Oaksford and Chater, 1994). Jern et al. (2014) challenge the classical belief polarization findings, arguing that the 'neutral evidence principle' that they are based on is not always normative, and providing a more formal analysis to show how responses might sometimes be considered in line with normative prescriptions. There is some disagreement about what the 'correct' normative model is in the pseudodiagnosticity tasks (Crupi et al., 2009, Tweney et al., 2010). Le Mens and Denrell (2011) argue that it's possible for individuals' beliefs to systematically favour one of two hypotheses even under purely 'rational' assumptions (i.e. their judgements are

predicted by Bayes' theorem), because the distribution of judgements across all people can end up skewed but technically 'unbiased'.

It's also worth noting here that there's a distinction between saying that a given judgement strategy is biased and saying it is biased in the *specific* way we are interested in - i.e. towards the focal or current hypothesis. Some of the discussion about whether a positive test strategy is really a confirmation bias (e.g. Klayman and Ha, 1987) is not necessarily claiming that a positive test strategy does not lead to systematic errors - but simply that those errors do not necessarily always favour the focal hypothesis.

Disagreements about the term 'bias', if not made explicit, can therefore fuel a great deal of confusion. In particular, the main sources of disagreement seem to be the following:

- Whether 'bias' carries normative implications - or whether it simply describes a tendency or inclination that may be overall harmless (related to Klayman's 1995 distinction between bias as inclination and faulty judgement, which I discussed in chapter two);
- If 'bias' does carry normative implications, whether it should be judged relative intuitive principles, or whether standards should be grounded in more formal normative models;
- If a formal normative model, what the appropriate formal model actually is;
- What it means for a strategy to deviate 'on average' from that normative model.

To reduce some of this confusion, we could introduce some new terms: using 'bias-as-inclination' to describe 'biases' that may or may not be non-normative, and 'intuitive bias' to refer to biases that deviate from intuitive principles of rationality but may not deviate from some more formal normative standards. In what follows, therefore, I will use the term 'bias' to refer to a systematic deviation from a normative model. But this still leaves issues unresolved and room for possible disagreement - in particular, what justifies using a given normative model as the standard against which bias is judged, and what do we do when different competing normative models are proposed? The next section will consider these questions in more detail.

4.3 Normative models

I suggested that bias in judgement and decision making should be defined more precisely as a property of a given strategy or heuristic, which results in systematic deviation from a normative model, holding across a range of contexts. In chapter two, I also discussed different normative models that have been used to make attributions of bias in the confirmation bias literature. I found that confusion around normative models in general, and disagreement about the appropriate normative model, made it very unclear whether so-called demonstrations of confirmation bias showed a genuine bias (as opposed to a bias-as-inclination or intuitive bias, as I distinguished at the end of the last section.) The question of what this 'normative model' is, and what justifies its status, is therefore very important. So where do normative models in psychology come from?

4.3.1 The use of normative models in psychology

Baron (2012) overviews several different kinds of normative model used in different areas of psychology, for different kinds of judgement and decision making tasks. In the simplest cases, the normative standard is simply the objectively correct answer - if someone is using a heuristic to estimate a quantity (the population of a country, say), then we can judge whether their strategy is biased by looking at whether their estimates deviate from the known correct answer. However, for most aspects of reasoning and decision making psychology is interested in, it's much less straightforward than this - there's no single, known, 'correct' answer. If I'm trying to figure out what information is most likely to help reduce my uncertainty in making a decision, then the answer depends on what I already know, the costs of obtaining different kinds of information, and how much time I have, among other things. The question of what inferences I should draw from new information and how to update my beliefs relies similarly on what I already know, what different explanations there might be of the information, how reliable the source is, and so on.

Rather than being able to judge what the 'correct answer' is in some objective sense, what we want to do here is instead ask: what problem are people trying to solve here, and what's the optimal solution to this problem? Sometimes, we can do this formally: expressing the problem mathematically, and then calculating the solution to the problem

- providing us with formal normative models against which to judge human performance. For example, the problem of drawing inferences from new data is essentially a problem in probability theory - given that I have observed data D, what probability should I put in my hypothesis H? The formal solution to this problem is given by Bayes' theorem, which provides a formal answer for how to calculate $\Pr(H \mid D)$ given some other relevant probabilities (more on this later.) We can then judge people's inferences by how closely they follow Bayes' theorem.

A different but closely related approach to developing normative models in psychology is to ask what minimal standards we think reasoning should meet, and then find ways to formalise those requirements. For example, certain logical principles formalise the idea that beliefs should be *consistent* in certain ways - I cannot believe both P and $\neg P$, and beliefs should be closed under implicature (if I believe A, and that A implies B, I should also believe B.) More recently, normative standards for beliefs have moved away from logical principles and towards probability theory (Oaksford and Chater, 2007), on the recognition that beliefs tend to be graded, not binary, and that we have to deal with uncertainty in most of the problems we are trying to solve. The basic rules of probability theory formalise the requirement that people's degrees of belief need to be consistent in certain ways so that they are not open to exploitation. The *Dutch book theorem* (De Finetti, 1964, Ramsey, 1926) says, for each of the laws of probability theory, that violating them would leave an agent open to making bets they cannot win - the converse Dutch book theorem then says that human reasoning should follow these laws because doing so prevents people from taking self-defeating actions.

The risk, of course, with grounding normative models in minimal standards we expect reasoning to meet, is that this can start to slip into territory where people disagree about what these minimal standards are, and start invoking intuitions about more 'general reasoning principles' as discussed above. The further we move from situations where people are simply estimating known quantities, or there is a 'correct answer', the more difficult it becomes to agree how people *should* reason, what the appropriate normative model is. We may also need to start factoring in other considerations such as what the individual's goals are, what cognitive constraints they are operating under, and how features of the environment affect what the optimal strategy is - complex issues which we will come to when discussing different notions of 'rationality' in the next section.

4.3.2 Normative models in the study of confirmation bias

We have said a bit more about how normative models in psychology might be justified, and where they come from - suggesting that most such models are based in attempts to formalize mathematically the problem a person is trying to solve. More general principles such as Bayesian probability theory are justified in that they formalize the most minimal requirements we think reasoning should fulfil, in particular, some notion of consistency and not being open to exploitation.

Disagreements about whether a strategy is biased might therefore arise from disagreements about what the correct normative model for a task is. A key example of this is the debate around the correct normative interpretation of Wason's hypothesis-testing tasks - with Wason originally using logical principles of inference as normative standards, and others later arguing that the normative solution should instead be understood in terms of optimal data selection, grounded in probability theory (Austerweil and Griffiths, 2008, Oaksford and Chater, 1994), as well as taking into account constraints based on the kinds of 'real-life' hypothesis testing tasks people face.

Focusing for now just on the 'inference' aspect of confirmation bias, Bayes' rule then tells us how we should, ideally, update our beliefs in light of new evidence (see 4.1, 4.2, and 4.3 below for three equivalent formulations of Bayes' theorem).[2] The claim that people exhibit a confirmation bias (in inference) can therefore be understood more formally as the claim that people update their beliefs more towards their current hypothesis (or less away from it) than Bayes' theorem prescribes.

[2]It can be shown that obeying the laws of Bayesian probability theory has lawful connections with accuracy - Leitgeb and Pettigrew (2010a,b), for instance, argue that for a suitable measure of accuracy, Bayesianism follows from the simple premise that an agent ought to approximate the truth.

$$Pr(H \mid D) = \frac{Pr(D \mid H)\, Pr(H)}{Pr(D)} \tag{4.1}$$

$$Pr(H \mid D) = \frac{Pr(D \mid H)\, Pr(H)}{Pr(D \mid H)Pr(H) + Pr(D \mid \neg H)Pr(\neg H)} \tag{4.2}$$

$$\frac{Pr(H \mid D)}{Pr(\neg H \mid D)} = \frac{Pr(D \mid H)}{Pr(D \mid \neg H)} \times \frac{Pr(H)}{Pr(\neg H)} \tag{4.3}$$

However (as discussed earlier), many of the classic papers on confirmation bias do not refer to formal normative models at all - perhaps one of the main challenges for the confirmation bias literature has been that it tends to focus on the kinds of beliefs which do not have 'correct answers' (as opposed to, say, the overconfidence literature), making normative standards much more complex. Studies of confirmation bias have typically fallen into one of two categories to deal with this issue, as Eil and Rao (2011) point out. Those in the first category use intentionally simple, abstract tasks (such as the 2-4-6 task or having people estimate the proportion of balls of a given colour in a bag), with clear priors and objective signals that make the normative response easy to compute and compare participants' responses to (Edwards, 1982, Wason, 1960). The downside here is that these findings have unclear relevance to more 'realistic' situations. Studies in the second category, by contrast, focus on more 'realistic' beliefs, such as opinions on important issues (Lord et al., 1979, for example) - but it is very difficult to develop a normative standard for comparison here, since we do not have an objective measure of participants' prior beliefs, and the signals they receive from new information are often ambiguous (i.e. they could be interpreted differently given different background assumptions.) A few recent studies have attempted to bridge this gap, looking at how people update their beliefs about their own intelligence or attractiveness (more relevant/important beliefs than the number of balls in an urn!), given objective signals (ranking relative to others on an IQ test, for example) (Eil and Rao, 2011, Möbius et al., 2014).

Fischoff and Beyth-Marom (1983) document a number of different ways in which reasoning might deviate from Bayes' theorem - using the third form of Bayes' rule, known as the odds ratio form (4.3). (From left to right, the components of this theorem are: the posterior odds that H is true in light of data D; the likelihood of observing data D if H is

true relative to alternative hypotheses, and the prior odds that H is true before observing the data.) We summarise the potential sources of bias that Fischoff and Beyth-Marom (1983) list in table 4.1 below.[3]

Task	Potential bias
Hypothesis formation	**Hypothesis is untestable**, e.g. because it is ambiguous; Alternative hypotheses are **poorly defined**
Assessing component probabilities	**Misrepresentation**: people may give the response that is expected of them rather than what they actually believe; **Incoherence**: sometimes Pr(H) and Pr($\neg H$) may not equal one if not evaluated simultaneously, or if the beliefs themselves are not well thought-through; **Miscalibration**: failure of ones confidence to correspond to reality - overconfidence, for example; **Nonconformity with expert judgements**: due to reliance on availability or representativeness; **Objectivism**
Assessing prior odds	**Poor survey of background**: not treating the probabilities for different hypotheses equally; **Failure to assess**: i.e. base rate neglect
Assessing likelihood ratio	**Failure to assess**; **Distortion by prior beliefs**; **Neglect of alternative hypotheses**: taking the current hypothesis as a given, treating it as definitely true
Aggregation	**Wrong rule**: for example, averaging rather than multiplying the likelihood ratio and prior odds; **Misapplying right rule**: e.g. making a computational error
Information search	**Failure to search**: perhaps due to premature conviction; **Nondiagnostic questions**; **Inefficient search**: particularly failure to ask potentially falsifying questions; **Unrepresentative sampling**
Action	**Incomplete analysis**: neglecting certain consequences, for example; **Forgetting critical value**: confusing acting as if H were true (as a best guess) and actually believing H is true

TABLE 4.1: Ways reasoning can deviate from Bayes' theorem - adapted from Fischoff and Beyth-Marom (1983)

It could be helpful to ask, therefore, how these potential sources of bias correspond to the ways in which confirmation bias has been said to occur - and which of these potential errors might result in a bias towards confirming the present hypothesis. For example, based on the potential errors in the table above, a confirmation bias could result in some of the following ways:

- **People may initially overestimate** Pr(H) **relative to** Pr($\neg H$)**.** This could happen for a number of reasons: the fact that people simply have difficulty translating their subjective beliefs into probability judgements; failing to consider enough

[3]This is based on the table in the original paper - we have simply added a bit more detail to explain some of the potential sources of bias.

alternative hypotheses and so under-weighting their joint probability; or information supporting H may be more immediately accessible than the opposite.

- **People may seek out data D such that** $\Pr(D \mid H)$ **is higher than** $\Pr(D \mid \neg H)$: that is, seek out data that is more likely to support the hypothesis, and not account for this bias in search when calculating these probabilities.
- **People may miscalculate** $\frac{\Pr(D|H)}{\Pr(D|\neg H)}$: either by neglecting to calculate the denominator entirely, or perhaps underestimating the relative value of the denominator because one struggles to think of relevant alternative hypotheses.
- **In general, people may neglect** $\neg H$**, and so overestimate both items on the right hand side of the equation**: either not realising that it is important to consider the alternative hypothesis at all, assuming that their current hypothesis is essentially true, or failing to bring to mind alternatives to the current hypothesis.

Research has done relatively little to relate confirmation bias to these kinds of more specific errors in applying Bayes' rule - or to discuss what kinds of heuristics people might be using to approximate Bayes' rule that might lead to bias. We might get a clearer picture of confirmation bias by attempting to specify more clearly what these heuristics might be, and to what extent they result in incorrectly estimating various important aspects of the equation, or whether they are estimating different quantities. We can then ask whether these errors occur on average across all scenarios, and at what cost.

4.4 What does it mean to be 'rational'?

4.4.1 The relationship between bias and rationality

The terms 'biased' and 'irrational' have often been used interchangeably in the psychology literature, and their meanings often do overlap, but there are also some additional issues that arise from different meanings of the term 'rationality'. I will take it for now, based on discussion in the earlier section, that a given heuristic or strategy is *biased* if it systematically deviates from some normative model. What I will discuss in this section is the possibility that even if people broadly agree about whether a bias *exists*, they

may still disagree about whether that bias is *rational* or not. Arguments that a bias is 'really rational' often center around claims that the normative model used is inappropriate in some way, that the wider consequences of the bias are less problematic than they might seem, and/or that the strategy being used is the best possible one given various constraints.

There is no clear agreed meaning for the term 'rational' - and no clear consensus on how it relates to bias. Often debates about whether a purported bias in reasoning or decision making is 'actually rational' arise when different people use the term in different ways - one person or group claiming a given behaviour is irrational by one standard, and someone else arguing that it is not necessarily as irrational as it seems, if we accept slightly modified standards of rationality. Often in these exchanges, I'll argue, the disagreement largely comes down to two people failing to make their different interpretations of the word 'rational' explicit. I'll also try to clarify what kinds of substantive disagreements remain beyond these more terminological disagreements.

I mentioned in the earlier section on bias that there's a relevant distinction between (a) places where judgements deviate systematically from normative standards and (b) 'bias' in the sense of a systematically flawed judgement that causes negative consequences for an individual or society. A lot of the discussion around whether a given bias is rational or not seems to turn on this question of the wider consequences of the bias: taking into account the multitude of goals people have, the cognitive constraints they face, and the complex environments they are generally operating in. Something that looks like a bias when studied in an abstract lab setting and compared to a formal normative model, might be part of a strategy that actually performs well on average 'in real life', given the goals people have and constraints they face.

This is essentially the kind of argument being made by those who suggest that the positive test strategy demonstrated by Wason and others is actually more rational than it seems (Wason, 1960, 1968). Though a positive test strategy may lead to systematic errors in the very specific paradigm Wason used, when we consider the broad range of hypothesis-testing scenarios people encounter in their day-to-day lives, this strategy may actually perform relatively well.

Rather than getting caught up in whether a bias is 'rational' or what this means, it may be more helpful to ask more specific questions about the consequences of a given bias.

What kinds of costs does this bias lead to in different environments, relative to different goals, and are there alternative strategies people could use that would better achieve their goals, given the constraints they face? Rather than disagreeing about whether a bias is 'rational' or not, we can then point to more specific disagreements about which goals are relevant or important, what consequences a given heuristic leads to in different environments, and whether, given cognitive constraints, it's possible to improve upon existing strategies.

In this section, I will overview some different ways that the term 'rationality' has been used - recognising that some of these different types of rationality overlap with one another and are vaguely defined - before suggesting some clearer ways we might think about these issues.

4.4.2 Different types of rationality

Table 4.2 summarises the main different ways that the term 'rationality' has been used in the psychology literature, which I expand on in more detail below.

Type of rationality	Summary	Links/overlaps	Questions
Normative rationality (Elqayam and Evans, 2011)	Conforming to a formal normative model e.g. Bayesian probability theory (Oaksford and Chater, 2007)	Is essentially a minimal form of instrumental rationality - where the goal is consistency/avoiding exploitation (Oaksford, 2014, Stanovich, 2011)	Do people reason in ways that systematically deviate from Bayes rule, in a way that favours the current hypothesis?
Epistemic rationality (Stanovich et al., 2008)	Reasoning in ways that lead to accurate beliefs	Specific type of instrumental rationality, though sometimes contrasted with it (contrasting accuracy with other, more personal, goals)	Does a confirmation bias (if it exists) come at a cost to accuracy on average?
Instrumental rationality (Stanovich et al., 2008)	Reasoning in ways that lead to effectively achieving ones goals	Suggests normative models should take into account additional/different goals	If a confirmation bias exists, does it hinder progress towards goals? Might confirmation bias be considered rational with respect to goals other than accuracy?
Bounded rationality (Gigerenzer and Goldstein, 1996, Simon, 2000)	Reasoning that is optimal (relative to some standard - formal model or some goal) given cognitive constraints	Suggests normative models should factor in realistic bounds on cognition	Do people reason in ways that systematically deviate from normative models, if those normative models take into account cognitive constraints?
Ecological rationality (Todd et al., 2000, Todd and Gigerenzer, 2007)	Reasoning that is optimally adapted to the specific environment	A form of instrumental rationality - which says that goals depend on the environment. Suggests normative models should be narrower, more specific	Might a confirmation bias, if it exists, be well-adapted to certain kinds of environments, and the goals of those environments?
Evolutionary rationality (Tooby and Cosmides, 1992)	Reasoning that is adapted to furthering genes of the organism	A form of instrumental rationality where goals are construed at the level of genes, not the organism	Might a confirmation bias be adaptive from an evolutionary perspective?
Prescriptive rationality (Stanovich and West, 2000)	Strategies that people could realistically use (given cognitive constraints) to make better judgements (by some standard)	Attempts to generate normative standards that are also prescriptive - i.e. can be followed given cognitive constraints	Are there alternative reasoning strategies people could actually use, that would do better by some agreed normative standard?

Table 4.2: Different types of rationality

4.4.2.1 Summary of different types of rationality

Normative rationality (or 'rationality' simpliciter) says that behaviour or reasoning is rational if it conforms to the prescriptions of a formal normative model (e.g. inference is rational if it follows Bayes' rule.) This is essentially the same as saying that reasoning

is unbiased - and so this notion of rationality doesn't draw any clear distinction between what it means to be biased and what it means to be irrational.

One first way in which people might disagree about what is rational is, therefore, that they might disagree about what the correct normative model is. As Elqayam and Evans (2011) point out, it is becoming increasingly rare to find 'single norm paradigms' in reasoning and decision making research - tasks where a single normative model is undisputed. Evans (1993) refers to this as the 'normative system problem', and Stanovich (2011) similarly talks of the 'inappropriate norm argument'. This fuels some of the disagreement I discussed in chapter two - around what the correct normative standard against which to judge confirmation bias is.

To say that a reasoning strategy is **epistemically rational** is to say that it reliably leads one to form accurate beliefs about the world (Stanovich et al., 2008). This essentially adds the importance of accuracy to the basic notion of normative rationality - in addition to asking whether a strategy leads to systematic deviation from a normative model, we also want to ask whether those deviations come at a cost to accuracy. As I discussed in the earlier section on bias, because there is a tradeoff between bias and variance, it's possible for a reasoning strategy to be systematically biased and yet still be more accurate than alternative, higher variance, strategies.

Another notion of rationality commonly discussed is that of **instrumental rationality**, which is based on the idea that rationality should take into account the various different goals an agent may have (Stanovich et al., 2008). To determine whether a strategy is instrumentally rational, therefore, we need to ask not just whether it deviates from some normative model, but whether it actually comes at a cost to important goals. A heuristic might seem to result in bias relative to an abstract normative model, but yet be highly effective at achieving certain goals. For example, though a tendency to ignore or underweight unpleasant information might be viewed as irrational by Bayesian standards, it might be instrumentally rational if the goal is to maximise personal utility, at least in the short term. Buchak (2010) argues that it is not always instrumentally rational for a risk averse decision maker to seek out more information before making a decision - even though seeking more information might be considered the normatively rational response.

Of course, there's a lot of room for disagreement here about what the relevant goals are, and perhaps what goals people *should* have in different situations. Our reasoning and decision making strategies face conflicting goals on multiple levels - there's often a conflict between my personal goals and the 'goals' of my genes (what is evolutionarily adaptive isn't necessarily what's best for me - see Stanovich, 2009), a conflict between my short-term and long-term interests, and conflicts between the goals of individuals and larger groups. Which of these goals we should define 'rational' behaviour relative to when conflicts arise is not an easy question to answer.

Discussions of **bounded rationality** (Gigerenzer and Goldstein, 1996, Simon, 2000) emphasise the importance of taking into account the cognitive constraints people face when making judgements and decisions: we clearly do not have the computational power, or time, to always be calculating Bayes' rule, and so proponents of bounded rationality argue that many normative models are simply inappropriate standards by which to judge reasoning. A given strategy is boundedly rational if it is effective at achieving the relevant goals (whether accuracy or other goals) given these cognitive constraints - or put differently, if there is not clearly any alternative strategy that would do better that people could feasibly use. So even if I err towards interpreting ambiguous information in ways that favor my current hypothesis, the only way for me to avoid this may be to consider multiple different counterfactuals and alternative hypotheses, going far beyond the cognitive capacity and time I have available.

Proponents of the closely-related notion of **ecological rationality** similarly suggest that the normative models typically used are inappropriate standards by which to judge human rationality, because they fail to take into account the relevant features of the environments in which people are making judgements and decisions (Todd et al., 2000, Todd and Gigerenzer, 2007). Proponents of an ecological theory of rationality suggest that we need to study human behaviour and judgement in real-world domains, observe what heuristics and processes they use, and then assess whether this enables them to get the 'correct answer', and/or to successfully attain their goals in those domains. This contrasts with the classic heuristics and biases approach, which typically studies judgement and decision making in much more general and abstract contexts, and in comparison to much stricter normative standards.

Evolutionary rationality focuses specifically on whether strategies are well-adapted

for evolutionary purposes: a given strategy might look 'biased' relative to some normative model, but be rational from the standpoint of the genes (Tooby and Cosmides, 1992). For example, some have argued that a kind of confirmation bias might have been beneficial in the ancestral environment: much better to believe that a predator is around the corner and be wrong than the opposite.

Finally, **prescriptive rationality** asks not just whether reasoning deviates from some normative standard in theory, but whether in practice there are alternative strategies people could actually use to move closer to these normative standards (Stanovich and West, 2000). Just because Bayesian probability theory provides the correct formal solution to the kinds of inference problems people are often trying to solve, doesn't mean people should literally use Bayes' rule when drawing inferences - the cost of the time and effort involved might well outweigh the benefits (in this sense, this is closely related to bounded rationality.) What we can ask, however, is whether there are strategies people can use that are different from those currently or automatically employed, which might bring judgements closer to these normative standards - and these would provide prescriptive standards for rationality. For example, the prescription to 'consider the opposite' under some circumstances might help people to more accurately assess the diagnosticity of a given piece of information, by helping them to consider how likely it is under an alternative hypothesis that they might otherwise ignore (Lord et al., 1984).

One useful way to understand the distinction between prescriptive and normative rationality is by saying they operate at different *levels of description* (Marr, 1982): normative rationality operates at the 'computational level' (i.e. it describes the kinds of problems reasoning is trying to solve, without saying anything about the actual algorithms that implement those solutions in the brain), where prescriptive rationality operates at the 'algorithmic level' (trying to say something about the actual algorithms or strategies people might use to solve problems.)

4.4.2.2 Links and overlaps between different types of rationality

There are a number of links and overlaps between these different notions of rationality (summarised earlier in table 4.2). Both epistemic and evolutionary rationality might be thought of as specific types of instrumental rationality, where the primary goal of interest is made explicit. This also highlights that people might disagree about whether

certain goals are 'more rational' than others: proponents of epistemic rationality might argue, for example, that reasoning *should* help us to reach true conclusions, and any reasoning strategies that do not should be classed as irrational - even if they serve other goals. Proponents of evolutionary rationality might argue that so long as reasoning does what it evolved to do - to further the genes of the organism - then we shouldn't hold it to a higher standard. Others might argue that rationality is always a relative notion, defined in relation to some goal, without saying anything about what those goals should be.

Bounded and ecological rationality both suggest that there is additional information we should take into account when applying normative standards to human reasoning - that we should factor in cognitive constraints and/or relevant environmental features when modelling the problem people are trying to solve. Normative models that do not take into account these features are thought to be either putting unrealistic standards on human reasoning, or simply incorectly framing the problems they are trying to solve.

The idea that epistemic rationality says something beyond normative rationality - that a strategy might deviate from a normative model but not at a cost to accuracy - also seems to implicitly assume some notion of bounded rationality. If there were no constraints on human reasoning, presumably we could simply apply normative models perfectly, and so deviations from those normative models would always come at relative costs to accuracy. It is only when we assume that reasoning is imperfect and the strategies we use have to navigate tradeoffs, that we start to see situations where an increase in bias may not decrease accuracy relative to other feasible strategies. Similarly, any theory of prescriptive rationality must also implicitly acknowledge constraints on reasoning and ecological factors, in order to come up with realistic prescriptions for strategies people can actually use to improve their reasoning. Prescriptive rationality differs from bounded and ecological rationality, however, in that it does still assign some normative status to the more abstract normative model - acknowledging that this provides the relevant benchmark against which to compare human reasoning (where bounded and ecological rationality would argue this is simply the wrong benchmark), but accepting that this benchmark doesn't actually provide actionable prescriptions. It's unclear how much difference this makes in practice, however.

Even normative rationality might be considered a minimal form of instrumental rationality

- raising the question of whether it ever makes sense to talk about 'rationality' in an unqualified way at all. Elqayam and Evans (2011) object to the standard picture of normative rationality, pointing out that it is often not clear what the appropriate normative standard is, especially in increasingly complex areas of reasoning, where various competing norm paradigms are often proposed. They argue that research programs attempting to establish normative models risk drawing a controversial is-ought inference: justifying how people *should* reason on the basis of how they *do in fact* reason. Oaksford (2014) and Stanovich (2011) respond that this distinction between normative and instrumental rationality is much more permeable than Elqayam and Evans' argument requires - what appear to be unconditional statements about rationality, are actually just conditional statements with a very broad antecedent (so broad that it is often assumed and not stated.) Stanovich points out that the standard justification for using probability theory as a normative model is that violating the laws of probability theory leaves an agent open to exploitation: open to making bets one cannot win (the *Dutch Book Theorem*, Vineberg (2011), as discussed earlier.) In this sense, we might consider normative rationality a minimal kind of instrumental rationality - with the minimal assumption that people want to reason in ways that prevent them from being open to exploitation or making bets they are bound to lose.

4.4.3 Disagreements about rationality: summary

What all of these different notions of rationality are trying to do is provide some standard against which to assess reasoning - disagreements arise about rationality because there are multiple ways to disagree about what that standard should be. To a large extent, this is all just disagreement about what the word 'rational' means, but some more substantive disagreements may still remain. Before asking how disagreements about rationality have influenced the confirmation bias literature, therefore, I will first attempt to clarify where the more substantive disagreements about rationality lie - and what is just terminological confusion. Which disagreements would be resolved if people could clarify that they were talking about different 'types' of rationality, and which would remain?

4.4.3.1 Substantive disagreement or terminological confusion?

In the earlier section on normative models, I talked about how we might best think of normative models in terms of (mathematical) formalizations of the problems that people are trying to solve: the normative model against which we judge reasoning or behaviour is then the optimal solution to that formalised version of the problem. For example, Bayes' rule might be thought of as the formal solution to the problem of updating beliefs based on evidence under uncertainty.

Similarly, I think it can be helpful to view judgements of rationality as judgements about how well someone is solving a given problem. Disagreements about rationality can therefore arise in two ways: disagreements about what the solution to a given formal problem is, or disagreements about what the appropriate formalisation of the problem is in the first place. I think most disagreements about rationality are actually of the latter kind: about what the appropriate formalisation of the problem is, even though they often look more like the former: disagreements about the right solution to the problem. Given a mathematical formalization of a problem, there's simply not much room for disagreement about what the solution is (which is not to say that finding the solution is easy.)

I think a lot of disagreement can arise, therefore, simply because researchers don't realise this: they think they're arguing about different solutions to the same problem, but actually they're solving subtly different problems. And if they were able to clarify the ways in which they're framing the problem differently, they could agree that yes, given the other's formulation, their solution makes sense. For example, when someone claims a behaviour is 'actually rational' based on arguments about the specifics of the environment a person is operating in, the cognitive constraints they're operating under, or assuming certain goals, they are basically building in additional assumptions to the problem - and asserting that, given these assumptions, the solution looks different. The person who originally claimed the behaviour was irrational can accept that yes, given these additional assumptions and different framing of the problem, the solution is different and so the behaviour is not so irrational by different standards. This can also be thought of as essentially terminological confusion - confusion about the 'kind' of rationality that is being referred to. To some extent, we may be able to simply accept

that there are different notions of rationality, different ways to formulate the problems people are trying to solve, and do our best to distinguish between them.

However, I may agree with you that given how you have formulated the problem, your solution is correct (given your definition of rationality, the behaviour is rational), but still disagree with your formulation of the problem (disagree that your definition of rationality is appropriate.) Give our discussion of different types of rationality, I think there are three main ways in which this disagreement can arise. There is, however, considerable overlap between these different dimensions of disagreement, and they might even better be thought of as subtly different ways of thinking about what is essentially the same disagreement.

1. **Disagreement about the level of generality at which problems should be framed**:

   For example, theories of normative rationality use broad, abstract models such as probability theory as the standard against which to judge reasoning on a very wide range of problems. These are construing the problem people are trying to solve very broadly - in terms of adhering to very general principles or standards (such as consistency.) By contrast, theories of ecological or bounded rationality are essentially suggesting that we should frame the problems people are trying to solve much more narrowly: factoring in specifics of the situation, and judging rationality in terms of how well people solve a specific problem, not how well they solve a range of problems in general.

2. **Disagreement about the relevant goals against which rationality should be judged**:

   Theories of normative rationality either seem to assume no specific goals at all, or only the most minimal goals that all people could be expected to have across all situations, such as consistency/avoiding exploitation. Theories of epistemic rationality suggest that the main goal against which rationality should be judged is accuracy, whereas theories of instrumental rationality suggest that rationality should be more often judged relative to whatever the goals of the agent are in a given situation - implying that goals should be more specific than normative rationality suggests, and more subjective than epistemic rationality suggests. Theories

of evolutionary rationality imply that the relevant goals are evolutionary ones, goals at the level of the gene - and we should judge a person's rationality relative to whether their behaviour is evolutionarily adaptive. We might also imagine further distinctions where rationality is judged relative to goals at different levels: rationality relative to more short-term versus long-term goals, rationality relative to individual-level goals versus goals at a larger group or societal level.

There is certainly some room to 'agree to disagree' here - we can simply say a behaviour is rational with respect to one goal and not to another. But I think there is also some substantive disagreement over whether the concept of rationality should, in the most basic sense, take as given certain goals, or prioritise certain goals over others when they come into conflict.

3. **Disagreement about how much we should factor in various constraints (environmental and cognitive)**:

   Theories of normative rationality take very little, if any, account of the constraints people actually face when making judgements/decisions - whereas theories of ecological and bounded rationality suggest that taking these constraints into account is crucial for formulating the relevant problem. Again, there may be room to simply agree that these are describing different types of rationality - and this also seems closely related to point 1.: whether rationality is defined relative to a specific narrow problem or more broadly. I think the substantive disagreement here is whether the notion of rationality should take into account cognitive constraints - which I think is essentially a question about whether the standards against which we judge human reasoning should be realistically attainable or not. Theories of normative rationality seem to be under no pretense of providing actual processes that people could realistically follow. Whereas theories of prescriptive rationality (closely related to ecological and bounded rationality), suggest that we should formulate the problem in a way that's realistically solvable by humans, and therefore judge human reasoning relative to a standard that's realistically attainable.

At root, I think the main disagreement may come down to whether there is any one thing we can call rationality, simpliciter, without it needing to be expressed in relative or instrumental terms. Is there such a thing as being simply 'rational', or are all attributions of rationality simply saying behaviour is rational relative to some goal or given

some specific problem one is trying to solve? We often talk about rationality without qualification, suggesting we think of it in this broad sense - but it's not clear exactly even what this would mean, and as Elqayam and Evans (2011) argue, the idea of an unqualified notion of rationality raises problematic issues around drawing is-ought inferences. At best, we might follow Oaksford (2014) and Stanovich (2011) and suggest that a broad kind of rationality is one which is judged relative to very general problems and the most minimal goals and assumptions. But even this kind of rationality is relative.[4]

I suspect that some substantive disagreement will remain, nonetheless, over what it is appropriate to call 'rational'. I'd find it hard to accept someone using the term 'rational' to describe how effectively someone was achieving the goal of self-deception, for example, even if strictly speaking they were behaving rationally with respect to that goal - our intuitive concept of rationality is certainly more closely tied to certain goals than others. However, I think we could make much greater progress understanding human reasoning and how it goes wrong, if claims of 'rationality' were made more specific - if, when attributing rationality, we could clarify whether this is meant in a broad or more specific sense, and talked more in terms of the specific problems people are trying to solve either well or poorly.

4.4.3.2 Disagreements about rationality in the confirmation bias literature

I will now discuss a few examples of how disagreements about rationality have led to disagreements within the confirmation bias literature, beyond some of the examples already discussed earlier in the section on bias.

As I summarised in the last section, many disagreements about the 'rationality' of confirmation bias might be understood in terms of different construals of the problem being solved. On the broadest level, the relevant problems are how to choose new information to test hypotheses, and how to draw inferences from that information to update existing beliefs. From this standpoint, the relevant normative models, understood as the formal

[4]Though consistency might naturally be thought of as the most minimal standard rationality should meet, in practice this might not be as minimal a requirement as it seems. Given the wide range of situations we encounter, keeping our beliefs and decisions consistent across those situations may be an incredibly demanding task. Sometimes it may be that some degree of inconsistency is a price worth paying to save time/energy, or to better achieve some other goal. This helps explain why normative rationality, though in a sense the 'simplest' type, is also often the most demanding.

solutions to these problems, are Bayesian inference and optimal data selection. A confirmation bias then exists and is a sign of irrationality if people's reasoning processes systematically deviate from these normative standards, and do so in a way that favour prior beliefs. In an earlier section, we listed a number of different ways judgements might deviate from Bayes' theorem that could lead to a confirmation bias, building on discussion in Fischoff and Beyth-Marom (1983). A confirmation bias in this sense also seems likely to be epistemically irrational - i.e. to come at a cost to accuracy.

However, there are other ways we might understand the rationality of a confirmation bias. People often have other important goals than accuracy and consistency, and it's possible that in some of the environments people typically encounter with the cognitive constraints they face, the strategies they use may be effective at achieving those goals. For example, some have argued that forms of confirmation bias, though irrational from a purely epistemic perspective, might help people to achieve other goals - such as protecting one's ego (Hart et al., 2009), or mental health (Nickerson, 1998). Though strategies like falsification might be normatively appropriate in certain abstract rule-discovery experiments, something more like a positive-test strategy might be ecologically rational given features of the kinds of hypotheses we typically have to test. Perfors and Navarro (2009) bring together some different perspectives on the rationality of hypothesis-testing behaviour by suggesting that a positive-test strategy may be rational on the assumption that hypotheses are 'sparse': that they are rare - true for less than half of the logically possible entities (Oaksford and Chater, 1994) - or in the most extreme form deterministic rules that only predict a single possibility at each trial (Austerweil and Griffiths, 2008).

Taking into account cognitive constraints, it has also been argued that some forms of confirmation bias might optimally balance the costs of different kinds of errors. Nisbett and Ross (1980) point out that, given practical time constraints, the tendency to persevere in one's current hypothesis might be a 'stabilising hedge' against changing one's mind too frequently. Friedrich (1993) suggests that human inference processes are designed to identify potential rewards and minimize costly errors, a very different task from pure truth-detection - given this goal and cognitive limitations, he argues, a confirmation bias may be viewed as rational. Friedrich gives the example of an employer with the hypothesis that extroverts make good salespeople. If his main goal was testing the truth of this hypothesis, then he would want to seek potentially disconfirming

evidence by trialling an introverted salesperson. But in practice, false-positive errors (hiring an introvert who turns out to be a poor salesperson) are much more costly than false-negative errors (hiring an extrovert who is good, and missing out on an introvert who would also have been good.) So seeking to minimise the probability of committing the former type of error, by seeking to confirm one's hypothesis, might be rational from a more pragmatic perspective.

Tooby and Cosmides (1992) have emphasised understanding confirmation bias from an evolutionary perspective - suggesting that selection pressures would favour strategies that solved biologically significant problems rather than those that were perfectly consistent or truth-seeking. So we might similarly expect mechanisms that minimise false-negative errors (undetected predators) by tolerating false positives (assuming all bears are dangerous) to out-reproduce mechanisms more driven by falsification testing.[5]

It may also be that the strategies we use are not so effective at achieving the relevant goals, but since calculating probabilities and value of information and Bayes' rule is incredibly cognitively demanding, we need to come up with realistic strategies that people could use that would lead to better outcomes (prescriptive rationality.) For example, some studies have found that simply asking people to consider-the-opposite - to consider hypotheticals such as how they would interpret evidence if they believed the opposite of what they currently believe, say - can minimise supposed confirmation biases (Lord et al., 1984). Mckenzie (2004) uses Monte Carlo simulations to investigate the accuracy of several intuitive strategies for inference, and finds that some of them perform almost as well as the normative prescriptions of Bayes' rule, even though they are simpler, more intuitive strategies. In particular, the 'relative likelihood average' strategy, which involves ignoring base rates and simply estimating the relative likelihood of new data under alternative hypotheses, and then averaging this with the base rate of the focal hypothesis, correlates almost perfectly with Bayes' rule (of course, this is still a fairly complex strategy to actually implement and so perhaps not prescriptive - but it does demonstrate that following strict normative standards is not necessary to get very close to 'optimal' performance.)

[5]Note here though, that acknowledging how evolutionary pressures have influenced reasoning is very different from saying that this makes behaviour *rational* - this requires the further step that we define rationality relative to evolutionary goals.

4.4.4 Why does this matter? Rationality and improving human reasoning

One might naturally ask how much these different meanings of terms like 'rational' really matter - isn't this just pedantic quibbling over definitions? I think it does matter, firstly because it can help us to better understand many of the disagreements that arise in the psychology literature about whether a given behaviour is 'actually rational' or not. Clarifying different uses of these terms can help us to clarify where disagreement is just terminological confusion, and where the more substantive disagreement arises, as we have discussed. But I think it also matters for something even more important than this: because clarifying whether a given behaviour is 'biased' or 'irrational', and what this really means, has implications for the possibility of improving human reasoning.

Given the complexities of the normative issues we've discussed so far, it's not surprising that this question - 'can, and should, we try to improve human reasoning?' - does not have a straightforward answer, and has generated a fair amount of debate. There are various different viewpoints on this issue, which arise from differing views on the following questions:

- What is the appropriate normative standard against which to judge human reasoning?
- Where and to what extent do people's reasoning strategies fall short of this standard?
- Is it possible, within the constraints of cognitive and biological capacity, for people to use better strategies as judged by this standard?

These questions are closely related to the disagreements raised in the previous section on rationality - whether or not one believes that it's possible to improve reasoning will depend on whether one thinks there are standards people are failing to meet which are realistically attainable. Another way of framing this, linking back to the earlier section, is to ask whether there are problems that people could solve better or more effectively if they were to reason in different ways.

If we think of rationality too broadly, in terms of just the most minimal goals, and trying to solve very general problems - as theories of normative rationality tend to -

then we are likely to arrive at the conclusion that people frequently fall short, but not feel particularly optimistic about people's ability to do any better. For example, Tversky and Kahneman (1974) take the view that reasoning falls short of these very broad normative standards, and Kahneman in particular seems sceptical that there is any way to really 'debias' judgement against these errors, saying, at the end of *Thinking Fast and Slow* that, "little can be achieved without a considerable investment of effort... Except for some effects that I attribute mostly to age, my intuitive thinking is just as prone to overconfidence, extreme predictions, and the planning fallacy as it was before I made a study of these issues." (Kahneman, 2011, p.417)

By contrast, if we think of rationality too narrowly, then there is always some way that we can frame people's behaviour as rational, by specifying more and more constraints and assumptions people are operating under - and so the idea of improving human reasoning becomes simply unnecessary or undesirable. For example, some evolutionary psychologists believe the appropriate standard against which to judge human rationality is adaptiveness on the level of the genes (Tooby and Cosmides, 1992), and others argue that human reasoning can be viewed as basically optimal when we look at how well-adapted it is in specific domains (Gigerenzer and Goldstein, 1996, Todd et al., 2000).

I think both of these perspectives go too far. If we construe rationality too broadly, we are expecting too much of human reasoning - we cannot be expected to have general-purpose reasoning strategies that work well without error across a broad range of situations, given the constraints we're operating under and different scenarios we're operating in. Though it's interesting and useful to ask how we fare relative to this 'broad' notion of rationality, I'm not sure this is the appropriate standard against which to assess the feasibility of improving human reasoning. But if we construe rationality too narrowly, then we may be asking too *little* of reasoning: I think it's highly unlikely that human reasoning is as well-adapted as some proponents of bounded or ecological rationality suggest. I think we do want to judge human reasoning against standards other than purely evolutionary ones, for example. There are various different goals we might care about as individuals and as part of larger groups, some which will be pretty much universally accepted, some which might conflict with one another (my short-term and long-term goals might conflict, or my personal goals might conflict with those of a group I'm part of, for example). Given that each person has various different, often competing, goals in different scenarios, and that different people's goals often come into conflict with one another, it seems

very unlikely that our natural judgement and decision-making processes are the best they could possibly be: that each of us couldn't realistically achieve our personal goals better, that different strategies couldn't help us to better navigate tradeoffs and conflicts within groups of people. Especially given how many people in this world seem to be unsatisfied, misinformed, or both, it seems to me hard to deny that there is room for reasoning to be improved in various ways, even given biological and cognitive constraints.

However, the task of improving human reasoning is much more complex than simply identifying biases relative to formal normative models and then somehow trying to 'fix' these biases. Rather than simply showing how reasoning deviates from broad normative models, we need to understand the constraints that people face in reasoning, what different goals they have, the tradeoffs that arise given those constraints and goals, and how people navigate those tradeoffs. Once we understand this we can then ask whether people could navigate those tradeoffs better in certain ways. We need to take into account all of these things - but be careful not to fall into the trap of being able to describe any behaviour as rational, construed narrowly enough. In many ways this is much more complex and difficult a project than the standard normative rationality project. Understanding how different reasoning processes actually result in errors we care about, in the real world, is far from straightforward - but ultimately much more important and potentially useful.

It might be more useful, therefore, to focus on much more specific questions than "Can we make people more rational?". Questions like "Where do the strategies people use seem particularly ill-suited to their goals?", "Where does this actually cause problems in the world?", "What alternative strategies might people realistically use in these contexts", or "How could judgements and decisions be made easier for people?"

When it comes to confirmation bias, the suggestion here is that we might be better off asking first what kinds of strategies people use to seek out and draw inferences from information, and then ask in what contexts and relative to what goals those strategies might perform particularly well or poorly.

4.5 Summary

I have discussed a number of different interpretations of what it means to say that someone is biased or irrational, and some of the disagreements and confusions that arise from different uses of these terms. I will now briefly summarise this discussion and what seem to be the implications for confirmation bias.

4.5.1 What does it mean to be biased?

Colloquially, the term 'bias' is generally taken to mean a lack of impartiality, and often used interchangeably with the term 'irrational.' Hahn and Harris (2014) make the case for using a more precise definition of bias: bias as a property of an estimator (or heuristic), which occurs when that estimator deviates systematically from a normative standard. To say that reasoning is biased in a certain way, therefore, it is not enough to observe a single judgement: bias occurs when the same reasoning process leads to a systematic pattern of error on average, relative to some clearly defined standard.

I also discussed how bias isn't necessarily always the same as inaccuracy: given the constraints we are operating under, it's possible that bias might be the optimal solution to some tradeoff (as we see with the bias-variance tradeoff in statistics.) Even if we tend to use reasoning processes that are biased, there may not be alternative processes that we could feasibly use which would be more accurate without incurring other costs. This means we need to ask not just whether a reasoning process or heuristic is biased, but what costs that bias incurs - and what costs an alternative strategy might incur.

In the confirmation bias literature, we can understand some confusion and disagreement as stemming from different uses of the term 'bias.' In the social hypothesis testing and selective exposure literature, for example, the term 'bias' refers to a general tendency or preference - rather than bias in the strict, systematic deviation sense. Lord et al.'s (1979) study on biased assimilation demonstrates that people fail to conform to intuitively-based rational principles - but also does not talk about bias in a strict sense. In relatively few places in the confirmation bias literature is 'bias' talked about in the more precise sense - as systematic deviation from some normative standard.

4.5.2 Normative models in psychology

I next discussed the use of normative models in psychology - and how these standards are justified. I suggested thinking about normative models as attempts to formalise the problem people are trying to solve, and to then provide ideal solutions to those problems against which we can compare people's actual judgements and behaviour. We might also ground normative models in certain minimal requirements we expect reasoning to meet - following the laws of probability theory, for example, captures the minimal requirement that reasoning is consistent and does not leave one open to exploitation.

In the study of confirmation bias, Bayesian inference is generally accepted as the appropriate standard, where normative models are discussed. However, a large number of papers on confirmation bias do not discuss formal normative models at all. There is a tension between studies which are sufficiently simple and abstract that it's easy to compare human judgement to normative standards - and those which have greater relevance to important judgements in real-world domains. Navigating this tradeoff is particularly difficult since confirmation bias research has tended to focus on questions which don't have clearly-defined 'correct' answers.

4.5.3 Different types of rationality

A simple, often-used definition is that to be rational is for one's reasoning to conform to the standards of some normative model - I called this 'normative rationality.' By this definition, bias is just the same thing as systematic irrationality. However, this notion of rationality raises all kinds of questions about how we determine the correct normative model - with various aspects of reasoning and decision making, it's not totally clear what the appropriate normative standard is. Some, such as Elqayam and Evans (2011), have challenged whether it really makes sense to talk about normative rationality at all, suggesting that this involves a contentious *is-ought* inference. Oaksford (2014) and Stanovich (2011) imply we might think of normative rationality as being a very minimal kind of rationality - that is, how one should reason assuming one has very basic goals such as consistency and not making bets one is certain to lose.

An alternative way of thinking about rationality is as instrumental to certain goals - to be 'instrumentally rational' is to reason in ways that best lead one to achieve one's goals,

whatever those goals are. This is sometimes contrasted with 'epistemic rationality': reasoning in ways that lead one to form accurate beliefs about the world (though we might also consider this a specific kind of instrumental rationality, as having accurate beliefs might be thought of just one goal a person could have.) However, epistemic rationality seems like a particularly important kind of rationality, since rationality has often been considered closely related to accuracy - and it might be argued that a rational person should prioritise accuracy very highly amongst their goals. This is potentially the source of some disagreement: some claiming that behaviour is irrational by accuracy standards, and others arguing it is 'really rational' if we assume other, non-accuracy goals. It seems, however, that we could simply distinguish between instrumental and epistemic rationality more explicitly and dispel a lot of confusion, without necessarily having to resolve whether one is more important than the other.

Others believe that a notion of rationality should more explicitly take into account the cognitive constraints people operate under ('bounded rationality'), the context in which behaviour evolved ('evolutionary rationality'), or the specific demands of the environment/context within which people are operating ('ecological rationality'.)

A lot of disagreement - in the confirmation bias literature and beyond - seems to stem from people disagreeing about which of these different notions *really* means 'rationality.' Some have argued that confirmation bias is 'actually rational' if we assume that people are optimising for different goals than accuracy (such as preserving one's self-worth), or that it effectively navigates tradeoffs we have to deal with in many real-world environments, such as balancing the costs of different types of errors.

I suggested that it might helpful to stop using the term 'rationality' simpliciter, and instead explicitly label these different kinds of rationality as different standards on reasoning. Instead of getting caught up in different meanings of rationality, we could then more simply ask different questions about the consequences of a confirmation bias, if it exists: in what environments and for what kinds of problems might a confirmation bias effectively navigate difficult tradeoffs given cognitive constraints, for example?

4.5.4 Bias, rationality, and improving reasoning

Finally, I discussed the implications of all of this for the project of improving reasoning, or trying to make people 'more rational.' The main questions here seem to be (a) whether reasoning is suboptimal relative to some standard - whether that's a normative model or some more instrumental goal/outcome, and (b) whether we think it's possible, within biological constraints, for people to move closer to this standard. Some argue that reasoning is suboptimal relative to formal normative models, but fail to say enough about how, prescriptively, reasoning could move closer to these standards. Others retort that since we can't literally reason using Bayes' rule, defining rationality by these standards is unreasonable - but go far in the other direction, lowering the standards to a point where human reasoning can be called totally normative, and there's no room or need for improvement.

It seems very unlikely to me that there's no room for improvement in human reasoning and decision making - but it's also not clear that teaching people to better understand Bayes' rule is the answer. I suggested we might be better off focusing more on instrumental and epistemic rationality - asking what goals people have in different scenarios, and how default processes of reasoning and deciding might hinder progress towards those goals. We can then ask whether different, realistically learnable strategies might lead to improvements relative to given goals without introducing new tradeoffs, and whether the benefits are worth the cost of teaching those strategies.

4.6 Implications for confirmation bias

In chapter two, I discussed how the evidence for a confirmation bias is much more mixed than it might first seem, owing largely to a lack of clear normative standards in the studies often considered evidence of confirmation bias. In this chapter, we've looked in more detail at this issue, outlining some of the different ways normative standards might be applied, and some of the disagreements that arise from different views about what it means to be 'biased' or 'irrational'.

All of this suggests that the case for confirmation bias is yet more complex than it seemed at the end of chapter two. Not only is it unclear from existing research whether

people are systematically biased in favour of their current beliefs, this discussion has raised additional issues. Even if people are biased in certain systematic ways to favour what they already believe, to appreciate the wider consequences of this we need to ask a number of questions about the impact this has in different real-world environments, given different goals people might have, and taking into account cognitive constraints. It's easy to naively say that something is a 'bias' or 'irrational', and infer from this that there is a problem that needs to be solved - but there are a lot of missing steps and assumptions here. If we want to improve human reasoning, we need to understand the nuances of why people use the strategies they do, what benefits those strategies might have, and whether it's practically feasible for them to do anything better. This is not to say that the way people reason given their prior beliefs is not sometimes problematic, and that there isn't room for improvement - but simply that we need to tread carefully.

To begin with, it might help to break down the broad question, "do people exhibit a confirmation bias?" into smaller constituent parts, recognising the issues arising at each stage:

First, we can ask whether there's evidence of a bias in the strict sense defined: do people's reasoning and judgement strategies deviate from normative standards, and do so in a way that supports their current hypothesis over alternatives?

Though it might seem on a naive view like the strategies people use are 'biased' in such a way towards what they already believe, there's certainly not conclusive evidence that a confirmation bias when defined in this more specific sense. There are a number of difficult issues that arise in answering this question:

- What is the correct normative standard, and how is it justified?
- How can we study human reasoning in ways that allows us to compare performance to clear normative standards, but also has relevance to real-world situations?
- What kinds of reasoning strategies are people using in the first place, that would result in bias?
- When determining whether a deviation is systematic, how broad should the scope be? It's clearly not enough to show that a deviation exists in one very specific type of task - but in how many different domains and situations does a bias have to be established in order to count as a bias more broadly?

Second, we might ask whether, even if a bias exists, it actually leads to problematic consequences in real-world scenarios.

This is closely related to the question of what the appropriate normative model is - even if people's judgements seem to deviate systematically from a simple normative model, it's possible that when we factor in various other assumptions the picture looks quite different. This raises additional questions:

- Might people's reasoning strategies be optimised for different goals than we've assumed, and therefore be considered instrumentally rational with respect to these goals?
- Can we say that some goals are normatively better than others? What should we do when goals conflict at different levels - e.g. the goals of the 'genes' vs those of the 'organism', goals on different time horizons, the goals of individuals vs goals of society?
- Might people's reasoning strategies be the best possible ones within the bounds of cognitive and biological capacity?
- Might such strategies be well-adapted to specific real-world scenarios even if seemingly ill-suited to abstract lab tasks?

It's worth acknowledging here that it's also possible that people's reasoning strategies are not biased in the first, strict sense - but that the goals and incentives of certain environments mean that these basic strategies result in problems or errors in specific situations. This is the reverse of a point that's often been made in psychology research - that we see errors in abstract lab situations, but that 'in the real world' people actually reason pretty effectively. It's possible that people are not biased towards confirming whatever they already believe in an abstract, strict sense - but that for certain kinds of beliefs we hold, there are many more incentives for us to continue believing whatever we currently do than there are to be accurate - leading to a kind of domain-specific confirmation bias. One domain in which we might expect this to be the case is political beliefs or other areas where beliefs seem to be strongly tied to identity and social groups - a point I will return to later.

Having now discussed in detail the evidence for confirmation bias, and the complex normative issues arising around it, I'm now going to look at these issues from a slightly different angle. The basic idea behind confirmation bias is that we let what we already believe influence our reasoning too much - creating a dangerous circularity where we end up reinforcing our prior beliefs. The reason that the normative issues quickly get very complex here is that it's not clear how much is 'too much': if we believe anything at all, those beliefs have to influence us in some ways. The most extreme way to avoid confirmation bias would be to simply have no beliefs at all, to be highly uncertain of everything, or to find some way to entirely 'set aside' what we already believe when considering new evidence. We might refer to this as 'open-mindedness': being able to consider all possibilities, set aside one's preconceptions, and view issues from a blank slate-perspective. Just as it's often assumed that we fall prey to aconfirmation bias, I think it's often assumed that it would be better if people were more 'open-minded'. But is this really the answer? What does it really mean to be open-minded, and is it always a good thing? In the next chapter I will focus on these questions - on what it means to be open-minded and whether it's necessarily good - in the context of the rest of what I've discussed in this thesis.

Chapter 5

Open-mindedness

5.1 Introduction

In this thesis so far, I've focused on the notion of confirmation bias - the idea that what we already believe can 'bias' our thinking processes in various ways: influencing what information we seek out, how we interpret it, and how we form beliefs. I've talked about how the evidence for this 'bias' is not as strong as it might first seem: that the definition of confirmation bias is itself confused, and that there are unresolved questions regarding how much one's prior beliefs *should* influence subsequent reasoning. We cannot approach every question or situation as a blank slate, and to some extent it is rational for prior assumptions to guide how we make sense of the world. It is not clear exactly how we separate out what is bias from what is not; an issue that is further complicated by different meanings of the word 'bias'.

I have argued that the existing literature on confirmation bias does not do nearly enough to address these issues, and so has been too quick to attribute bias in cases where it may not apply. At this point, I'd like to turn to a closely related concept which has been explored in both the psychological and philosophical literatures - that of 'open-mindedness'. Just as it is generally assumed that people fall prey to a confirmation bias, it has also been commonly stated that people should be 'more open-minded': better able to step back from what they already believe, set aside assumptions, and consider different perspectives.

Open-mindedness is considered important both in academic discussion and in more popular discourse. Philosopher Wayne Riggs points out that "open-mindedness is typically at the top of any list of the intellectual or 'epistemic' virtues." (Riggs, 2010, p.172) In psychology, open-mindedness is considered a character strength (Peterson and Seligman, 2004) and interventions to promote open-mindedness have been considered a kind of 'positive psychology' intervention (Seligman et al., 2005). In educational theory, the idea that we should be teaching young people to be open-minded has received a great deal of attention (Harding and Hare, 2000, Hare, 1993, Miri et al., 2007). A number of more popular books and articles have been written on why and how to be more open-minded: including "How to be critically open-minded" (Lambie, 2014), "Teaching tolerance: raising open-minded, empathetic children" (Bullard, 1996), and articles with titles like, "Open your mind to let happiness in" (Lian, 2017). We hear talk about the importance of making people and society more open-minded, of teaching children to be open-minded, and of the problems this would solve.

But just as with confirmation bias, I think this concept of open-mindedness needs some closer evaluation. The belief that we need to promote open-mindedness seems very closely related to, even dependent on, the idea that people fall prey to a confirmation bias. Open-mindedness may sometimes be thought of as a kind of 'antidote' to confirmation bias: the reason we need to be more open-minded is that we tend to be biased towards what we already believe. But if as I have suggested here, the case for confirmation bias is not as straightforward as it seems, then perhaps this should give us grounds to step back and question the widespread assumption that 'people should be more open-minded.'

This chapter has two main aims. First, to bring together two research literatures that deal with very closely related questions but have rarely been explicitly linked - the literature on confirmation bias, and on open-mindedness. Both are essentially concerned with ways in which our prior beliefs and assumptions might constrain our thinking, from different perspectives - and so bringing them together may help shed new light on this issue. Second, to more closely examine the concept of open-mindedness - a notion frequently invoked yet rarely questioned - within the context of the issues discussed in the rest of this thesis. What exactly is open-mindedness, how does it relate to confirmation bias, and why do we think it's so important? Might open-mindedness sometimes be a bad thing - is it possible to be *too* open-minded? What might everything I've discussed

in this thesis imply for open-mindedness, both as it is studied in psychology and more broadly?

The chapter will be structured as follows. First, I will review how the term 'open-mindedness' has been understood in different contexts, to get a clearer picture of what it really means and how it relates to confirmation bias. I will then consider whether discussion of open-mindedness does enough to consider the potential downsides of open-mindedness, both in psychology and in more popular discourse. I will argue that the claim that people 'should be more open-minded' suffers from some similar challenges as claims of confirmation bias do, resulting in overly simplistic normative claims that cannot be backed up. Finally, I will talk about the implications of these issues for how we think about and study open-mindedness.

5.2 What is open-mindedness?

To get a clearer picture of what open-mindedness is, I will start by attempting to capture some intuitions about open-mindedness as it occurs in everyday usage - before reviewing how the term has been treated in both the psychological and philosophical literature.

5.2.1 An intuitive picture of open-mindedness

When we think of someone who is open-minded, we tend to think of someone who actively tries to engage with a whole range of views - perhaps reading different newspapers from across the political spectrum, and wanting to understand all perspectives on an issue, even those (perhaps especially those) they find difficult to agree with. An open-minded person does not 'switch off' when someone challenges them, but seriously tries to consider what they might learn from the challenge. An open-minded person is willing to admit they were wrong and change their mind when the evidence stacks up against them. Earlier I suggested that confirmation bias needs to be understood in terms of every stage of the reasoning process: not just how we seek out new information, but also how we interpret and update our beliefs based on new information (as illustrated in diagram 2.1 earlier in the thesis.) We might say something similar here about open-mindedness: to be open-minded it is not enough to simply seek out different perspectives, for example. It also matters how one pays attention to, and updates one's beliefs as a result of these

different perspectives. This also ties in with some of our discussion in the earlier chapter on selective exposure. I suggested that selective exposure has sometimes been assumed to roughly measure how 'open-minded' a person is, but that we can only learn a limited amount without understanding how people interpret and use the information they seek out.

A closed-minded person is commonly thought of as quite the opposite: fixed in their views; uninterested in trying to understand why others might disagree with them or what they might learn from other perspectives; dismissive and even outright hostile towards anyone who might challenge them. A closed-minded person is very unlikely to change their views, except in light of overwhelming evidence. We also tend to associate closed-mindedness with a lack of tolerance, and perhaps prejudice: a tendency to assume that those with differing views are stupid, evil, or crazy.

We might worry that a person is *too* open-minded if open-mindedness starts to look like gullibility, or an unwillingness to take a stance on any issue. As I said at the beginning of this thesis, having a completely 'blank slate' - having no beliefs or assumptions - is neither possible nor desirable, even though in a sense such a person might be considered totally open-minded. What seems best, then, is some balance of open-mindedness with the ability to critically scrutinise and evaluate different positions, to hold firm beliefs where doing so is useful and appropriate. Doing this and not falling into the trap of getting too attached to those beliefs - and therefore becoming too closed-minded - seems challenging. Attaining the perfect amount of open-mindedness seems like a very delicate balance, and leaning too far in either direction can have its costs (an idea I will return to later in the philosophical discussion.)

Open-mindedness has a clear link to the idea of confirmation bias: if we think that people in general fall prey to a confirmation bias, then we might think of more open-minded people as being those who are less vulnerable to the bias, or have taken more steps to avoid it. And just as it seems possible to be too open-minded, as we have seen in the earlier discussion, some amount of confirmatory reasoning might sometimes be helpful (or at least is sometimes necessary.)

5.2.2 Open-mindedness in psychology

5.2.2.1 Early accounts of open-mindedness

Discussion of open-mindedness in the psychological literature has attempted to characterise and measure open-mindedness as a broad personality trait, in contrast with closed-mindedness. Some of the earliest discussions of these concepts focus more on concerns about people being too closed-minded, and measuring this. Adorno et al. (1950), for example, created a set of criteria intended to capture the 'authoritarian personality type', initially motivated by wanting to understand the conditions that allowed for something like Nazi-ism to gain foothold. Adorno's scale included elements such as 'blind allegiance to conventional beliefs', 'belief in aggression towards those who disagree', and 'black-and-white thinking'. Rokeach (1960) similarly developed a 'dogmatism scale' intended to capture how open- or closed-minded a person was, as well as tendencies towards authoritarianism and intolerance.

Both Adorno et al. and Rokeach's scales were criticised, however, for failing to measure closed-mindedness independently of political ideology - both their measures were correlated with right-wing views. The implication here is that open-/closed-mindedness should capture a certain way of holding one's beliefs, independent of whatever those beliefs are. Both scales were also criticised for trying to measure too many different things, and failing to clearly distinguish between closely related notions - closed-mindedness does not seem to be exactly the same thing as authoritarianism or intolerance, though closely related.

5.2.2.2 Openness as a personality dimension in the five-factor model

Perhaps the best-known account of open-mindedness in more recent psychological literature is the construct of 'openness to experience' in the Big Five personality inventory (also known as the five-factor model - see Digman, 1990, for a review and discussion of the theory). Openness is considered a very broad personality trait, encapsulating a variety of different kinds of openness: openness to ideas, fantasy, aesthetics, feelings, actions, and values. Though when we think of 'open-mindedness', we tend to focus mostly on ideas and beliefs, personality research has found these different types of openness are highly correlated (hence including them in just one dimension.) However, there

does seem to be disagreement about how exactly to conceptualise this dimension (with some suggesting it is better referred to as intellect or culture - see McCrae and Costa (1997).

McCrae and Costa suggest one important and useful way to characterize openness is in terms of "the structure of consciousness." (McCrae and Costa, 1997, p.838) They suggest that people may fundamentally differ in the extent to which they are able to hold in mind or access multiple thoughts or feelings at once: with more 'open' individuals finding this much easier than more 'closed' individuals. We might expect that individuals who are more 'open' in this sense to find it easier to consider alternative interpretations of new information not necessarily because they are less biased by their prior assumptions, but simply because they find it cognitively easier to consider multiple perspectives at once than a more 'closed' person.

This perspective suggests that open-mindedness is rooted in basic cognitive abilities - but McCrae and Costa also suggest that *motivation* is also crucial for understanding openness. Open-minded people don't seem to be simply unable to screen out ideas and experiences - they seem to actively seek out new and varied experiences. McCrae and Costa therefore suggest that the 'openness' personality dimension is characterised both by a particular "permeable structure of consciousness" (cognitive ability) and "an active motivation to seek out the unfamiliar" (motivation). (McCrae and Costa, 1997, p.839)

It's also worth noting here a number of individual difference variables that seem closely related to the concept of openness (table 5.1). There seems to be a fair amount of overlap across the different terms used here. In particular, the Need for Closure (NFC) variable seems very broad and many of its composite parts close to existing notions. Neuberg and Newsom (1993) argue, using their own data and factor analysis, that the NFC scale seems to fail as a unidimensional construct. They suggest that the NFC scale in fact masks (at least) two largely independent motives: (i) the preference for quick, decisive answers to questions, and (ii) the need to create and maintain simple structures (specific closure.) However, even accounting for this, Neuberg and Newsom (1993) argue that the scale fails to exhibit discriminant validity relative to other measures that already exist - particularly the personal need for structure (PNS) scale.

Bringing these individual differences together with McCrae and Costa's characterization of openness, there does seem to be a common thread: the idea that individuals differ

in their willingness and ability to hold multiple ideas and experiences in mind at once, meaning they respond differently to ambiguity/uncertainty, have differing levels of motivation to seek out new and varied experiences, and differing needs to structure and organise ideas simply. If this dimension does, as McCrae and Costa suggest, have both a cognitive/structural and a motivational component, there is an interesting question of how these are related. Structural differences may influence motivation - people who find it cognitively easier to hold multiple ideas in mind at once may also find it easier to experience the rewards of deeper understanding, resulting in greater motivation to seek out varied ideas. Or do differences in motivation come first, resulting in different ways of structuring beliefs that match those incentives? Possibly there is some effect in both directions.

Construct	**Summary**
Intolerance of ambiguity (IA) (Frenkel-Brunswik, 1949)	Tendency to interpret ambiguous situations as threatening, and to respond to novel, complex situations with discomfort and avoidance.
Intolerance of uncertainty (IU) (Dugas et al., 1997)	Tendency to consider it unacceptable that a negative event may occur, no matter how unlikely. Closely related to IA, but differs in that it focuses on discomfort with uncertainty about the future, whereas IA focuses on discomfort with being unable to interpret current situations. IU is more closely related to anxiety/worry than IA.
Uncertainty orientation (Sorrentino and Short, 1986)	Uncertainty-oriented people view uncertainty as a challenge, and enjoy approaching and resolving uncertainty. Certainty-oriented people view uncertainty as something to be avoided, and cling to the familiar, predictable and certain.
Personal need for structure (PNS) (Neuberg and Newsom, 1993)	The extent to which an individual is inclined to cognitively structure their world in simple, unambiguous ways. 'Cognitive structuring' refers to the creation and use of abstract mental representations that are simple generalisations of previous experiences, and is one way people might reduce their cognitive load (along with avoidance strategies that limit the information they are exposed to.)
Need for cognitive closure (NFC) (Webster and Kruglanski, 1994)	The desire for an answer on a given topic as opposed to confusion and ambiguity. The researchers who proposed NFC as a psychological construct suggest that it be treated as a single variable that is manifested through several different aspects: desire for predictability, preference for order and structure, discomfort with ambiguity, decisiveness, and closed-mindedness.
Need for cognition (Cacioppo and Petty, 1982)	The extent to which an individual needs to structure relevant situations in meaningful, integrated ways - a need to understand and make reasonable the experiential world. This has also sometimes been referred to/thought of in terms of differences in individuals to engage in and enjoy cognitively demanding tasks and thinking.

Table 5.1: Individual difference variables related to openness

5.2.2.3 Open-mindedness as behaviour

The psychological research on open-mindedness I have discussed so far tends to characterise open-mindedness as a broad personality trait, and to measure open-mindedness using self-report surveys. However, we might also think of open-mindedness in a narrower sense: as a certain way of reasoning about a given idea or topic. In this sense, open-mindedness is essentially a kind of behaviour, and the extent to which one person is open-minded might vary depending on the situation or issue. This isn't necessarily incompatible with the personality-trait view of open-mindedness: we could say that people can differ in how open-minded they are generally, but that a given person might also vary in how open-minded they are in different situations (in the same way we might attribute honesty as a broad trait and also to more specific situations.)

It also seems useful to develop ways of measuring open-mindedness behaviourally, as opposed to using self-report scales, as these measures may be more objective. Since open-mindedness is generally thought of as a desirable trait, self-report measures are vulnerable to social desirability biases: people answering questions based on how they would *like* to come across, rather than based on what is actually true of them (Paulhus, 1991). People may exaggerate their responses to questions out of a desire to *appear* open-minded.[1] Ideally we would like to measure open-mindedness by actually observing people's behaviour and seeing whether they do what they claim to do.

However, research on open-mindedness-as-behaviour and ways to measure it is limited. Baron (1985, 2000) discusses the importance of 'actively open-minded thinking' (hereafter AOT), which seems to characterize open-mindedness less as a broad personality trait and more as a way of thinking about an issue. Baron defines AOT as thinking where "search is sufficiently thorough for the question, search and inference are fair to possibilities under consideration, and confidence is appropriate to the amount of search that has been done and the quality of inferences made." (Baron, 1996) This seems to be a very broad definition - almost describing something closer to 'good thinking' more generally.

[1] Recall that in our analysis of people's responses to the question, "Why did you choose to read this balance of arguments?", we found that over 50% of people mentioned being fair/unbiased - whether this indicates that people actually *are* fair and unbiased, or just that they are aware of the importance of *appearing* unbiased, is not entirely clear (and a sceptical interpretation could easily draw the latter conclusion.)

The notion of AOT, and methods used to measure it, seems closely related to confirmation bias - Gurcay-Morris notes that "most behavioural measures of AOT are designed to assess myside bias." (Gurcay-Morris, 2016, p.9) (As we saw in chapter two, myside bias is one of many phenomena in the psychological literature that has come under the heading of 'confirmation bias'.) For example, Baron (1995) assessed myside bias in thinking about abortion, by asking students to generate lists of arguments on either side of the issue, and to evaluate the arguments produced by others. Stanovich and West (1997) similarly develop an 'argument evaluation test' where subjects are asked to evaluate fictitious individuals arguments, and their responses are then compared to the evaluations made by expert judges, to see if they give more favourable evaluations to arguments that support their prior beliefs. However, as discussed in an earlier chapter, one issue with these measures is that it is difficult to compare people's responses to clear normative standards. We can draw descriptive conclusions about how people generate and evaluate arguments, and we could define this 'open-mindedness' - but if we want to say anything about how people should reason, and whether people should be more open-minded, we run into the same issues discussed in the earlier chapter.

5.2.2.4 Open-mindedness in psychology: summary

In much of the psychology literature, open-mindedness is characterized as a broad personality trait. It is often contrasted with concepts like authoritarianism and dogmatism, indicating that a key part of what it is to be open-minded is to not take things as given, to question things, and consider the possibility of being wrong. More recently, personality psychology has suggested that open-mindedness as it relates to beliefs and ideas might be part of a broader trait of openness: the willingness and ability to hold multiple feelings/experiences/ideas in mind at once, and the extent to which one actively seeks out depth and variety of experience. The concept of open-mindedness also seems closely related to the ability to tolerate and even enjoy uncertainty, whereas closed-mindedness is related to an aversion to ambiguity and uncertainty, and the need to structure, control and understand.

Open-mindedness in psychology has been loosely linked to confirmation bias, insofar as open-mindedness has sometimes been characterised as the absence of certain biases. In particular, Baron's notion of 'actively open-minded thinking' has often been measured

as the absence of myside bias - the ability to generate and evaluate arguments from multiple perspectives, not constrained by one's prior beliefs.

What's perhaps surprising is that nowhere in the psychological literature is there a precise, agreed-upon definition of open-mindedness. It's generally assumed that we know and agree what open-mindedness means. And yet if we want to be able to measure it, and discuss important questions about whether it should be promoted, we need more than a vaguely agreed upon concept - a problem which arises in the philosophical literature, which we turn to next.

5.2.3 Open-mindedness in philosophy

Much of the philosophical literature on open-mindedness is in the philosophy of education - discussing whether, and how, promoting open-mindedness should be a goal of education. This generates a great deal of debate around whether it is possible to be too open-minded, and the possible downsides of promoting open-mindedness. I will turn to these issues in the next section - for now, I will simply review how the concept of open-mindedness has been characterized in the philosophical literature. Of course, as we will see, these two questions - of what it means to be open-minded, and to what extent it is a good thing - are closely related.

5.2.3.1 Open-mindedness as uncertainty

One of the most vocal proponents of open-mindedness and its importance in education is William Hare (Hare, 1985, 1993, 2003, 2006, Hare and McLaughlin, 1994, 1998). Though he has written extensively on the topic, Hare's definition of what it means to be open-minded varies and is somewhat vague. In some places, Hare simply defines open-mindedness in terms of the ability to change one's mind - "an open-minded person is one who is able and willing to form an opinion, and revise it, in the light of evidence and argument." (Hare, 1985, p.251) Elsewhere, he speaks of open-mindedness more as those specific attitudes and habits of thought which enable one to change one's mind: the readiness to give due consideration to relevant evidence and argument; being critically receptive to new perspectives and alternative ideas; remaining committed to reconsidering views in light of new questions, doubts, and findings (Hare, 2006). The key thread

here is the idea of being able to both: (a) critically evaluate one's own position, and (b) give due consideration to other positions, in order to be able to revise one's position when necessary.

Hare's account of open-mindedness is challenged, however, by Gardner (1993, 1996) - who believes that uncertainty is more central to open-mindedness than Hare suggests (and that this, in turn, causes problems for the basic view of open-mindedness as a good thing.) Gardner states that "to be open-minded is to have entertained thoughts about an issue but not to be committed to or hold a particular view about it." (Gardner, 1993, p.39) A disagreement between Hare and Gardner ensues, concerning whether or not it is possible to be open-minded about a belief to which one is firmly committed. This disagreement largely seems to stem from different notions of what it means to be open-minded. Hare responds to Gardner's criticism by arguing that open-mindedness is a certain attitude towards one's beliefs that is independent of how certain one is - but does not make it clear exactly what it means to have such an attitude, and how it is compatible with firm belief.

Much of the subsequent discussion of open-mindedness in philosophy revolves around attempts to dispel this apparent tension between open-mindedness and firm belief. Is it possible to characterise open-mindedness in such a way that it iss separate from uncertainty - so that it is possible to be simultaneously very certain about something, but still open-minded about it?

5.2.3.2 Open-mindedness as intellectual humility

Adler (2004) is one of the first to pick up on this unresolved tension between Hare and Gardner, and to propose a solution. Adler suggests we think of open-mindedness not as an attitude one holds towards any specific belief, but rather an attitude towards oneself as a believer more generally.

He unpacks the tension between open-mindedness and belief as depending on the following assumptions:

1. That if one is strongly committed to a position, one must regard it as not seriously possible that it is wrong;

2. That being open-minded about a position requires one to consider it seriously possible that one is wrong. (Adler, 2004, p.129)

Adler suggests that we should reject the second assumption, arguing that it is possible to be open-minded about a position without considering it seriously possible that it is wrong. He starts by saying that the reason we care about open-mindedness is that we want to have true beliefs, but recognise that we are fallible as thinkers. Given this fallibility, we must acknowledge that some of our beliefs are incorrect - and open-mindedness allows us to discover which ones they are. "Open-mindedness is then a second-order attitude towards one's beliefs as believed, and not just towards the specific propositions believed... fallibilism is a second-order doubt about the perfection of one's believing, not a doubt about the truth of any specific belief." (Adler, 2004, p.130)

On Adler's account, then, to be open-minded is to recognise one's fallibility as a believer. This is consistent with not having any reason to think any specific belief is false, and therefore consistent with holding many beliefs in a 'full', 'committed', or 'strong' sense. Riggs (2010) agrees with Adler's account, but suggests that it needs to be extended: to be genuinely open-minded, this recognition of one's fallibility needs to be supplemented by certain habits of thought that lead to genuine open-minded inquiry. More specifically, Riggs suggests that being open-minded requires one to (a) develop *self-knowledge*: awareness of one's biases and the circumstances in which one is most likely to be led astray; (b) *self-monitor* in order to identify when one is in such a scenario where open-mindedness is needed. If Adler is saying that open-mindedness means recognising the ways in which one is biased, Riggs' point is that recognising biases in a broad sense is not necessarily enough to actually counteract them - since it can still be very difficult to identify *when* one is biased in the moment. Riggs' account therefore says that open-mindedness requires recognising one's fallibility, and also developing a more specific understanding of (a) what one's biases are, and (b) when these biases are particularly likely to arise. However, this feels like it is beginning to slip into the trap discussed before where 'open-mindedness' gets equated 'good reasoning' more broadly.

One question we might ask here is whether Riggs' (or Adler's) account actually captures how we think of open-mindedness intuitively. Awareness of one's fallibility in general, the specific biases one is vulnerable to, and when they are likely to come up, will certainly

help a person to identify situations where it might be particularly important to be open-minded. But when we say someone is being open-minded, I think we mean something more than that they are aware of their biases - we mean that they are specifically good at seeking out alternative viewpoints, and taking counter-arguments seriously. The attitudes and skills Adler and Riggs describe certainly seem like they *help* a person to be open-minded when it is most important, but it seems a little confused to claim that what they are describing is itself open-mindedness.

Spiegel (2012) makes this point, arguing that Adler does not really resolve the conflict between open-mindedness and firm belief. While recognising one's fallibility clearly explains how one can generally be open-minded while holding firm beliefs, it cannot account for belief-specific accounts of open-mindedness. Spiegel argues further that in fact, what Adler is describing should not be termed 'open-mindedness' at all, but is rather a separate, related, intellectual virtue, which he calls 'intellectual humility'. However, even if Spiegel is correct about this, he does not provide a better alternative. He suggests returning to Hare's original conception of open-mindedness, but does not offer any resolution to the initial problem of a tension between open-mindedness and firm belief.

5.2.3.3 Open-mindedness as detachment or engagement

Finally, I will briefly discuss two more recent characterisations of open-mindedness in the philosophical literature that are closely related: Baehr's account of open-mindedness as 'detachment' (Baehr, 2011), and Kwong's account of open-mindedness as 'engagement' (Kwong, 2016).

Baehr (2011) begins by listing a number of different situations in which a person might be said to be being open-minded: fairly evaluating arguments that run counter to one's views; being impartial when assessing two sides of an argument; setting aside any pre-conceptions in order to understand a new idea. Baehr criticises prior accounts of open-mindedness in that they focus too much on the first case: cases of open-mindedness that arise as a result of a direct conflict between one's beliefs and new evidence, and fail to acknowledge that open-mindedness can also be attributed in cases without such a conflict.

Looking at these different examples of open-mindedness, Baehr argues that the 'conceptual core' of open-mindedness is that in each case, the person *detaches* from a certain default or privileged standpoint. In addition, Baehr suggests, open-mindedness requires that one do so with a certain motivation: to be detaching from a default perspective with the aim of taking an alternative or new perspective seriously. Open-mindedness therefore requires both a certain *motivation* - to seriously consider and understand different perspectives - and the *ability* to do so independent of what one currently believes (echoing McCrae and Costa (1997) earlier.) Baehr offers the following specific definition of open-mindedness: "an open-minded person is characteristically willing and (within limits) able to transcend a default cognitive standpoint in order to take up or take seriously the merits of a distinct cognitive standpoint." (Baehr, 2011, p.202)

Kwong (2016) agrees with much of Baehr's analysis, but is concerned that his account does not allow one to be open-minded about a strongly held belief - returning us to Gardner's original objection. If I am fully committed to a belief, is it really possible to 'detach' from it in the way Baehr believes is necessary for open-mindedness?

Kwong suggests that the conflict between open-mindedness and belief might be more easily resolved if we construe open-mindedness in terms of *engagement*: "a willingness to make room for novel ideas in one's cognitive space and give them serious consideration." (Kwong, 2016, p.71) According to Kwong, 'engagement' is a broad term constituting a wide range of cognitive activities, including but not limited to the notion of transcendence or detachment. Detaching from one's current perspective is one way in which one can engage with a different perspective, and is sometimes but not always necessary. What is key to engagement is 'making room for' a viewpoint in one's cognitive space, not necessarily to consider why it might be true but at least to see how it might relate to, or connect with, one's existing network of beliefs. By contrast, to be closed-minded is to completely dismiss a viewpoint without even seriously attempting to understand it or how it relates to what one already believes. Kwong (2016) suggests that it's possible to engage with a belief while leaving one's 'epistemic commitment' intact - for example, by trying to demonstrate that it is false. However, whether this kind of engagement really constitutes open-mindedness is controversial.

Perhaps what this highlights is that open-mindedness comes in degrees: we can be more or less open-minded. To try to demonstrate, through reasoned argument, why p is

false, is certainly more open-minded than simply dismissing p as false without proper consideration. Someone who does this repeatedly is, in the long run, more likely to end up changing their beliefs than someone who does not (sometimes they may realise that demonstrating p's falsity was not as easy as they thought.) However, it is not clear that we would call someone who does this a particularly 'open-minded' person, if they never seriously try to consider why any opposing view might be true.

5.2.3.4 Open-mindedness in philosophy: summary

The philosophical discussion of open-mindedness has done more to try to pin down what precisely the concept means than has been done in psychology. Attempting to do so inevitably raises more issues and disagreements: while psychologists are broadly agreed on what open-mindedness is and why it's good, philosophers spend much more time discussing these questions. In particular, philosophers are concerned that being open-minded is by nature incompatible with being firmly committed to beliefs. This begins to hint at potential downsides to being open-minded which the psychology literature largely fails to acknowledge, and which I will discuss in more detail in the next section.

Several philosophers have attempted to characterize open-mindedness in a way that avoids these difficulties, but none entirely convincingly. Adler (2004) suggests we think of open-mindedness as higher-order recognition of one's fallibility, not necessarily requiring one doubt any specific beliefs. However, it seems like recognising one's fallibility should simply give one reason to have some minimal degree of uncertainty in all one's beliefs - and Adler may simply be talking about a related, but different intellectual virtue ('intellectual humility'). Baehr (2011) and Kwong (2016) suggest we might characterize open-mindedness in terms of certain ways of detaching from existing beliefs or engaging with new information, but it's still unclear whether these things are actually possible for firmly committed beliefs.

Is this whole attempt to resolve the open-mindedness/committed belief tension really necessary? I'll suggest that it's only necessary if we think that open-mindedness should be a *normative concept*: that open-mindedness is always a good thing, and that more open-mindedness is always better. However, I don't think this is actually how we think of open-mindedness. Open-mindedness describes ways of thinking that may well improve reasoning in various ways, but does not describe *ideal* thinking (at best, open-mindedness

is a *prescriptive* notion.) I'll discuss this distinction in more detail in the next section, suggesting that avoiding acknowledging the potential downsides of open-mindedness, has resulted in a confused view of whether we should be open-minded, and why.

5.2.4 Summary: what is open-mindedness?

The different accounts of open-mindedness that I've discussed share a lot of common themes. It is generally agreed upon that open-mindedness involves a willingness and ability to consider different perspectives and their merits, and to not immediately disregard everything one disagrees with. What seems central to open-mindedness is that it is concerned with ensuring we don't get too 'stuck' in any one perspective, and helping us discover if our beliefs are wrong.

Open-mindedness can be considered as both a broad *personality trait*: capturing a fairly fundamental way that people differ, and as a *behaviour* more narrowly defined: such that a person might be said to be more open-minded about a specific topic or in a specific situation than others. Being open-minded seems to require both a certain *ability*: having the cognitive capacity to consider multiple ideas or switch between alternatives, and a certain *motivation*: actively wanting to seek out different perspectives.

Discussions of open-mindedness seem obviously connected to confirmation bias, but this link is rarely explicitly made. Open-mindedness might simply be thought of as the opposite of confirmation bias: reasoning without being unduly influenced by prior beliefs - or perhaps as a way of 'correcting for' confirmation bias: ways of thinking that can be developed to counteract such a bias. Either way, the idea that open-mindedness is important and should be promoted seems to be rooted at least implicitly in the assumption that we fall prey to a confirmation bias.

Despite a fair amount of attention in both the psychological and philosophical literature, what precisely it means to be open-minded is still vague, and some disagreements remain. In particular, it's not clear whether it's possible to be open-minded about something one firmly believes, or whether being open-minded is the same as being undecided or uncertain. Relatedly, there is disagreement about whether open-mindedness is always a good thing, and whether more open-mindedness is always better. There is also limited

research on how we might measure open-mindedness behaviourally, and how exactly to identify open-minded reasoning.

Given how often open-mindedness is said to be important, therefore, it seems there are aspects of open-mindedness that deserve more attention in the research. What precisely does it mean to be open-minded about a topic, and how is this different from simply being uncertain, or other related notions such as tolerance? Is it always a good thing to be open-minded, or might 'too much' open-mindedness sometimes be a problem? It is this latter question to which I will now turn - suggesting that, in a similar way to how the literature on confirmation bias has not done enough to demonstrate genuine bias, the literature on open-mindedness has not done enough to justify the common view of open-mindedness as an unqualified virtue.

5.3 Should we be more open-minded?

The benefits of open-mindedness are fairly obvious, and frequently discussed: being open-minded prevents us from getting too stuck in one perspective, allowing us to change our minds as we learn more information, helping us to avoid false or unhelpful beliefs. There are also a number of more social benefits to open-mindedness: it helps us to get along with people we disagree with and avoid conflict.[2]

The potential costs of open-mindedness are less clear, and less commonly discussed. Might it sometimes be useful to 'close' one's mind on an issue - to decide something is not worth further consideration? It's often said that people should be more open-minded, but what exactly does this mean? In this section I'll argue that, like with confirmation bias, this normative claim that we should be more open-minded is much more complex and hard to defend than it might first seem.

I'll start by reviewing some potential downsides of open-mindedness: ways in which it might be possible to be too open-minded. I'll then consider two different ways we might interpret the claim that people should be more open-minded. First, we might interpret

[2] As a side note, it's worth being careful here not to confuse open-mindedness with the closely related notion of tolerance. It seems possible to be tolerant of different views: that is, to not think badly of those who have them - while also failing to be genuinely open-minded about those views: not considering it possible they are actually true. Open-mindedness certainly makes it easier for us to be tolerant, by helping us to see the potential benefits in others' perspectives - but being tolerant does not necessarily make us open-minded.

it as a *normative* claim - saying that open-mindedness is a normative concept, an ideal to be attained, and that the more open-minded we can be, the better. Alternatively, we might interpret it as a *prescriptive* claim - acknowledging that more open-mindedness is not *necessarily* always better, but given the kinds of errors that people currently tend to make, pushing more in the direction of open-mindedness would improve reasoning (in relation to some other normative standard.) I will argue, drawing on earlier discussion in this thesis, that neither of these claims that people 'should' be more open-minded are easily defended.

5.3.1 The costs of open-mindedness

Though the psychology literature almost exclusively talks about the benefits of open-mindedness, there is some limited discussion of the potential costs of being too open-minded (e.g. Kruglanski, 2013). Kruglanski emphasises the difficulty of the fact we have to make decisions and take actions, and yet our judgement and decision-making processes have no 'natural' termination point. Given this, Kruglanski argues, the ability to sometimes "shut our minds" is important - allowing us to focus on just one belief or viewpoint and "get on with our lives." This seems to be an argument about conserving attentional resources: given we have limited attention, we obviously can't be maximally open-minded about everything. In this basic sense, it seems like of course it's possible to be too open-minded.

We also touched earlier upon some philosophical discussion of the potential downsides of open-mindedness: particularly the work of Gardner (1993, 1996). Gardner is concerned that there may be certain issues that it is not appropriate to be open-minded about, especially moral issues. He worries that encouraging people to be open-minded might lead people to give more weight to certain views than they deserve - leading people to consider 'crackpot' views or conspiracy theories, or to question important moral principles, for example. Sometimes, Gardner thinks, it might be better to close one's mind on an issue than to risk falling into a relativist position where all views are equally worthy of consideration. The concern here seems to be particularly rooted in the idea that open-mindedness alone may not always be a good thing, if it's not accompanied by other virtues, such as the ability to critically evaluate arguments and be appropriately discriminating.

Both Kruglanski (2013) and Gardner (1996) make good points about the potential downsides of open-mindedness: open-mindedness can go too far if the attentional resources it consumes outweigh the benefits it brings, and open-mindedness alone can lead to problems if not accompanied by other skills such as critical evaluation. There is also some psychology research suggesting that people who score high on the personality trait of 'openness' are more likely to suffer from certain psychological problems, such as chronic nightmares and symptoms associated with schizophrenia spectrum disorders, including dissociation and perceptual aberration (disturbance in one's continuous experience of space and time.) (McCrae and Costa, 1997). McCrae and Costa acknowledge that having highly 'permeable' cognitive systems is not always adaptive, and suggest that individuals very high on the trait of openness "may be so easily drawn to each new idea or belief that they are unable to form a coherent and integrated life structure." (McCrae and Costa, 1997, p.841)

Given that open-mindedness does seem to have downsides not commonly recognised, we might start to question the simple idea that it's always better to be more open-minded.

5.3.2 Why open-mindedness is not a normative concept

In light of these potential downsides of open-mindedness, several philosophers have attempted to develop an account of open-mindedness that dodges these criticisms: to characterise open-mindedness in a way that means it's not possible to be too open-minded, in a way that avoids these downsides. For example, Adler (2004) suggests open-mindedness is the ability to recognise one's fallibility, that is always useful regardless of how certain one may be of any specific belief - but I don't think this adequately explains why this higher-level uncertainty wouldn't just 'trickle down' to make one less certain of any specific belief. Others like Baehr (2011) and Kwong (2016) suggest that to be open-minded is more about being able to engage with new information in a certain way, unconstrained by prior beliefs - but I do not think they go far enough in explaining how it is really possible to, for example, 'detach from' something one very strongly believes. Many accounts of open-mindedness that attempt to dodge these criticisms end up characterising open-mindedness in such a broad way that they're essentially equating it with 'good thinking' more generally. For example, Hare (2003) - who has tried very hard to rebut all criticisms against open-mindedness - says that part of being

open-minded is "to be concerned to defuse any factors that constrain one's thinking in predetermined ways" (Hare, 2003, p.5), which sounds suspiciously close to simply 'being unbiased.' And in responding to Gardner's concerns, Hare and McLaughlin clarify that "our conception of open-mindedness is strongly related to and presupposes the norms of rationality." (Hare and McLaughlin, 1994, p.287) This feels a little too much like a 'get out of jail free card' - if the concept of open-mindedness presupposes rationality, then of course one should be open-minded - but what does the concept of open-mindedness add to the notion of rationality?

What these accounts are essentially attempting to do is to define open-mindedness as a normative concept. They want to define open-mindedness in such a way that makes it *categorically good*: so that more open-mindedness is always better, and that total open-mindedness is an ideal to be attained. I'll explain in this section and the next why I think characterising open-mindedness in this way is both misguided and unnecessary. A first reason to think that trying to characterise open-mindedness normatively is not a fruitful approach, of course, is that attempts to do so so far have not been particularly successful - but I think there is more to the problem than this. Attempts to characterise open-mindedness as a normative concept end up equating it with 'rationality' more broadly, and so the claim that people "should be more open-minded" risks feeling rather trivial - saying nothing over and above, "people should be more rational."

If we want to think of open-mindedness as a normative concept, an interesting question to ask is how it relates to other theories of normative rationality. Would a totally rational agent be maximally open-minded? I think its clear that the answer here is no. Even a perfect Bayesian (i.e. an agent who always updates their beliefs in perfect accordance with Bayes' rule) faces constraints - limited time and processing power - meaning that seeking out more information, considering more hypotheses, and making fewer assumptions, is not necessarily always better. A perfect Bayesian updater still faces two challenges: the challenge of deciding what and how much additional information to seek out, and the challenge of assessing the expected value of different actions given their competing goals.

When it comes to these two additional challenges, the normative response is clearly not always to be 'more open-minded.' The concept of the *value of information* (Howard,

1966) - how much you would be willing to pay for additional information prior to making a decision - makes clear that more information is not always better, if the costs of obtaining it outweigh the amount it allows you to improve your decision-making. And *Bayesian decision theory* (James, 1985) makes explicit the idea that we have to consider and make tradeoffs between different goals when making decisions - which might sometimes (or even often) mean that being 'more accurate' is not optimal.

Its not clear, therefore, that a perfectly rational agent would be maximally open-minded - or that such an agent gets to avoid the trade-offs weve discussed between greater exploration (more open-mindedness), and 'exploiting' what one already knows. If one somehow had unlimited time and cognitive capacity, then more open-mindedness might always be better - but this does not seem like a realistic or even helpful ideal. This suggests that it is not appropriate to think of open-mindedness as a normative concept, as categorically good, in the way that much of the philosophical discussion has tried to. Rather than representing some kind of ideal to be attained, then, open-mindedness is just one side of a certain tradeoff that we face, between the benefits of having a certain, fixed, viewpoint, and the benefits of being able to change one's mind.

5.3.3 Why open-mindedness does not need to be a normative concept

I've argued that many accounts of open-mindedness attempt to characterise it in a way that avoids all downsides - so that more open-mindedness is always better - and that this misguidedly assumes that open-mindedness should be a normative concept. But saying someone is open-minded seems very different from saying they are a perfect reasoner. I think the confusion here arises because there is a common impression that open-mindedness is beneficial or virtuous, which leads one to think that it must *always* be beneficial. But to claim that open-mindedness is a virtue, and that more open-mindedness would be better, we do not need to claim that it's a normative concept, that it's somehow categorically good. It's sufficient to say that, given certain conditions that hold for humans in the real world, a greater degree of open-mindedness would produce better outcomes.

In philosophy, in particular, the focus is on whether open-mindedness should be considered an 'intellectual virtue.' Many of the philosophers we discussed seem to be concerned that, if open-mindedness has downsides, it cannot be an intellectual virtue. But this

seems misguided - if we look more closely at what philosophers generally mean by 'intellectual virtues', they do not seem to be normative concepts in this sense. Intellectual virtues are thought of as those thinking habits/dispositions which help one to form accurate beliefs about the world (closely related to the idea of epistemic rationality, reasoning in ways that lead one to form true beliefs about the world.) We might then think of intellectual virtues as attempting to capture ways of thinking that help people to become more epistemically rational, and open-mindedness as characterising a specific cluster of these ways of thinking which make it easier to change one's mind.

To say that something is a virtue isn't to say that more of it is always better, or that it never comes into conflict with other virtues. It is generally accepted that honesty is a virtue even though it's possible to sometimes be too honest, and even though honesty can sometimes come into conflict with other virtues such as kindness. As Schwarz and Sharpe (2006) argue, virtues should not be considered in isolation, and more of one virtue on its own is not necessarily good - instead we need to consider how different virtues interact with one another. Aristotle's conception was that virtue lies at the mean between two extremes, between two vices - open-mindedness might be said to lie at the mean between dogmatism and indifference, or gullibility.

The fact that open-mindedness might sometimes conflict with certainty or conviction, and that sometimes open-mindedness can go too far, doesn't threaten its status as an intellectual virtue any more than saying it's possible to be too honest threatens its status as a moral virtue. Intellectual virtues aren't supposed to describe ideal standards to be maximised. They seem to be describing something closer to the prescriptive strategies I discussed in the last chapter - ways of reasoning that people can reasonably be expected to develop, given cognitive constraints, that seem likely to improve human reasoning relative to certain goals.

Intellectual virtues such as open-mindedness might then be thought of as describing prescriptive strategies that specifically seem likely to help people attain epistemic goals - to form more accurate beliefs. We do not need to say that open-mindedness is a normative concept, that more open-mindedness is always better, in order to call it an intellectual virtue.

5.3.4 Open-mindedness as an explore-exploit tradeoff

If we accept that open-mindedness is not a normative concept, and that closed-mindedness also sometimes has its benefits, then we might more usefully think of these concepts in terms of a tradeoff. Given the constraints we are operating under as reasoners, we face a tradeoff between the benefits of being able to easily change our minds, and the benefits of making assumptions that save time and effort, that help us make sense of the world. The concepts of open- and closed-mindedness have often been considered opposites, with open-mindedness perceived as good and closed-mindedness as bad - but it might be more appropriate to think of a spectrum. Going too far in either direction is likely to be problematic, with different tradeoffs arising as you move in either direction.

The 'optimal' point on the spectrum will depend on the person, situation, and the relevant goals. If some people's brains are structured such that they find it easier to consider multiple viewpoints at once, or to switch between perspectives, then open-mindedness will be less cognitively costly for them than for others. If some people find uncertainty more aversive or stressful than others, then open-mindedness will be more emotionally costly to them. For certain topics, it may be more important to be able to make quick decisions with conviction - in which case, the costs of open-mindedness go up as the benefits of closed-mindedness increase - whereas for others, deliberation and accuracy may be crucial - in which case the costs of closed-mindedness outweigh the benefits. While we might talk about some people simply being 'more open-minded' than others, it might actually be more appropriate to say that different people face different tradeoffs or incentives, and therefore where the optimal balance between open- and closed-mindedness lies depends on the individual and situation.

The idea of a tradeoff between open- and closed-mindedness seems closely related to the concept of an 'exploration-exploitation tradeoff' in (machine) learning. The tradeoff here is between exploiting what you already know - going to a restaurant you have been to before and know will be pretty good, for example - and exploring to learn more and potentially get better options in future - trying a new restaurant that might be better than your standard one, but which could also be worse. In decision making, the tradeoff is between making a decision now based on the information you have, and delaying the decision to spend more time getting information (buying the first house you see that seems decent versus spending weeks looking at many houses and then deciding.)

There's no way to avoid this tradeoff - no way to get the best of both worlds - and no fully general solutions (though the study of these tradeoffs in computer science has come up with algorithms which provide the optimal solution under certain specific assumptions.) How much one should explore versus exploit depends on your goals, and various features of the situation. We might think of the open- versus closed-mindedness tradeoff as a kind of explore versus exploit tradeoff for forming beliefs: the tradeoff between learning and exploring in order to ensure one forms the most accurate beliefs possible, and 'exploiting' one's current best guess: acting based on what one knows, and saving the extra time and cognitive effort. So just as there's no fully general solution to the explore-exploit tradeoff, there's no fully general answer to how open-minded one should be: it depends on the situation and on your goals.

Exploration is valuable because it helps one avoid getting stuck at a local optimum: a point that looks better than all those surrounding it, but which might not be the best possible option in the entire search space. A helpful visualisation/analogy here is to think of mountain climbing - one might get 'stuck' at the top of a peak and be unsure whether there are higher peaks elsewhere, vision clouded by fog. To explore and avoid getting stuck at a local optimum, one sometimes has to go 'downhill', to where conditions are clearer and it's easier to see all the peaks. Applying this analogy to our discussion of open-mindedness: we might sometimes get 'stuck' in a certain viewpoint, and if we want to learn may need to sometimes do things that feel like going downhill - considering perspectives that don't make sense to us or we don't like, even if this leaves us confused for a while, or having to abandon assumptions that are helpful or comforting. The difficult question is when and how much to explore, especially if we're not really sure where higher peaks are, or if they even exist at all.

5.3.5 Open-mindedness and science

Some of the ideas I've explored in this chapter, and in this thesis more broadly, also seem to echo some discussion in the philosophy of science, and particularly the work of Kuhn (1962, 1963, 1979). Philosophers of science have long been concerned with how theories and paradigms guide knowledge but can also constrain our viewpoints, how to treat anomalies and how many anomalies have to build up before we consider revising or abandoning a theory, and what kind of open-mindedness and creativity are required

for scientific innovation. Here, as with thinking and learning more broadly, there is no straightforward answer: Kuhn talks about what he calls an "essential tension" between tradition and innovation in science - we need tradition and established theories in order to make 'normal' progress, but innovative scientific discoveries require the ability to break from this tradition (Kuhn, 1962, 1979).

In particular, Kuhn suggests that the importance of tradition for scientific progress has been relatively undervalued, compared to the amount of focus there is on open-mindedness, creativity, and innovation - asserting that, "both my own experience in scientific research and my reading of the history of sciences lead me to wonder whether flexibility and open-mindedness have not been too exclusively emphasised as the characteristics requisite for basic research." (Kuhn, 1979, p.139) Kuhn argues that science ideally progresses in two distinct modes: 'normal science', which progresses firmly on the basis of past discoveries and widely accepted base assumptions - and 'revolutionary science' - where the most fundamental discoveries and assumptions are questioned, occasionally resulting in a complete overthrow or rethink of the currently prevailing paradigm (Kuhn, 1962). Though he emphasises how crucial scientific revolutions are - indeed, much of his most famous work focuses on them - he believes it is essential that they are accompanied by extended periods of 'normal science'. In "The function of dogma in scientific research", he emphasises the importance of paradigms, theories, and assumptions: "nature is vastly too complex to be explored even approximately at random... something must tell the scientist where to look and what to look for." (Kuhn, 1963, p.363) This seems closely related to many of the points we have discussed about the importance of assumptions for making sense of the world - and the notion of an "essential tension" between tradition and innovation closely analogous to the idea of a difficult tradeoff between the benefits of open- and closed-mindedness.

One idea we might usefully take from Kuhn's work is that thinking needs both 'normal' and 'revolutionary' stages - but both do not need to occur (and perhaps *cannot* occur) at the same time. Just as science might go through long periods of incremental progress based on a prevailing paradigm before that paradigm is challenged, we might do something similar in our own thinking. Most of the time, we can go about our lives not questioning our most fundamental beliefs and theories, building on our assumptions and learning within those constraints (to be constantly questioning these seems impossibly cognitively demanding and perhaps even incapacitating). However, we also

need occasional 'revolutionary' periods where we question even our most fundamental assumptions, and genuinely explore the possibility that an alternative perspective might be better. Of course, the really challenging question is when these 'revolutionary' periods should occur, how often, and what prompts them. As Kuhn suggests with science, we might look out for anomalies - things that don't quite fit with our current beliefs - and when a certain number build up, recognise the importance of taking a step back. Or we might even decide (as individuals, groups, or society) to schedule times at regular intervals - every few months, every year, or longer, depending on the issue - for challenging our assumptions.

Before we even get to the question of when and how to challenge our assumptions, there is a simpler challenge: simply being aware of what our assumptions are. By their nature, assumptions are things we rarely think about, and are often not aware we're making. If we're not even aware of what our assumptions are, then it seems impossible that we'll ever change them. Rather than saying that people should be "more open-minded" or "less biased", therefore, I think a more actionable and clearly beneficial goal would to help people recognise what their assumptions *are*, when they are making them - so that they can actually notice when they have experiences that conflict with them, when anomalies arise.

A final interesting point that Kuhn makes is that rather than expecting individual scientists to balance this tension between tradition and innovation, we might simply want different kinds of scientists for different kinds of research. The 'inventive personality', he suggests, may simply be a very different kind of person from the basic scientist - and both are equally valuable to the progress of science in different ways (Kuhn, 1979). Similarly, from the perspective of societal progress, perhaps it is less important that each individual person get the perfect balance between open- and closed-mindedness - and more crucial that we have societies and institutions that balance different kinds of personalities on these dimensions: open-minded, innovative thinkers on the one hand and traditionalists who are very good at working within constraints on the other. We might be concerned that certain kinds of institutions and jobs are particularly likely to attract one or the other kind of person - more closed-minded and traditionalist types being attracted to large, bureaucratic institutions like government (and thus continuing those traditions), and more innovative and open-minded types being attracted to more creative and novel industries. If there genuinely is a stable personality difference here

that we could measure, capturing the benefits of both open- and closed-mindedness and not necessarily suggesting one is better than the other, then we might be able to explore this concern, and the possibility that a better balance here might lead to better outcomes.

5.3.6 Summary: would more open-mindedness be better?

I've suggested that the claim that people should be more open-minded is ambiguous and creates confusion. On the one hand, we might interpret this claim normatively - as saying that open-mindedness is some kind of ideal to be attained and more open-mindedness is always better. This claim is hard to defend, however, because open-mindedness does genuinely seem to have downsides in some scenarios - and attempts to characterise open-mindedness in a way that avoids these downsides risk saying little more than "people should reason well." A second way to interpret the claim that people should be more open-minded, then, is as a prescriptive claim: suggesting that, given the constraints people are operating under and the kinds of errors they tend to make, more open-mindedness would, in general, result in better outcomes.

More generally, I suggested thinking about open- and closed-mindedness in terms of a tradeoff between the benefits of certainty and the ability to change one's mind. This is a tradeoff we face as imperfect reasoners, and there is not necessarily any general, 'correct' solution to this tradeoff. Whether or not more open-mindedness is better, therefore, depends not on some normative theory but on how we think people actually navigate this tradeoff. There is no optimal point on the spectrum across all scenarios - what is optimal depends on the situation and on what one's goals are. It is therefore very difficult to defend the broad claim that people "should be more open-minded", even prescriptively - since whether open-mindedness is beneficial depends both on the specifics of the situation, and what goals one is measuring 'benefits' relative to.

Of course, one *might* argue that we can defend the general prescriptive claim on the grounds that on average, across the range of situations people generally encounter, and across the goals generally shared by people, people systematically err towards being too 'closed-minded'. I said earlier in this chapter that the claim that people should be more open-minded seems to implicitly assume that we fall prey to something like a confirmation bias. One way to defend the claim that people should be open-minded

is therefore by saying that open-mindedness helps to counteract a pervasive tendency to make a certain kind of error. But as we have argued in the rest of this thesis, the evidence for this pervasive 'bias' is much weaker than it first seems.

This isn't to say that more specific claims about open-mindedness might not have merit - we might argue that given certain goals, pushing more towards the 'open' side of the tradeoff would better help achieve those goals. In particular, returning to the notion of open-mindedness as an intellectual virtue, it seems plausible one could argue that more open-mindedness is generally better if one's primary goal is forming accurate beliefs - that across a range of scenarios, people tend to err more towards certainty and closed-mindedness than is optimal if one's goal is accuracy. This would be an interesting claim to explore in more detail, both experimentally and theoretically. But of course, we need to be careful to be aware of this 'if' clause - people do, I think quite reasonably, have many other goals than simply having accurate beliefs. And even if more open-mindedness were good for accuracy, it's another step from that to saying that more open-mindedness would be better for the world.

5.4 Implications

I've suggested that our view of what it means to be open-minded - both in academic discussion, and more broadly - is often confused and vague, and the narrative that people 'should' be more open-minded is overly simplistic. If, as is generally agreed, open-mindedness is crucially about thinking and reasoning in ways that make it easier to change and avoid getting stuck in one perspective, this will sometimes have downsides. Changing one's mind is often important but also costly - and sometimes making assumptions and having a fixed perspective can be useful. Rather than putting open-mindedness up on this pedestal as a categorically good thing, I argued, we have to understand the tradeoffs between open- and closed-mindedness, how people navigate these tradeoffs in different situations, and when and how we might do better. In this section I will discuss a few implications of this view for how we think about open-mindedness - particularly how it is studied experimentally in psychology, and more practical implications related to teaching or promoting open-mindedness.

5.4.1 Implications for the psychological study of open-mindedness

In general, psychology research could do more to look at how people navigate the tradeoff between open- and closed-mindedness in different situations, and the costs and benefits that might arise. In what kinds of contexts do people find it particularly easy to remain open-minded and consider a range of perspectives, and in what contexts do people seem particularly likely to 'close' their minds quickly and ignore alternative perspectives? How might the features of different scenarios explain these differences, and to what extent does this correspond with the actual costs and benefits of open- vs. closed-mindedness in those scenarios? Are there situations in which it seems people would benefit from a little more openness in certain ways, or a little more closure? Theoretical and experimental work could help to more clearly lay out a theory of the kinds of factors that influence how easy or difficult it is for people to consider different perspectives and change their minds, and the different costs and benefits of being able to do so. With a clearer background theory like this, we could then begin to identify cases in which being more or less open-minded might be particularly likely to shift the benefit-cost ratio in a positive direction.

In particular, the study of open-mindedness in psychology has largely focused on its benefits: cases where open-mindedness is useful, and where people may fail to be sufficiently open-minded. It might be interesting and informative, therefore, to take a different perspective and study cases where people may be too open-minded, and where more closure may be useful. One way to do this would be by studying open-mindedness as a personality trait on which individuals differ, and then looking at whether more open-minded people find certain tasks or situations more difficult than those who are less open-minded. For example, we might expect more open-minded people to have more difficulty in situations where making quick decisions or judgements is important, or in situations where a lot of highly ambiguous information is available - perhaps performing more poorly by some measures than those scoring higher on 'closed-mindedness' due to over-deliberating. We might also look at situation- or topic-specific open-mindedness and try to explore whether there are contexts in which people in general may err towards being too open-minded.

There are a few areas of psychological research that have looked at related tendencies, but have not commonly been associated with open-mindedness. The literature on

exploration-exploitation tradeoffs in learning and decision-making is obviously relevant, though has not previously been linked to the concept of open-mindedness (as far as we are aware.) For example, Wilson et al. (2014) look experimentally at different strategies people use to make explore-exploit decisions; Lee et al. (2011) develop a psychological model of how people navigate 'bandit problems' (a specific type of explore-exploit problem) and test this model experimentally, and Mehlhorn et al. (2015) comprehensively survey the literature on human and animal behaviour navigating these tradeoffs. Drawing on this and related literature might help us to understand explore-exploit tradeoffs in general, which could provide a useful basis for modelling the tradeoff between open- and closed-mindedness and understanding how people navigate these tradeoffs.

It might also be helpful to draw more on discussion of explore-exploit tradeoffs and related problems in machine learning literature. Research on solving optimization and learning problems recognises that we need assumptions to learn, but that any set of assumptions will create an 'inductive bias' (Mitchell, 1980). Mitchell argues that a totally unbiased system that makes no *a priori* assumptions about the data it sees cannot learn as effectively as a system that uses additional information or certain kinds of 'justifiable' biases. Also closely related are 'no free lunch' theorems in machine learning (Wolpert and Macready, 1997), which broadly imply that, given different assumptions about the environment, some learning algorithms will perform better than others - but none are better across all situations. Therefore, there is no *a priori* justification for what and how many assumptions to make (Forster, 2009). There are clear links here with our discussion of the tradeoffs between making assumptions and being 'open-minded', and the idea that there is no general solution to this problem - no free lunch, as it were. Future research could explore these parallels with machine learning research and what psychology might learn from them in more detail.

Finally, psychological research on how people manage attentional resources might also be useful to draw on, in order to understand the extent to which 'closed-mindedness' may sometimes be a reasonable solution to limited attention. For example, Ansburg and Hill (2003) found that creative and analytical thinkers use attentional resources differently - the former pay more attention to 'peripheral' cues that are not necessarily immediately relevant. If we think creativity is related to open-mindedness, then this might suggest that open-mindedness is similarly underpinned by specific ways of managing attention. Biesanz et al. (2001) find that when highly distracted, participants were much more

likely to interpret information in a way that was consistent with their expectations - suggesting that under limited attentional resources, people make more assumptions and are less able to consider alternative explanations, consistent with the idea that 'closed-mindedness' is in part an attempt to manage limited attention.

A remaining challenge here, of course, is actually measuring open-mindedness in a consistent and reliable way. This is difficult partly because open-mindedness is such a broad concept, and we'd like to be able to measure it both as a personality trait and more behaviourally: to have measures that tell us how open- or closed-minded a person is in a general sense, and how open- or closed-minded a person is being in a given scenario. Another challenge is wanting open-mindedness not just to capture certain behaviour but also to capture a certain motivation - not just reading different viewpoints, but also genuinely trying to consider their merits (an issue that came up in our earlier discussion of selective exposure.) Existing measures of open-mindedness don't seem good enough by these standards: many of them use self-report scales which have their own issues, or only capture certain narrow parts of open-mindedness behaviourally - such as the choice of what information to pay attention to, or how easy it is to generate different arguments. Ideally, I think a measure of open-mindedness would combine both self-reported answers (to at least partially capture general motivation) and multiple different behavioural measures: not just what information a person pays attention to, but also what they do with that information, how frequently they change their minds, and what kinds of assumptions they seem to be making. If open- or closed-mindedness refers to certain different ways of forming and updating beliefs, then it needs to be measured at all stages of this process (similarly to how we suggested confirmation bias needs to be studied at all stages of this process.)

Thinking of open versus closed-mindedness in terms of a kind of explore-exploit tradeoff could potentially simplify this problem somewhat, by helping us to define open-mindedness more precisely. A big part of the challenge for measuring open-mindedness is that it's still a somewhat vaguely defined concept. As an explore-exploit tradeoff, what seems crucial is whether and at what point a person decides to 'fix' on a certain belief - to take it as an assumption, acting as if it were definitely true. On this basis, we could measure how open- or closed-minded a person is in terms of how frequently and quickly they make assumptions or 'fix' beliefs in this way (how, precisely, to measure this is another challenge - but this at least narrows the challenge somewhat.) Experimental

paradigms could be designed specifically to explore how people navigate the tradeoffs of making assumptions versus leaving one's mind open - giving people the option to continue exploring more information or to stop at some point and take what they currently believe as assumptions, and investigating how different incentives, scenarios, and participants navigate this differently.

In general, it might be helpful to pull apart some of the more specific behaviours and attitudes associated with open-mindedness, which are themselves more clearly defined: things like the ability to hold multiple ideas in mind at once, the motivation to seek out and understand varied perspectives, and the degree to which one is quick to make assumptions when handling a new problem. None of these seem to individually quite capture what it means to be open-minded, and each seems to capture something the others do not. But it is much easier to begin to ask how we might measure someone's ability to consider multiple ideas at once, say, than how we might measure this broad, vague, construct of 'open-mindedness'. Research could then focus on these more specific, clearly-defined tendencies falling under the broad umbrella of open-mindedness, understanding how they relate to one another and what outcomes they result in. We might find that then, a cluster of specific traits and behaviours arise that are closely correlated and might together be called something like 'open-mindedness'.

5.4.2 Implications for promoting open-mindedness

I've also suggested that the idea that people "should be more open-minded" expands far beyond abstract psychological and philosophical discussion: it's a pervasive view in wider society, more applied research, and policy suggestions. What are the implications, then, of all we've discussed here for the idea that we should be teaching or promoting open-mindedness?

The most important thing, I think, is that we are clearer about what exactly we mean by promoting open-mindedness, in what situations, and why. Gardner (1996) is probably right that basic moral principles are not a priority here: it does not seem particularly helpful, and might even be harmful, to teach young people to question the wrongness of murder. By contrast, it might be much more helpful to teach young people to be open-minded about different religious or political views: to consider a variety of different perspectives before simply agreeing with those around them, to be open to the possibility

that what they currently believe is wrong. I won't make the argument here, but it certainly seems plausible that politics is an area where people are particularly prone to close their minds too quickly.

As well as focusing on the importance of open-mindedness in specific situations as opposed to in a general sense, I think it would be useful to clarify why open-mindedness is particularly likely to be important in a given situation: with respect to what goal? What, specifically, are the costs if people are not more open-minded in this situation? As I discussed in the last chapter, there are a few broad categories of goal that open-mindedness might help with. Open-mindedness might help us with epistemic goals: helping us to more accurately understand the world. Open-mindedness might also help us with personal, instrumental goals: considering a wide range of possible options before making up our minds will often help us to make better personal decisions. Open-mindedness might also help us with moral or social goals: helping us to better understand and get along with other people. When we talk about how it would be better if people were more open-minded, I think it's always helpful to ask why - better with respect to what goal? In some situations, being more open-minded might be better with respect to one goal - accuracy, say - but come at a cost to another - personal goals, perhaps. Asking this question could help to clarify and resolve some of these tensions, and ensure we don't promote open-mindedness in a general blanket way without considering both the upsides and downsides.

It's also worth sometimes questioning whether it's really open-mindedness that is needed in a given situation, or whether some other closely related notion might be more relevant. In particular, I'm thinking of how open-mindedness is closely related to, but distinct from, concepts like tolerance or 'epistemic empathy'. It's possible to be tolerant: to accept that others have different views and not judge them as bad or stupid because of it - without being all that open-minded: not actually giving those views serious consideration. Similarly, I think it's possible to develop 'epistemic empathy': to put a lot of effort into understanding why different people believe the things they do - without actually being open to being convinced by those viewpoints. Open-mindedness often comes alongside things like tolerance and epistemic empathy, and very likely makes these things easier, but is conceptually distinct. And arguably, especially when one's goals are social, promoting things like tolerance and epistemic empathy may be more useful than open-mindedness itself.

Finally, I think it may be equally if not more important that we teach people *when* and *why* open-mindedness is important, than teaching the skill of open-mindedness itself, since this will help people to identify themselves when it is particularly important to be more open-minded. Sometimes simply telling people to "be more open-minded" may not be all that helpful, if they do not have the cognitive resources or motivation to do so. Rather than simply saying people should be more open-minded, it might be more better to focus on helping people better manage their limited attentional resources given their goals, or somehow providing people with genuine incentives to engage with and understand novel perspectives.

It's not that we can't or shouldn't talk about improving open-mindedness, but that these claims need to be more specific, tailored to specific circumstances and goals, and with a greater understanding of why more open-mindedness is helpful in these circumstances.

5.5 Conclusion

I have argued that, in light of the discussion in the rest of this thesis, we may need to rethink the simple view that open-mindedness is always good. Just as I've argued that it's not clear that reasoning in 'confirmatory' ways is necessarily irrational, it's also not clear that 'closing one's mind' is always a bad thing. Just as the psychology literature (and more popular discussion) has tended to take it for granted that we fall prey to a confirmation bias, it has also been too quick to claim that the antidote to this bias is that we should all be 'more open-minded'. This fails to acknowledge the complex tradeoffs we often face when deciding what to pay attention to and what to believe. It's simply not realistic to be totally open-minded and unconstrained by prior beliefs - being a completely 'blank slate' is neither possible nor desirable.

I discussed the literature on open-mindedness in both psychology and philosophy, noting that the term is vague, and its relationship to confirmation bias unclear - despite the fact the literature on open-mindedness and confirmation bias seem to be tackling very similar questions and problems. I argued that it doesn't make sense to think of open-mindedness as a normative concept, even though it is often talked about as if it were one. This isn't to say that open-mindedness isn't sometimes or even often beneficial, but its benefits need to be weighed against its costs. Rather than thinking of open-mindedness

as good and closed-mindedness as bad, I suggested we might more appropriately think of a spectrum where we face tradeoffs between the opposite benefits of each - much as there are tradeoffs between exploration and exploitation in learning. At best, open-mindedness is a prescriptive notion - specific ways of reasoning that might help to bring human reasoning closer to normative standards. However, this relies on the assumption that in general, people err towards being too closed-minded - an assumption which I don't think is currently backed up by solid evidence.

Psychology research and practical recommendations, I suggested, would do better to focus on the tradeoffs between open- and closed-mindedness in specific situations, relative to explicitly stated goals - and understanding how people do in fact navigate those tradeoffs. It may well be the case that there are specific situations and goals with respect to which people could benefit from being more open-minded. But we also need to consider that the opposite could also be true: that in some situations, with some goals, it might plausibly be better to be more closed-minded - lest we risk being too closed-minded about open-mindedness.

Chapter 6

Summary and discussion

In this final chapter, I will begin by summarising everything I've covered in this thesis so far, tying together some of the key threads. I'll then discuss three remaining issues. First, if the evidence for confirmation bias is really as weak and mixed as I have suggested, why is a belief in confirmation bias so pervasive? Intuitively, I still struggle to give up a belief in something confirmation bias-like entirely - but is this just irrationality on my part, ironically sticking to my belief in confirmation bias despite evidence to the contrary? Second, what does this imply for understanding various 'real-world' problems that confirmation bias has often been invoked to explain? Finally, what are the implications of all I've discussed here for future research on these issues?

6.1 Summary

6.1.1 When is confirmation a bias?

In chapter two, I argued that the case for confirmation bias - that people irrationally confirm whatever they already believe - is much more complex than it first seems. The literature commonly cited as evidence for confirmation bias faces problems on multiple levels. To begin with, the term 'confirmation bias' has not been clearly defined, and has been used by different people in different ways. It's not clear what exactly constitutes confirmation of current beliefs, how to determine what someone's 'current beliefs' even are, or what kinds of reasoning/behaviour we're pointing to here (with some using

'confirmation bias' to refer to bias in the search for information, others using it to refer to bias in evaluation of information, and others arguing that it has to refer to some combination.)

When we define confirmation bias more precisely, it becomes evident that much of the research commonly cited does not show what it claims to. Most (if not all) of the evidence we reviewed fails to show a confirmation bias in the sense of a systematic deviation from normative standards that leads to favouring the focal hypothesis. This is often because it is not clear what the relevant normative standards for the task are, the tasks are too specific to generalise to a 'systematic' tendency, and/or because it's unclear whether supposed 'biases' actually lead to strengthening confidence in the focal hypothesis. In addition, almost all research has looked at biases at different stages of reasoning independently - but we need to understand all stages of the reasoning process, up to the point of actually updating beliefs, if we want to be able to conclude that reasoning actually systematically favours the focal hypothesis.

6.1.2 The mixed evidence for selective exposure

In chapter three, I looked more closely at a specific form of confirmation bias: selective exposure, the tendency to search for information that supports what one already believes. Though selective exposure makes sense intuitively, and is often discussed as a potential source of belief polarization in areas such as politics, the evidence for selective exposure is actually quite mixed. Studies of selective exposure find that whether or not the effect occurs seems to depend on a long and growing list of moderating factors, suggesting the effect is not particularly strong or robust even if it does exist. I reported the results of six online experiments which attempted to investigate the robustness of selective exposure effects with political issues, a domain where it seems particularly likely to occur and be problematic. In fitting with selective exposure research more broadly, we found very mixed results: including the fact that a very subtle manipulation to how information is presented to people seems to reduce or eliminate the selective exposure effect. I conclude that selective exposure effects, even in this specific domain, seem to be very sensitive to experimental changes and therefore not all that robust.

I discussed a number of issues this raises about the phenomenon of selective exposure, and understanding how people seek out information more generally. One potential

problem is that the selective exposure paradigm is very abstract and does not model real-life decisions particularly well, as well as possibly creating demand effects: people are likely more motivated to appear 'balanced' in these experiments than they would be when seeking out information more naturally. Even if we do not see a strong selective exposure effect in the lab, it may be that people still end up displaying selective exposure 'in real life'. However, if this is the case, it may be a result of complex environmental factors rather than a basic preference for information that supports what one believes.

I also discussed the relationship between selective exposure and bias, and suggested that the literature on selective exposure has been too quick to draw 'normative' conclusions: that selective exposure is problematic, irrational, and something to be prevented. However, it is not clear that this is the case, especially given the abstract ways in which selective exposure has been studied. As I discussed in the review of confirmation bias, whether or not a search strategy leads to bias depends on how that information is then interpreted, and the motivation behind seeking out that information. It's possible that a perfectly rational person might seek out more apparently 'supportive' information if they have good reason to think it will be more informative than counterarguments. Similarly, just because someone reads balanced sources of information does not necessarily mean they are being unbiased, since they might evaluate these arguments in biased ways. This might also help explain the mixed selective exposure results: we don't find a strong effect because the balance of arguments people read is dependent on various other factors and motivations. I also suggested that 'selective exposure', as typically measured, does not actually capture the phenomenon that's most interesting: whether people are biased towards their current beliefs.

6.1.3 Bias, rationality, and confirmation

A deeper problem with the confirmation bias literature, I argued, is ambiguity and confusion surrounding terms like 'biased' and 'irrational'. Failing to clarify different interpretations of these terms, I suggested, underpins a lot of disagreement about whether different reasoning strategies are 'actually rational'. I therefore tried to clearly distinguish some of these different meanings, the disagreements that arise, and the implications for confirmation bias. Building on earlier discussion, I argued that most research on confirmation bias has not done enough to establish that a bias exists in the sense

of a systematic deviation from normative standards. Most studies either fail to discuss normative standards at all, or only show that such a bias arises in relatively narrow domains. Beyond this, different views of what counts as 'rationality' - disagreements about what we should take into account when defining standards against which to compare human reasoning - lead to various ways in which we might say that confirmation bias, even if it exists, is 'actually rational'.

I suggested, based on this, that the normative question, "does a confirmation bias exist, and is it irrational?" needs to be broken down into a number of smaller parts. First, we might ask whether the strategies people use at different stages (search, inference, updating) actually lead to a bias in the sense of systematically deviating from a normative standard, resulting in favouring the focal hypothesis. We can then also ask what the consequences of these strategies are in different situations, taking into account different contexts, goals, and constraints - and whether it seems possible to do 'better' by various standards. The most important thing here is to clearly articulate and distinguish the different standards that human reasoning is being compared to.

6.1.4 Open-mindedness

Finally, I argued that just as it's easy to oversimplify confirmation bias, it's also easy to assume that open-mindedness is always a good thing - but that this assumption may not be entirely justified. I discussed how the term 'open-mindedness' has been covered in both psychology and philosophy, before considering whether it's possible to be *too* open-minded. I argued that it is, and that attempts to define open-mindedness as a normative concept, as something that is always good and an ideal to be attained, are misguided.

Instead of thinking of open-mindedness as good and closed-mindedness as bad, I suggested it might be better to imagine a tradeoff: acknowledging that there are benefits to having firm beliefs and making assumptions on the one hand, and to being flexible and willing to change on the other. Moving further in one direction inevitably comes at the expense of the other. Another way we might then interpret the claim that people 'should' be more open-minded is as a *prescriptive* claim: saying that, given how people reason in practice, they overall tend to err too far in the direction of closed-mindedness (even if more open-mindedness is not *necessarily* better in theory.) However, this sounds

suspiciously similar to the claim that people fall prey to a confirmation bias - which I've spent the rest of this thesis challenging.

Stepping back, the broad point running throughout this thesis, drawing together research on confirmation bias, selective exposure, rationality, and open-mindedness, is the following. It seems, intuitively, like it is irrational for our existing beliefs to influence how we reason about the world. But when we actually try to pin down more precisely how this is irrational, things very quickly get complicated. Presumably it's acceptable for our prior beliefs to influence how we think about new information *sometimes* and in some ways - but it's not clear where to draw the line, where this becomes irrational or problematic. Whether or not we judge something as 'irrational' also depends on what the normative standard is - whether we're judging rationality relative to strict normative models, to what extent we think these standards should take into account cognitive constraints, and what the relevant goals are. It's also often very difficult in practice to actually compare people's judgements to normative standards, without resorting to abstract and contrived scenarios which then have questionable applicability to 'real life'. Ultimately, it may well be that there is no fully general answer to how much one's prior beliefs should influence subsequent thinking, how many assumptions it is reasonable to make, or how one should navigate tradeoffs between exploring alternative possibilities and exploiting existing beliefs. The best we can do is think clearly about the costs and benefits of different strategies in different situations, given different constraints, and with respect to different goals.

6.1.5 Confirmation bias, open-mindedness, and Bayes' rule

At various points in this thesis, I've tried to discuss the concepts of confirmation bias and open-mindedness with reference to Bayes' rule, commonly considered the appropriate normative standard for belief updating. Here I'll briefly summarise how we might understand both open-mindedness and confirmation bias with reference to Bayes' rule, and how this can help us understand the relationship between the two and the issues that arise for each.

I suggested we think of how 'open-minded' someone is as determined by three things: (a) how many different sources of 'data' they seek out, (b) how many different alternative hypotheses they consider, and (c) how few background assumptions they make that

influence their interpretation of the data (and its implications for the hypothesis.) This makes it clear that there is not necessarily any 'optimal' amount of open-mindedness across all situations - all of these things are cognitively costly and time consuming, and it's impossible for us to seek out all the information, consider all alternative hypotheses, or drop all assumptions.

Recall again Bayes' rule in odds ratio form:

$$\frac{Pr(H \mid D)}{Pr(\neg H \mid D)} = \frac{Pr(D \mid H)}{Pr(D \mid \neg H)} \times \frac{Pr(H)}{Pr(\neg H)} \tag{6.1}$$

I suggested thinking of confirmation bias as capturing ways of reasoning that systematically lead one to update more towards the focal hypothesis H than prescribed by Bayes' rule. Looking at Bayes' rule, there are a few different ways this could arise more specifically: (a) seeking out one's data D in a way likely to favour the focal hypotheses, so that Pr($H \mid D$) is likely to be high; (b) failing to consider sufficient alternative hypotheses that might explain the data, and so underestimating Pr($\neg H$); (c) making background assumptions that make the data seem more likely under the focal hypothesis than it in fact is, essentially over-estimating the likelihood ratio. These seem to reflect fairly well some of the different ways confirmation bias has been discussed in the psychology literature: selective exposure being commonly thought of as a case of (a), overconfidence being an example of (b), and interpreting supportive data as more convincing than conflicting data a case of (c) (the background assumption here perhaps being that people with certain views are more trustworthy than others.) Thinking of confirmation bias in this way, it's clearer why actually demonstrating that confirmation bias exists is problematic. Since we cannot possibly seek out *all* available information, or perfectly balanced information, what really counts as biased search? Similarly, since we cannot possibly consider all alternative hypotheses, how many is too few? Short of saying we should make no background assumptions, how many is too many?

Thinking about confirmation bias and open-mindedness in this way also highlights the close relationship between the two. We might start by thinking of confirmation bias as deviating from Bayes' rule in a way that favours the focal hypothesis - but as I have discussed, the problem with considering Bayes' rule the 'ideal' standard for updating is that it's not realistically attainable. Bayes' rule also doesn't capture *everything* that's

required of a rational reasoner - a perfect Bayesian still faces the challenge of how much and what types of information to seek out. When we start to think about the kinds of errors people might actually make, and the constraints they actually face, it's less clear what ideal reasoning looks like: how much information to seek out and what makes it balanced, how many alternative hypotheses to consider, how many assumptions to make. The more we do all these things, the more 'open-minded' we might be said to be, but there's no simple answer to how open-minded we should be - just a question of how to navigate tradeoffs in specific situations with different goals.

6.2 Discussion

In the remainder of this chapter, I will discuss a few particularly interesting questions and issues that arise from this conclusion - that the case for confirmation bias is not as straightforward as it seems, and that it is not necessarily better to be more 'open-minded'. First, if this is the case, then why is a belief in confirmation bias so pervasive and widely accepted? Second, and relatedly, what are the implications for various 'real-world' problems that confirmation bias has often been invoked to explain? Finally, what does this imply for future research on confirmation bias and related tendencies?

6.2.1 Why is the belief in confirmation bias so pervasive?

If the case for confirmation bias is really as weak I've suggested, then why is belief in it so pervasive? Klayman (1995), for example, reviewed the literature on confirmation bias at the time and acknowledged many of the issues I've discussed here - and yet still seems unwilling to give up the idea of a confirmation bias entirely, ultimately concluding that "when people err, it tends to be in a direction that favours their hypotheses." Despite everything I've covered here, *I* still have a hard time entirely rejecting the idea that something like a confirmation bias exists, as a broad human tendency. It still seems to me, based on my experiences, that what people already believe influences their thinking and reasoning in problematic, if not strictly speaking biased, ways. It's tempting to half-jokingly suggest that confirmation bias may have simply fallen prey to its own problem - we're confirmation-biased about confirmation bias, seeking out and interpreting evidence to reinforce the concept while ignoring the other side of the issue. I pointed out that

research focuses almost exclusively on cases where we are 'too influenced' by prior beliefs - and not enough on situations where we are influenced by prior beliefs the right amount, or even too little. But this creates an almost-paradoxical situation: if we're claiming confirmation bias isn't a robust phenomenon, then we can hardly invoke confirmation bias in order to explain why belief in it is so pervasive...

Of course, the claim made in this thesis isn't that confirmation bias, or something close to it, *doesn't* exist: but just that the evidence commonly cited for it is much weaker than is generally supposed, and there are a lot of complex issues that aren't adequately dealt with in the existing literature. This isn't to say that a strong case *couldn't* be made for something like a confirmation bias - just that the existing case is poorly made. One point I've repeatedly made is that the claim that 'there is a confirmation bias' is simply too broad - if a case for confirmation bias is to be made, I think it needs to be made in a more specific sense - i.e. that people err systematically towards confirming what they already believe under specific circumstances, and that this is irrational with respect to specific goals or some specified normative standard.

6.2.1.1 A theoretical case for confirmation bias?

Some might respond that, even if the evidence for confirmation bias is weak, this says more about the limitations of our experimental methods - and a theoretical case could still be made that something like a confirmation bias is likely to arise. Klayman (1995), for example, gives a learning-based argument for confirmation bias: we'd generally expect confirmation to be more rewarding than disconfirmation, and so the kinds of processes that lead to confirmation are more likely to get reinforced. Chater and Loewenstein (2015) make a similar argument - positing a 'drive for sensemaking', whereby behaviours that help us to simplify and make sense of the world are fundamentally rewarding and so more easily learned. (There is an additional step here, of course, to say that 'confirming' evidence is likely to lead to greater simplification, but this at least makes sense intuitively.) Others have proposed evolutionary explanations for a confirmation bias - Tooby and Cosmides (1992) suggest that since false-negative errors were more costly than false-positives in the ancestral environment (better to think a predator is there and be wrong than the opposite), we'd expect reasoning to develop an asymmetry, with greater focus on confirming than disconfirming current hypotheses. Mercier

and Sperber (2011) suggest that reasoning processes evolved not for truth-seeking but for the social purpose of producing and evaluating arguments - and, given this purpose, we'd expect reasoning processes to be better suited to case-building for a certain position than taking balanced and impartial views on a subject.

However, there are two problems with these theoretical arguments for confirmation bias. First, these theories at best argue that people would develop a *tendency* towards confirmation - but are yet to prove anything about the normative status of such a tendency. Klayman (1995) draws a distinction between confirmation bias as an 'inclination' and as 'faulty judgement' - the latter, but not the former, having normative implications. At best, these theories make the case for confirmation bias as an inclination - but not as faulty judgement, i.e. not as a genuine bias. (And in fact, some of these theories - specifically evolutionary ones - have by some authors been used to claim that an inclination towards confirmation is not so irrational as it might seem.) The second problem is that, if these theories genuinely predict a confirmation bias, then would we not expect them to be backed up by empirical evidence? Rather than asserting that these theories provide reason to believe in confirmation bias despite a lack of empirical evidence, we might conversely argue that the lack of empirical evidence for confirmation bias is evidence against those theories that predict such evidence would exist.

6.2.1.2 Confirmatory reasoning as a (not necessarily irrational) tendency

It's possible that, as Klayman (1995) suggests, confirmation bias exists as an inclination but not as faulty judgement - that is, we tend to reason in ways that confirm rather than disconfirm what we already believe, but there's no clear case that this is non-normative or irrational. This might help explain why a belief in confirmation bias is so pervasive - there's simply a confusion between these two interpretations. We see evidence for *confirmation-bias-as-inclination*, and unreflectively conclude that *confirmation-bias-as-faulty-judgement* follows, without thinking about the distinction between the two. There is a certain trivial sense in which of course we exhibit an inclination towards confirmation: obviously what we already believe influences how we then seek out and interpret information (how could it not?), and to some extent this is rational. It may be that the pervasive belief in confirmation bias is a combination of this basic observation that people have theories about the world, and those theories influence subsequent thinking

- and an inability to understand the difference between when this is rational and when it is not. (This is hardly surprising, since I have argued that there may be no clear, general solution to this question.)

6.2.1.3 Confirmation bias as a convenient explanation

If this is the case, then the next question is: why are we so quick to attribute irrationality here, without properly thinking about what this means? One possibility is that confirmation bias provides a convenient explanation for why others disagree with us: it's easier to write our opponents off as 'irrational' than to try and genuinely understand their perspectives. If I think someone's priors are wrong, then obviously it's going to look to me like they're over-weighting those priors when considering new information. But perhaps they really are reasoning rationally, *given* their prior beliefs. Confirmation bias as a theory may have arisen in part to explain why there is so little agreement among people, even when those people genuinely try to share information and convince one another. But it may be that people disagree largely because they simply have access to different information and background beliefs, and the amount of complexity involved means it's practically impossible for people to actually share all the information they have.

6.2.1.4 Confusing confirmation bias for something else

It's also possible that we've mistakenly attributed to confirmation bias what is better 'blamed' on other, closely related biases - such as overconfidence or motivated reasoning. I suggested before that research has not done enough to distinguish the effects of confirmation bias: unfairly privileging one's prior beliefs when seeking out/interpreting new evidence - from those of overconfidence: simply putting too much weight on those beliefs in the first place - or of motivated reasoning: unfairly privileging whatever one would most *like* to be true (which often, but not always, coincides with what one already believes.) Many studies that supposedly demonstrate confirmation bias could equally be interpreted in terms of one of these alternative tendencies.

However, it's actually unclear whether either overconfidence or motivated reasoning are any better established than confirmation bias is. Hahn and Harris (2014) argue

along similar lines that the literature on motivated reasoning has not done enough to establish a genuine bias, when this is defined more precisely. The strength and generality of supposed overconfidence effects has also been challenged (Erev et al., 1994, Klayman et al., 1999). Furthermore, if motivated reasoning or overconfidence were being mistaken for confirmation bias, we'd still expect to see lots of apparent 'confirmation bias' in experiments, which we'd then explain away as these other notions. The fact that we often don't find such evidence may in fact provide further indirect evidence against these other closely-related biases.

6.2.1.5 The role of emotions

Another issue that is worth brief discussion here is how emotions influence reasoning. Might emotions play a role in driving confirmation bias-like phenomena in the real world, in a way that hasn't quite been captured by the studies and research discussed in this thesis?

Theres certainly a substantial body of research to support the idea that emotions do influence how people seek out and process information, some of which has been mentioned in parts of this thesis - the literature on motivated reasoning suggests that people seek and process information to favour whatever they want to believe Kunda (1990) [1], and some researchers have hypothesized that selective exposure is more likely to occur when the materials used are especially emotionally triggering (e.g. Taber and Lodge, 2006).

The basic idea here is that people tend to seek out and process information in ways that lead to positive emotions, and avoid negative ones (Savolainen, 2014). This is a slightly different claim from that of confirmation bias, as what one wants to believe, or what makes one feel good, isn't always aligned with just preserving one's current perspective. But it certainly seems reasonable to suggest that emotional responses may play a role in confirmatory reasoning - that encountering ideas that conflict with what we believe, or abandoning deeply held assumptions, are very emotionally difficult experiences (and not just things we struggle with cognitively.) To support this idea, Chater and Loewenstein (2015) argue that humans have evolved a drive for sense-making - that we derive pleasure

[1] Though note that the robustness of this tendency has also been challenged by Hahn and Harris (2014)

from things that 'make sense', things our brains can simplify easily - and have an aversion to too much uncertainty or complexity.

However, I think this idea that emotions may play a role in confirmatory reasoning fits with the broader claims of this thesis - that confirmation bias is nonetheless not the consistent, simple phenomenon its sometimes been made out to be. Our emotions can push us in all kinds of directions, not just to confirm whatever we already believe, but sometimes also the opposite - driving curiosity and a desire to learn, for example (Loewenstein, 1994). It does seem that we can sometimes derive pleasure from learning new things and seeing new perspectives, and even sometimes challenging what we already know (Chater and Loewenstein, 2015, argue that we can derive pleasure from 'making sense of' new information and ideas, for example). It is therefore far from obvious that emotions always influence information processing in the direction of confirming existing beliefs.

It is also not clear that emotions role in information processing is entirely irrational. Pham (2007) reviews empirical evidence on emotions and rationality from multiple disciplines and concludes that, "any categorical statement about the overall rationality or irrationality of emotion would be misleading." Emotions may often be conveying important information (Oatley, 1996) - emotional resistance to changing ones mind may sometimes be a useful defence process against changing our minds too much and lacking the useful stability of beliefs, as discussed previously. In other cases, seeking positive emotions via certain ways of seeking out and processing information may be a worthy goal in itself. (Though in many other cases, the pursuit of positive emotions in the short-term may come at a long-term cost.) As discussed previously, the key here is balancing the costs and benefits of open-mindedness relative to different goals - and emotions can often give us more information about what those costs and benefits are, or even form part of the calculus themselves.

6.2.1.6 The role of trust in sources

A final possibility is that the appearance of a 'confirmation bias' can be at least partially explained in terms of how people decide how much *trust* to put in different sources of information. When we form judgements and deal with new information 'in real life', the source of a piece of information is often crucial for how much attention we'll

pay to it, and how we interpret any ambiguities in that information. By contrast, in many psychology experiments on confirmation bias, information source is absent or not considered (though of course there are some psychology studies looking at the importance of source credibility on persuasiveness - see Pornpitakpan, 2004, for a review). What most research does not acknowledge, however, is how what one already believes can influence judgements of *source credibility* - I'm more likely to think a source is reliable if they say something I agree with than if they do not.

An important but unresolved issue here is under what circumstances it is rational to allow one's prior beliefs to guide perceptions of source reliability. At least to some degree, this seems entirely reasonable - if someone is advocating for a view that I believe is very unlikely to be true, it seems appropriate for me to downgrade my assessment of how much to trust this source, on this and other matters. But there is a point at which this tendency begins to look like a bias and could lead to dangerously self-reinforcing beliefs: if I always judge people who agree with me as more credible and trustworthy than those who disagree with me, then I am going to end up strengthening my prior views even on the basis of objectively balanced evidence. One thing that seems key here is that my prior views should not be the *only* thing that factor into my assessment of how reliable a source is, especially not just with regards to the specific topic at hand: I also need to judge how often they have been right on similar issues in the past, what their motivations are, how much other people trust them, and so on.

It may also be the case that my judgements of source reliability aren't influenced by how much someone agrees with me so explicitly, but are based in other heuristics that correlate with agreement more indirectly. For example, our implicit judgements of who to trust might often heavily weight how much we like someone, and/or how much they seem 'like us', whether we feel they are part of our 'ingroup'. People who seem 'like me' or part of my social group seem highly likely to agree with me on important issues - and so deciding who to trust in this way seems likely to reinforce my beliefs, even if I'm not deciding who to trust on the basis of my beliefs more directly.

It might be that trusting those who agree with us either directly or more indirectly, and over-weighting this in our judgements of who is a reliable source of information, could easily drive confirmation bias-like effects in the real world. This could be the case even if there is no confirmation bias in the most basic sense: people do not selectively

seek out and interpret information to confirm what they believe, but end up doing so indirectly by being more trusting of sources whose viewpoints are likely to agree with them. This effect might not show up in many studies of confirmation bias, however, because arguments are presented abstractly without sources, and trust in sources is not measured as a relevant variable. If trust in information sources is a key driver of belief confirmation/persistence in real life, we'd expect to see a higher degree of 'bias' in experiments where information sources are made explicit than those in which they are not.

Of course, this is very close to the hypothesis I suggested to explain mixed selective exposure effects earlier in this thesis, and did not find particularly strong evidence for in my experiments. However, I have since discussed the problem that 'selective exposure' simply does not seem to be a good measure of confirmation bias, particularly because it doesn't capture the relevant motivations. We found that people were only slightly more likely to engage in selective exposure when information sources were made explicit - but I have also argued that selective exposure isn't actually a good measure of bias (because people might choose to read arguments from sources they disagree with but intend to ridicule or rebut them, for example.) The fact source reliability did not have a particularly strong effect in these experiments, therefore, does not tell us much about the relevance of source reliability more broadly - and so I think this hypothesis is still worth exploring further.

6.2.2 Confirmation bias and real-world problems

It's worth briefly discussing the implications of all of this for how we think about 'real world problems' that confirmation bias has often been invoked to explain. Confirmation bias hasn't just been discussed as an interesting experimental artefact - it's often suggested that it might underlie problems like prejudice, conflict in politics, and ideological extremism. I think part of the reason confirmation bias has received so much attention is that it really does seem closely related to real problems we see in the world, problems that arise when people have very strong beliefs, refuse to consider the possibility that they might be wrong, and struggle to engage with anyone who disagrees with them.

Nickerson opens his 1998 review by considering whether confirmation bias, "by itself, might account for a significant fraction of the disputes, altercations, and misunderstandings that occur among individuals, groups, and nations." (Nickerson, 1998, p.175) Lilienfeld et al. suggest that confirmation bias may arguably be "the bias most pivotal to ideological extremism and inter- and intragroup conflict." (Lilienfeld et al., 2009, p.191) In popular psychology books and articles, confirmation bias has been blamed for political conflict (Wolfers, 2014), segmented discourse and 'filter bubbles' online (Villarica, 2012), income inequality (Thompson, 2012), and even war (Wright, 2012). In the popular psychology book, "Don't believe everything you think: the six basic mistakes we make in thinking", Thomas Kida suggests that the failure to predict Japan's attack on Pearl Harbour was itself an artefact of confirmation bias: "Kimmel didn't think the United States was in any great danger, and since Hawaii was not specifically mentioned in the report, he took no precautions to protect Pearl Harbour... One hour before the attack on Pearl Harbor, a Japanese sub was sunk near the entry to the harbor. Instead of taking immediate action, Kimmel waited for confirmation that it was, in fact, a Japanese sub. As a result, sixty warships were anchored in the harbor, and planes were lined up wing to wing, when the attack came. The Pacific Fleet was destroyed and Kimmel was court-martialed. Our desire to cling to an existing belief in the face of contradictory evidence can have disastrous effects." (Kida, 2006, p.155)

If the evidence for confirmation bias isn't quite what it might seem, what does this imply for these real-world phenomena - conflict, extremism, dogmatism, prejudice - that it has been so closely linked to? On the one hand, if a confirmation bias genuinely doesn't exist in the sense that's often been supposed, this might suggest that we're thinking about these problems wrongly: that we need to explain them in different terms. In the extreme, the weak evidence for confirmation bias might even make us question whether these problems are as bad as we think they are: are people's political beliefs really as biased as they seem to be? On the other hand, we might consider the very existence of these problems evidence that something like a confirmation bias exists. No matter what we see in the lab, and despite genuine challenges for the study of confirmation bias, it seems very hard to look at the world and claim that people don't seriously struggle with considering viewpoints they disagree with, with changing their minds, and that people don't make selective use of evidence to support what they already believe.

I think there's something to both these perspectives. Sometimes being too quick to

make normative claims - to label something as a 'bias', as 'irrational', might prevent us from actually making a thorough effort to understand the problem. I think there's a risk of something like this happening with things like political polarization and conflict, extremism, and prejudice - to say that these things can be explained by a confirmation bias, that people are 'just irrational', and that's that. This might be helpful if it provided clear solutions - if understanding these problems in terms of confirmation bias made it easier for us to develop fixes, but it's not clear that we have made much progress here. In particular, explaining these problems in terms of confirmation bias can prevent us from trying to understand why certain kinds of domains and beliefs - politics, religion, identity and beliefs about other people - seem to be particularly problematic. Why do these problems seem to arise in politics in a way that they don't in, say, physics? Sure, people might disagree and become overly attached to 'pet theories' in science, but politics seems in a totally different league: scientists don't typically split themselves into directly opposed groups and express hatred and violence towards those who disagree with them in the way that can happen in politics. Clearly, something more than confirmation bias is needed to explain this - we need to understand what it is about the structure of incentives and feedback in certain domains that pushes strongly against truth-seeking and towards dogmatism and tribalism. This is not to say that serious attempts haven't been made to do this, but that sometimes broad cognitive biases can be overused, and that this can hinder further understanding.

However, I also think when we look at the features of a domain like politics there's a case to be made that something like a confirmation bias arises here, and is genuinely problematic, even if the more general case is weaker. The case 'against' confirmation bias made in this thesis is essentially that the current evidence doesn't do enough to demonstrate that there's a systematic tendency towards confirmation, that occurs across a wide range of scenarios, and comes at a genuine cost (specifically to accuracy.) However, this doesn't mean that something like a confirmation bias might not arise in a more narrow, domain-specific sense: that there aren't certain environments in which people seem particularly likely to err in the direction of confirming what they already believe, and in which this might be particularly problematic with respect to certain goals. That is, our reasoning processes might be well-adapted for a broad range of situations we encounter and environments we're operating in, such that it's not clear they result in any systematic errors, or that any very general class of strategies would

be better. At the same time, it's possible that they're very poorly adapted to certain, specific, environments, in the sense of compromising accuracy and/or leading to other problems.

6.2.2.1 Confirmation bias and politics

Before explaining why I think politics might be an environment to which our reasoning processes are poorly adapted, I want to briefly address a potential objection. In the earlier chapter on selective exposure, I made a similar suggestion that politics might be a domain in which selective exposure (and confirmation bias more broadly) might be particularly likely to arise. This drove my choice to use political issues in the studies I ran on selective exposure. However, I found no evidence of a consistent selective exposure effect - which seems to go against my claim (that I am now returning to) that politics may be a 'special case'. However, I firstly think the lack of evidence for selective exposure in politics can be fully explained by the fact that selective exposure is actually a poor measure of bias. As discussed earlier, it only captures a simple behaviour and not the motivations behind that behaviour, so is difficult to interpret one way or the other (someone might read many arguments they disagree with, for example, with the sole intention of coming up with the strongest rebuttals possible, therefore strengthening their views.) Secondly, as I will elaborate on below, my point is not that people are necessarily biased about politics in this very broad, vague, sense (having repeatedly warned against vague, broad attributions of bias.) Rather, my claim is that politics is a domain where there are very many competing goals at different levels (accuracy, social acceptance, long versus short-term, individual versus societal) - and that the way people reason about politics seems likely to favour some of these goals over others. I think a case can be made that this is irrational by certain standards, or at least problematic - if, say, reasoning favours short-term individualistic goals over long-term societal ones. But to some extent, of course, this depends on one's definition of 'rationality'.

With this in mind, there are two main reasons why I think a domain like politics might be worth further consideration as a special case. First, the incentives in politics are particularly likely to reward reasoning strategies that lead to confirmation, confidence, and consistency, and we're particularly unlikely to be rewarded for or receive feedback on the accuracy of our political beliefs. This is firstly because politics is such a social domain

- we seem to tie political beliefs to our identities in a way we don't with many other beliefs, and to form social groups based on these beliefs. This means we're rewarded for having political beliefs that are consistent over time (to keep our identities consistent), and for having political beliefs that are similar to those in our social circles. Secondly, political beliefs are incredibly complex, a lot of the relevant information is unclear and ambiguous, and it's therefore hard to say precisely what counts as solid evidence for or against a given political belief. This means that, in addition to being socially rewarded for having confident, consistent beliefs that other people agree with, we're also rarely punished for getting things wrong - because it's rarely obvious that we have. This combination means that in politics, our reasoning processes are very likely optimised for forming and maintaining political beliefs that help us socially, - and not for actually figuring out the truth about complex political questions.

We might say, therefore, that the way we reason in political domains is well-suited to certain goals - fairly individualistic and short-term ones - but not so well-suited to others - accuracy and the long-term goals of society. From a relatively short-term, individualistic perspective, the goal of maintaining social standing and identity can easily override any desire to be accurate, because we're much more clearly and immediately rewarded for the former than the latter. However, on a longer timescale, and when looking at the perspective of the group or society, this causes problems: such as conflict between groups and individuals, and failure to make progress on understanding important empirical issues. This comes back around and causes problems for individuals, too, as part of the group or society that's malfunctioning - but not quickly enough to influence how they reason and behave.

6.2.2.2 Improving forecasting accuracy

It also seems like a greater degree of 'open-mindedness' would be particularly valuable for improving forecasting in politics and other important areas. Seeking out more varied sources of information, considering a wider range of perspectives and hypotheses, and making assumptions more explicit, seem like they could markedly improve the accuracy of predictions in complex real-world domains, improvements that could well be worth a substantial cost in terms of time and effort. Small improvements in the ability of governments to predict political events and threats could potentially save millions of dollars

and in some cases even save hundreds or thousands of lives. Consider, for example, if the intelligence community had been able to more accurately estimate the probability that Iraq had weapons of mass destruction - it certainly seems plausible that considering more varied perspectives and data would have made a difference here.[2].

Research on forecasting (Mellers et al., 2014, Tetlock and Gardner, 2016) supports this idea that the ability to consider multiple different perspectives helps people to make more accurate predictions. In studies of forecasting, "superforecasters" (those who consistently performed in the top 1% of accuracy) scored highly on measures related to open-mindedness: need for cognition, openness to experience and active open-mindedness, and were judged to be self-critical, able to consider multiple perspectives, and value seeing multiple perspectives (Tetlock and Gardner, 2016).

Political forecasting seems quite obviously to be a case where more exploration could yield great benefits, at relatively low cost - and so it may well be valuable to find ways to push more in this direction. Its also much easier to assess the benefits of greater open-mindedness when applied to questions that have a 'correct' answer (or will have one in future, as is the case with predictions) - because we can judge how successful reasoning is based on how often it reaches the correct answer. A particularly promising avenue for more research, therefore, might be looking at the costs and benefits of increased open-mindedness for making political/real-world forecasts, where we can assess the accuracy of judgements against real-world outcomes.

6.2.2.3 A new way of thinking about bias

Most forms of 'irrationality' that seem genuinely important and problematic seem like they can be understood not necessarily as a fundamental bias that holds systematically across a wide range of scenarios, but rather as cases where there are conflicts between goals or interests at different levels, and where reasoning processes develop to optimise for goals at the lower level at the expense of a higher level. We might separate these into two cases: cases where *evolution* shapes reasoning in certain ways, and cases where *learning* shapes reasoning. In the former case, reasoning processes develop and change over generations based on which processes have the evolutionary advantage. This can

[2]In fact, its been suggested that a large mistake here was the simple failure to consider base rates - i.e. the base probability of any country in the world, or in the middle east, possessing such weapons (Chang et al., 2016)

cause problems when reasoning processes or behaviours that are 'good for the genes', are at odds with goals at the individual level. One example of this might be having an incredibly sensitive fear response - perhaps adaptive in an environment where there's constant threat of being eaten, but a real hindrance in the modern environment where people have anxiety attacks unprovoked by any real threat to their survival. In the latter case, reasoning processes and behaviour are shaped over shorter time frames, within the lifespan of an organism, as an agent gets rewards and punishments for taking different types of actions. Here behaviours that are immediately and obviously rewarding are likely to be reinforced over those that have longer-term benefits, and behaviours that benefit the individual are likely to be rewarded over those that benefit the larger group. I suggested that the way people reason about politics might fall in this category. Another classic example here is procrastination - arising from the difficulty we often have working on things where the reward is far-off and abstract.

I think this kind of explanation, though far from complete, is a much better start to understanding problems with how people reason and form beliefs in politics and related areas, than simply saying we have a 'confirmation bias'. It also provides more direction for solving these problems than simply saying we need to reduce confirmation bias: we might not want to change the way people reason in general, but rather somehow change the incentive structures in specific domains so that the kinds of conflicts discussed are less likely to arise.

6.2.3 Implications for future research

I'll finally make a few broad suggestions for future research based on what I've discussed here.

6.2.3.1 More attention and clarity around normative issues

One of the biggest issues I've commented on is a lack of clarity around normative issues - a lack of clear normative standards, ambiguous use of terms like 'bias' and 'irrational', and a tendency to unreflectively draw normative conclusions from purely descriptive findings. Many of the normative issues surrounding confirmation bias - i.e. to what extent one's prior beliefs should influence reasoning, and how to actually measure this experimentally

- are far from simple or resolved. My first suggestion for future research on confirmation bias is therefore that these normative issues should receive more attention, and that the normative standards being invoked should be made much more explicit.

More specifically, research in this area could greatly benefit from more work on developing experimental paradigms which allow explicit calculation of normative standards, from the very basic (i.e. bookbags and pokerchips-style) to those using more naturalistic materials (similar to those used by e.g. Eil and Rao, 2011, Harris and Hahn, 2009, Harris et al., 2012, Möbius et al., 2014). Claims of bias or irrationality would ideally be held to a higher standard - when making such claims, researchers should be pressed to disambiguate their use of these terms and/or specify the precise normative standard (or at least acknowledge when they are using 'bias' in a more colloquial sense.) Finally, more research could focus on understanding what the correct normative standard is in the first place - given certain cognitive constraints and assuming certain goals, for example, what is the 'boundedly optimal' way of seeking out and interpreting new information?

6.2.3.2 Studying different stages of reasoning in conjunction, not as isolated phenomena

I also mentioned the importance of more research looking at reasoning at all stages together, and understanding how processes at different stages interact: from forming initial beliefs, to seeking out additional information, interpreting and evaluating information, and incorporating that information to update beliefs. Most research on reasoning focuses on just one of these stages individually, which is understandable, given how complex even trying to understand what's going on at one of these stages can get. However, we at least need more communication between researchers studying these different stages of reasoning, because reasoning and belief formation are multi-stage processes - and what happens at one stage of the process inevitably influences what happens at later stages. If we want to understand when this process results in errors, we need to understand what's happening at every stage and how the different sub-processes interact.

6.2.3.3 Clarifying the relationship between confirmation bias and related concepts

I've also discussed how confirmation bias seems closely related to other tendencies, such as overconfidence, motivated reasoning, and later, the notion of open-mindedness. However, existing research has often not done a great deal to distinguish the effects of a purported confirmation bias from effects that could equally be explained in terms of these tendencies. In many studies of confirmation bias, for example, confirmation and valence are perfectly co-linear - that is, a person's *current* belief and what is most *desirable* for them to believe are the same thing, and so behaviour that looks like a bias towards confirmation might also be a kind of motivated reasoning. As Eil and Rao (2011) discuss, studies that separate these two variables - by constructing scenarios where information can be either *belief-consistent-but-undesirable*, or *belief-inconsistent-but-desirable*, can help to separate these two different potential effects.

It's also surprising how little research has looked at the relationship between confirmation bias and overconfidence, especially since confirmation bias has often been defined in terms of the latter - as reasoning in ways that leads one to put undue confidence in the current belief. There are a number of questions that could be further investigated here: to what extent do supposedly confirmatory reasoning strategies actually lead to overconfidence? Might overconfidence have other causes than confirmatory reasoning, and so arise independently of any biased information processing - simply because of how we represent our beliefs, or difficulties we have reasoning with probabilities? If overconfidence arises independently of confirmation bias, is it possible that we've simply understood causation the wrong way around - that overconfidence leads to the illusion of a confirmation bias, rather than confirmation bias making people overconfident?

6.2.3.4 More specific research directions

There are also a number of more specific research directions I think would be worthy of more exploration. One is the issue of source credibility: how people decide how much to trust different sources of information could plausibly explain confirmation bias-like effects. It would be interesting to explore in more detail how (a) different information about the source of information affects confirmation bias-like outcomess (building on the

minimal extent to which I did this in my studies of selective exposure), and (b) what we can say about when it is rational for one's prior beliefs to influence judgements of the reliability of a source. Hahn et al. (2009) look at judgements of source reliability and how these judgements, along with the person's prior hypothesis, influence interpretations of new evidence. In this research, the authors assume that prior hypothesis and judgements of reliability are independent - but presumably in reality this is often not the case. It would be interesting to extend some of this work, therefore, looking at what happens when we allow judgements of source reliability to be influenced by prior hypotheses.

Second, I think it would be useful to develop the idea of open/closed-mindedness as closely related to a kind of explore-exploit tradeoff, as discussed in the last chapter. One could draw on the existing literature on explore-exploit tradeoffs in learning (both in psychology and also in machine learning), and try to model the tradeoff between open- and closed-mindedness in this way: as a tradeoff between exploration and exploitation with the goal of forming accurate judgements in different environments. Doing this, one could then ask what the optimal solution to the tradeoff is given certain assumptions, and then design experimental tasks which compare human judgement to these optimal solutions, to see where people tend to err too far in one direction or the other.

Relatedly, I think psychology research could do much more to develop better ways to measure 'open-mindedness': breaking this broad concept down into more specific constituent parts (perhaps separating out things that measure cognitive ability and motivation at different stages of the reasoning process - e.g. ability to hold many possibilities in mind at once, ability to consider alternatives, and motivation to seek out new information.) One could then develop ways to measure these things individually and then do standard tests of correlation between them, internal and external validity, to establish whether it makes sense to combine them under this one construct, 'open-mindedness'.

Finally, I think it would be worthwhile to explore some of the questions related to confirmation bias and open-mindedness in more specific scenarios and under narrower assumptions. This is closely related to the idea of studying open-mindedness in terms of explore-exploit tradeoffs above. For example, we might ask: given a certain goal (figuring out the correct answer to a question, say) certain cognitive constraints (bounds on time and how much information one can process), and some ability to choose the information one seeks out and interpret it in different ways, how do people actually approach the

problem, and are there better approaches they could use that would better solve the problem? We could look at real-world domains where confirmation bias is thought to be an issue, and try to model the situation in this way: assuming that people have certain goals, are operating under certain constraints, and are able to make certain choices. With the problem stated more concretely like this, we can ask what good solutions to this problem look like (given the relevant goals), what tradeoffs people face, and then compare this to how people in fact approach such problems. Though this will of course involve making simplifying assumptions (especially for a domain as complex as political reasoning) it seems likely to result in much more useful insights about how reasoning could be improved than simply observing that what people already believe seems to influence how they think about new information.

Chapter 7

Conclusion: the costs and benefits of making assumptions

What unifies research on confirmation bias and open-mindedness is a concern: that people might sometimes be too slow to change their minds, and liable to get 'stuck' in viewpoints that are wrong. In challenging confirmation bias, I am not saying that we shouldn't be concerned about this - I think we should. However, I think focusing on the idea that people *reason* in ways that confirm whatever they already believe, and that this is irrational, may be misguided. Instead, I suggest, it might be more useful to focus on understanding what kinds of *assumptions* people make, and what the costs and benefits of those assumptions are.

In a strict sense, making assumptions is always irrational - from a purely normative or epistemic perspective, we should always have some degree of uncertainty, and should never assign any beliefs probability 1. However, as I've argued in this thesis, *given* the cognitive constraints we're often operating under, and the variety of goals we're optimising for, we may *have* to make some assumptions. This isn't to say that making assumptions can't sometimes go too far, and that the costs of these assumptions may not sometimes outweigh the benefits - but that making assumptions is not categorically irrational, and we have to weigh the costs and benefits of doing so.

Part of what it means to make assumptions (and one of the benefits of doing so) is that those assumptions then guide how we seek out and interpret information. *Given* that people have certain assumptions, then, the fact that they reason in ways that appear to

'confirm' them is perfectly rational. If people are too slow to change their minds and this causes problems, the problem may not be how they reason given their assumptions - but rather that they are making too many assumptions in the first place.

Consider, for example, how Jern et al. (2014) argue that many of the classic 'belief polarization' findings can be re-interpreted as rational. What they argue is that, *if* we understand people to be making certain background assumptions (about the reliability of different sources, or about a separate issue related to the question), then they might be interpreting evidence perfectly reasonably, *given* those assumptions. The relevant question is then not whether people are interpreting evidence in biased ways, but rather whether the assumptions they are making are reasonable.

I think this is actually a more useful starting point for addressing the concern that people may sometimes be too slow to change their minds. Instead of trying to somehow 'debias' people against this tendency to confirm whatever it is they already believe, we can ask: what kinds of assumptions do people make in different situations, and what are the costs and benefits of those assumptions? How can we teach people to recognise when they are making assumptions, when doing so is helpful, and when it might be more costly? We might learn to distinguish periods of 'normal' thinking, where we allow theories and assumptions to guide our thinking, from occasional 'revolutionary' periods, where we step back and question those assumptions - combining these to balance the benefits of both, as Kuhn (1979) suggests we do in science. It's easy to assume (pun not intended) that just because someone disagrees with us, or are unwilling to change their minds, that they are being irrational. But in fact, it's very difficult to conclude this unless we know exactly what assumptions they are making, and why. Focusing on understanding the assumptions that underlie disagreements in different domains, rather than attributing disagreements entirely to biased reasoning processes, might actually make it easier to resolve them.

Chapter 8

Final reflections

This PhD has been an interesting exercise for me in changing my own mind, and trying to set aside my preconceptions. I chose to study confirmation bias because I genuinely believed it was pervasive, and at the root of may of society's problems. I hoped that my research could help find a way to 'debias' people against it, to reduce this harmful source of irrationality. More generally, I had the impression that people are too slow and reluctant to change their minds, too 'closed-minded', and that pushing in the other direction - helping people to be more open-minded, was clearly a good thing.

However, over the course of my research, I've come to question all of these assumptions. As I begun exploring the literature on confirmation bias in more depth, I first realised that there is not just one thing referred to by 'confirmation bias', but a whole host of different tendencies, often overlapping but not well connected. I realised that this is because of course a 'confirmation bias' can arise at different stages of reasoning: in how we seek out new information, in how we decide what questions to ask, in how we interpret and evaluate information, and in how we actually update our beliefs. I realised that the term 'confirmation bias' was much more poorly defined and less well understood than I'd thought, and that the findings often used to justify it were disparate, disconnected, and not always that robust.

Reasoning that it made sense to start at the beginning of the process, I first focused my attention on selective exposure: this idea that people tend to seek out information they expect to confirm what they already believe. Though I knew that this was not all there was to confirmation bias, I thought that it was a good place to start: if people

don't even engage with different viewpoints at all, how are they ever going to be able to change their minds when they should? My focus therefore shifted from 'fix confirmation bias' to the only-mildly-less-ambitious 'fix selective exposure'. But as I began exploring the selective exposure literature further, and conducting my own experiments, this also began to look misguided: it wasn't clear from either the existing literature, or from the results of my first few studies, that selective exposure was actually a particularly strong or robust phenomenon. Was I trying to fix a problem that didn't exist?

Unsurprisingly, at this point I found myself feeling quite confused about what I was really trying to do. I spent several months trying to make sense of the mixed findings in the selective exposure literature, and trying to square this with a belief I still struggled to let go of: that outside of the lab, people do genuinely seem to have a hard time engaging with different perspectives. Eventually I realised that the problem was that selective exposure was far too narrow, and that my measures weren't really capturing the most important aspects of people's motivation and behaviour. Someone could display no or little selective exposure - reading a balance of arguments from both sides - but still not really be engaging with those arguments in an 'open-minded' way. Equally, the arguments a person chose to pay attention to might make them *look* biased, but actually be chosen for good reason - based on where they genuinely expected to learn more, for example. At this point I felt that further exploring the question of whether and when selective exposure occurs wasn't really going to help me make progress on the questions I was really interested in: whether people really are biased towards their existing beliefs, and what it really means to be open-minded.

This set me off along two closely related paths that would eventually converge, both involving taking a big step back.

First, I began exploring the broader literature on confirmation bias in more detail, along with the associated normative issues. My investigation of the selective exposure literature had made me realise that if I wanted to understand confirmation bias, I couldn't look at different aspects of reasoning independently: I needed to understand how bias might arise at all stages of reasoning, and how these stages interacted with one another. It made me wonder whether other findings I'd taken for granted, like selective exposure, might actually be less robust than I'd thought. I also realised that there were a number of normative questions that the selective exposure research did not adequately deal with

- whether selective exposure is genuinely a 'bias' or 'irrational', and what this really means - that other areas of research might address better. I had been interested in this broader debate around what it means to be rational, and whether it is possible to improve human reasoning, since the beginning of my PhD, so I decided to look into this further.

Second, I started delving into the question of what it really means to be 'open-minded' and how we might measure it. I was dissatisfied with the way that selective exposure was often implicitly taken to be a measure of 'open-mindedness': where open-mindedness seemed to me to be a much broader concept, a concept that selective exposure experiments were far from capturing. I also recognised that open-mindedness was closely related to confirmation bias, but that the term seemed to be somewhat vague, and I wasn't aware of good ways to measure how 'open-minded' someone was being. I therefore wanted to explore the literature on open-mindedness to see if I could get some more clarity on the concept and its relationship to confirmation bias, and to see whether there were better ways to measure open-mindedness than simply what arguments people select to read.

On the first path - exploring the confirmation bias literature and associated normative issues - I realised that most of the findings commonly cited as evidence for confirmation bias were much less convincing than they first seemed. In large part, this was because the complex question of what it really means to say that something is a 'bias' or 'irrational' is unacknowledged by most studies of confirmation bias. Often these studies don't even state what standard of rationality they were claiming people were 'irrational' with respect to, or what better judgements might look like. I started to come across more and more papers suggesting that findings classically thought of demonstrating a confirmation bias might actually be interpreted as rational under slightly different assumptions - and found often these papers had much more convincing arguments, based on more thorough theories of rationality.

On the second path, I realised that most of the interesting discussion around open-mindedness was taking place in the philosophical, not the psychological, literature. In psychology, discussion of open-mindedness largely took it for granted what it means to be open-minded, and focused on developing measures of open-mindedness as a personality

trait based on self-report scales. I was more interested in whether it was possible to measure open-mindedness *behaviourally* (i.e. how open-minded someone is in their thinking about a given topic), which required pinning down this vague term to something more precise. The philosophical discussion of open-mindedness seemed to be trying harder to elucidate what it means to be open-minded: but in doing so, found itself caught up in this tricky question of whether it's possible to be *too* open-minded, and if so, whether it is misguided for us to think we should teach open-mindedness. For a while, I myself got caught up in this elusive quest to define open-mindedness in a way that evades all possible downsides, before realising this was probably neither useful nor necessary.

All of this investigation led me to seriously question the assumptions that I had started with: that confirmation bias was pervasive, ubiquitous, and problematic, and that more open-mindedness was always better. Some of this can be explained as terminological confusion: as I scrutinised the terms I'd been using unquestioningly, I realised that different interpretations led to different conclusions. I have attempted to clarify some of the terminological confusion that arises around these issues: distinguishing between different things we might mean when we say a 'confirmation bias' exists (from bias as simply an inclination in one direction, to a systematic deviation from normative standards), and distinguishing between 'open-mindedness' as a descriptive, normative, or prescriptive concept. However, some substantive issues remained, leading me to conclusions I would not have expected myself to be sympathetic to a few years ago: that the extent to which our prior beliefs influence reasoning may well be adaptive across a range of scenarios given the various goals we are pursuing, and that it may not always be better to be 'more open-minded'. It's easy to say that people should be more willing to consider alternatives and less influenced by what they believe, but much harder to say how one does this. Being a total 'blank slate' with no assumptions or preconceptions is not a desirable or realistic starting point, and temporarily 'setting aside' one's beliefs and assumptions whenever it would be useful to consider alternatives is incredibly cognitively demanding, if possible to do at all. There are tradeoffs we have to make, between the benefits of certainty and assumptions, and the benefits of having an 'open mind', that I had not acknowledged before.

There's a nice irony to the fact that over the course of this PhD, I've ended up thoroughly questioning my own views about confirmation bias and open-mindedness: questioning my assumptions about the value of making assumptions, as it were. I haven't changed

my mind completely - I am still concerned that in some situations, and for certain topics, people really are too dogmatic and could do with exploring more. But I'm certainly more open-minded about this than I was. Whether my increased open-mindedness is a good thing, of course, is another question.

Appendix A

Factors influencing selective exposure

Hart et al. (2009) paper proposes a very basic theory of selective exposure, assuming people have two main motives: to defend their current beliefs, and to be accurate. Based on this assumption, they suggest that selective exposure will occur to the extent that the defense motive outweighs the accuracy motive.

Table A.1 below suggests some possible ways to expand on this model, suggesting that whether selective exposure occurs or not may be more complex than a simple balance of defense and accuracy motives. In particular, what Hart et al. (2009) fail to acknowledge is that defense motives may sometimes lead one to seek out conflicting info (if one expects to be able to rebut or ridicule those arguments easily, for example), and/or that accuracy motives may sometimes lead to seeking out more confirming info (if one has particular reason to want to question those arguments, say.) This means that greater defense motivation does not necessarily lead to greater selective exposure, and vice versa. In the below table I list a number of different relevant motives, under what circumstances they are an are not likely to increase selective exposure, as well as other factors that might moderate the strength/presence of this motive. Though far from a complete theory it does provide some ideas for how Hart et als theory of selective exposure might be extended.

Motive	**When does this motivate selective exposure?**	**When would this fail to motivate selective exposure?**	**Factors moderating the strength/presence of this motive**
Defense - desire to confirm/validate current position	When conflicting evidence is expected to be strong/threatening, when confirming evidence expected to be novel/helpful.	When conflicting evidence is expected to be weak, and expected that being able to rebut counterarguments might strengthen original belief.	Strength/importance of beliefs; Personal characteristics/traits such as need for closure and tolerance of uncertainty; Current emotional state, feeling of threat
Accuracy - desire to know the truth	If one wanted to check/scrutinise arguments for what one already believes	When one wants to challenge what one already believes	Personality factors - e.g. need for cognition; Social incentives - i.e. whether accuracy is socially rewarded; Whether there is a correct answer to the issue or not and whether one expects to discover it
Utility - wanting to learn whatever is most useful given a goal one has	When supportive info is expected to be more novel/informative/useful for given goal	When conflicting info is expected to be more useful	The goals one has at the time
Interest/curiosity - intrinsic desire to learn	When supportive info seems likely to be most interesting	When conflicting info seems likely to be most interesting	Personality factors
Social - desire to gain social approval	When the relevant social group agrees with what one believes, and is generally conformist and dislikes uncertainty	When the relevant social group has different views, or rewards understanding different views; When being able to ridicule or rebut different views might be socially rewarded	Presence or salience of relevant others; Personality factors - agreeableness/need for social approval

Table A.1: Factors influencing selective exposure

Appendix B

Selective exposure studies of social/political attitudes

Table B.1 below summarises the studies from Hart et al.'s (2009) review of selective exposure that look at social or political attitudes, as discussed in 3.3.2

TABLE B.1: Selective exposure studies of social and political attitudes

Paper	**Summary**	**Topic**	**Sample size**	**Measure**
Brannon et al. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology*.	Participants preferred opinion-supporting information on social issues such as abortion and the death penalty, but the effect was larger when the attitude was strongly held.	Social issues including abortion and the death penalty	139	Indicated on a 9-point scale how desirable it was for them to read different articles based on titles
Brechan (2002). Selective exposure and selective attention: the moderating effect of confidence in attitudes and knowledge basis for these attitudes. *Unpublished masters thesis*.	Find overall participants prefer opinion-supporting information, and this effect is slightly stronger for highly confident subjects (though the difference is not significant)	Abortion and euthanasia	105 (study 1), 86 (study 2)	Three measures: 1.Choice of text to read, between a text in favor or opposed to the issue, 2.Relative preference between the two texts on a scale of -3 (absolute preference for articles opposed) to +3 (absolute preference for articles in favor.), 3.Interest in reading each text on a scale of 1 to 7

Paper	Summary	Topic	Sample size	Measure
Clarke and James (1967). The effects of situation, attitude intensity, and personality on information-seeking. *Sociometry*.	Overall found participants were significantly more interested in supportive than conflicting information, but no difference between the two conditions (debate vs. discussion framing)	A variety of topics including political, moral, and religious issues.	79	Shown two magazine article titles for each issue (one on either side of the issue), and asked to indicate which article they would most like to read.
Cotton and Hieser (1980). Selective exposure to information and cognitive dissonance. *Journal of Research in Personality*.	When subjects felt they were forced to write a counter-attitudinal essay, they displayed greater selective exposure than when they felt they had done so willingly.	Attitudes towards nuclear power plants	64 (eight per condition)	Indicated how interested they were in receiving information pamphlets in favour of/against nuclear power plants, on a scale of 0 to 85
Feather (1969). Preference for information in relation to consistency, novelty, intolerance of ambiguity and dogmatism.	Found a significant selective exposure effect, and that this effect was higher when (a) intolerance of ambiguity, as a personality variable, was higher; (b) the information was expected to be novel	American intervention in Vietnam	158 (39 per condition)	Rated the degree and direction of their interest in reading different types of information (pro/anti intervention, novel/familiar)

Paper	Summary	Topic	Sample size	Measure
Hillis and Crano (1973). Additive effects of utility and attitudinal supportiveness in the selection of information. *Journal of Social Psychology.*	Found that the expected utility of information (whether or not it would help them with the later task of giving a speech on the topic) seemed to be a stronger determinant of information choices than selective exposure. I.e. peoples choice of what to read was largely determined by what speech they had been asked to prepare, rather than by what their attitude on the topic was.	Abortion	123 (15 per condition)	Number of pro- and anti-abortion arguments viewed.
Lavine et al. (2005). Threat, authoritarianism, and selective exposure to information. *Political psychology.*	Found a selective exposure effect only for those high on authoritarianism who had been exposed to a mortality salience intervention (asked to think/write about their death) - but not for those low on authoritarianism, or those who had not been exposed to the intervention.	Capital punishment	145 (36 per arm)	Rated interest in three different articles - in favour of, against, and neutral with respect to the death penalty - on a scale of 1 to 7

Paper	Summary	Topic	Sample size	Measure
Lundgren and Prislin (1998). Motivated cognitive processing and attitude change. *Personality and Social Psychology Bulletin.*	Found people exhibited selective exposure when given a 'defense' motive (to give their opinions for a boards decision) but not when given an accuracy motive (to show logic and reasoning abilities), and impression motive (to display agreeableness and other rapport skills) or no motive at all.	Tuition fee increase	63 (15 per arm)	Proportion of opinion-supporting attitudes read out of total
McFarland and Warren (1992). Religious orientations and selective exposure among fundamentalist Christians. *Journal for the scientific study of religion.*	Significantly greater interest in reading pro-fundamentalist articles than anti-fundamentalist articles	Religious beliefs	102 (selected for having fundamentalist Christian beliefs)	Indicated interest in reading articles based on titles, authors, and abstracts - six which clearly supported particular fundamentalist beliefs, and six which opposed them.
Rosenbaum and McGinnies (1973). Selective exposure: an addendum. *The Journal of Psychology.*	Naturalistic study looking at the attitudes of students attending two lectures by partisan speakers - found significantly more pro-Israeli students attended the lecture by the pro-Israeli speaker and the same for pro-Arab students (and all had equal information/opportunity to go to each)	Arab-Israeli conflict	50	Attendance at lectures on the topic

Paper	Summary	Topic	Sample size	Measure
Schulman (1971). Who will listen to the other side? Primary and Secondary Group Support and Selective Exposure. *Social Problems*.	Look at the effect of primary vs secondary group support on interest in supporting/conflicting information - primary group support means how similar ones views are to close friends, secondary group support how similar to a wider peer group. Found those whose views had low secondary group support were more willing to engage with opposing views, as were those with higher primary group support.	Multiple social issues	n/a	Asked to indicate whether they would rather discuss the issue with someone who agreed or disagreed with them.

Paper	Summary	Topic	Sample size	Measure
Schwarz et al. (1980). Interactive effects of writing and reading a persuasive essay on attitude change and selective exposure. *Journal of Experimental Social Psychology*	Found selective exposure was particularly strong when participants first wrote an essay arguing for their position, and then read an argument arguing for the opposite which did not acknowledge two sides to the issue. Found weaker selective exposure effects when people didnt first read the challenging essay, and when they didnt write an essay for their own position. When people first wrote an essay defending their position, and then read another argument in favour of that position which acknowledged two sides to the issue, they were then more interested in reading arguments from the other side.	Opinions on a mandatory year of social service for women	136 (15 per arm)	Reported interest in reading each of six communi-cations - two supporting initial attitude, two opposing, and two neutral
Smith et al. (2007). The role of information-processing capacity and goals in attitude-congruent selective exposure effects. *Personality and Social Psychology Bulletin*.	Find greater selective exposure when people are explicitly given the goal of expressing their attitudes vs. no goal, and when given a time restriction.	Death penalty	69 (study 1 - just goals), 264 (study 2 - goals and time restriction)	Proportion of pro-attitudinal articles selected from a list of 10 items (five on either side)

Appendix C

Ideology-based measures of selective exposure

Here I present the results of analysing data from the last two studies - looking at the impact of 'information source' on selectivity - using an ideology-based measure of selective exposure, as opposed to an attitude-based one. That is, rather than looking at whether peoples issue attitudes correlate with their selections, I look at whether these selections are in line with their broader political ideology C.1, C.2.

Political ideology is measured on a scale from 0 to 6 - where 0 corresponds to 'strongly conservative' and 6 'strongly liberal.' I assume here that people with a more conservative ideology will be more likely to oppose gun control/affirmative action and those with a more liberal ideology more likely to support gun control/affirmative action. These assumptions are backed up by the data - in both datasets and for both topics we find a strong positive correlation between political ideology and 'pro' gun control/affirmative action opinions (with $r > 0.6$ and $p < 0.001$ in all cases.) I then look at the correlation between political ideology and the number of 'pro' gun control or affirmative action arguments chosen, and how this differs between the two conditions (recalling that these differ in how arguments were presented - as simply 'for/against' the issue, or as coming from different interest groups, including political parties.)

Using this measure of selectivity, we do find some weak-moderate evidence of selective exposure in the condition where people made choices based on specific information sources. This lends some support to the hypothesis that selective exposure is more likely

to occur when articles are presented as coming from known sources. However, given the ideology-based measure of selective exposure, this is not particularly novel or surprising - since half the groups people were choosing articles from were political parties - this essentially tells us that republicans/democrats are more likely to choose to read articles from their own political parties (and more likely to do this than to choose articles they agree with on specific issues.)

	Affirmative Action	Gun Control
pro/con	0.07	-0.035
political groups	0.28*	0.31*
both conditions	0.17*	0.097

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE C.1: Correlation between political ideology and arguments selected, experiment 5

	Affirmative Action	Gun Control
pro/con	-0.06	0.086
political groups	0.24*	0.1
both conditions	0.08	0.09

$^{***}p < 0.01$, $^{**}p < 0.01$, $^{*}p < 0.0.5$

TABLE C.2: Correlation between political ideology and arguments selected, experiment 6

Appendix D

Materials used in experiments

D.1 Experiment 1

D.1.1 Opinion measures

Subjects answered the following questions to assess their opinions on the four topics, both before and after reading arguments (with three possible answers: Yes, No or I'm not sure):

(showing separate wordings for those in the "opinion" and "knowledge" conditions)

Income Inequality

- Do you believe that reducing income inequality would benefit society? (opinion framing)
- Does reducing income inequality benefit society? (knowledge framing)

Minimum Wage

- Do you believe that there would be some benefits for society if the minimum wage were abolished? (opinion framing)
- Would there be some benefits to society if the minimum wage were abolished? (knowledge framing)

Death Penalty

- Do you believe that it would it reduce costs to society if the death penalty were reintroduced? (opinion framing)
- Does the death penalty reduce costs to society? (knowledge framing)

Gun Control

- Do you believe that fewer people would be harmed if we had stricter gun laws? (opinion framing)
- Would fewer people be harmed if we had stricter gun laws? (knowledge framing)

Subjects also answered the following question for each topic to assess the strength of their opinion:

- How strong is your opinion on this issue, on a scale of 1 to 10? (opinion framing)
- How confident are you that your answer to the previous question was correct, on a scale of 1 to 10? (knowledge framing)

D.1.2 Arguments used

Subjects are shown just the first (italicized) sentence of each of the following eight arguments, from which they were able to select four to read in more detail.

(Arguments compiled from balancedpolitics.org and major new sources on each topic)

Income Inequality

The US economic system may be unequal, but it generates higher incomes overall than any alternative, so income inequality is not a problem. The US economic system may be unequal, but some argue that it generates higher incomes overall than any alternative. Ultimately, an economy should not be judged in terms of how equal it is, but on how it treats its poorest members. And evidence suggests that despite widening inequality, poverty in the US has been reducing over time - meaning that everyone is better off. High incomes at the top are the result of a free-market system that provides huge incentives

for performance. And the system that delivers that performance means that wealth doesn't come at the expense of the rest of us.

Some inequality may actually be needed to promote economic growth, so income inequality is not a problem. Some inequality may actually be needed to promote economic growth, some economists suggest. Without large financial rewards, things like risky entrepreneurship and innovation would grind to a halt. Since entrepreneurship and innovation substantially contribute to the growth of the economy, reducing the rewards for these endeavours in order to promote equality seems likely to hinder growth. In 1975 Arthur Okun, an American economist, argued that societies cannot have both perfect equality and perfect efficiency, and must choose how much of one to sacrifice for the other. The common-sense position is that there is a policy tradeoff between promoting growth and promoting equality: the tax-and-cost related policies that are associated with faster economic growth are also associated with larger increases in inequality. This suggests that policymakers - and society at large - have to make some tradeoffs when choosing policies, and that "reduce inequality" may not always be the optimal solution for everyone.

There's nothing wrong with some people being well-off if they've earned it, so income inequality is not a problem. Some argue that there's nothing wrong with some people being well-off if they've earned it.* Whether or not income inequality is unjust depends on /how/ the inequality came about, not just that it exists or the absolute magnitude of the inequality. If someone makes more money because he's making the world a much better place than he found it, then that seems ok. There's nothing wrong with someone becoming extraordinarily rich if they happen to provide products or services that are highly in demand. Social mobility is more important than inequality, and inequality doesn't necessarily inhibit mobility, so income inequality is not a problem. Inequality doesn't necessarily inhibit social mobility. To determine whether income inequality is bad, we also need to take into account economic mobility. Between 1975 and 1997 the wealthiest 20% went from receiving 43.2% of the national income to receiving 49.4% of it - whereas the bottom 20% went from receiving 4.4% to receiving 3.6% - this sounds pretty worrying. But it's much less worrying when you realise that the people who constituted the bottom 20% in 1975 are by and large not the same people who constituted it in 1997. In fact, most people who were in the bottom quintile in 1975 had moved out of it by

1997. Income inequality is much less of a problem if we know that people are relatively able to change their position in this economic hierarchy.

Money is worth more to you the less you have of it, so income inequality is a problem as it reduces overall welfare. Money is worth more to you the less you have of it. This is what economists call the 'declining marginal utility' of money - the exact same amount of money can mean very different things for the living standards of different people. A modest loss in money could mean forgone food for one person, forgone medical care for someone higher up the economic food chain, a foregone vacation for someone much richer than that, or be a completely imperceptible change to a genuinely rich person. In this respect, high levels of inequality clearly reduce overall welfare - the people at the bottom of the chain would benefit a great deal more from a small amount more money than those at the top do. Redistributing funds to reduce inequality seems like it would undoubtedly increase overall human well-being: those at the top would hardly miss what they'd lose, whereas it would make a huge different to those at the bottom.

Inequality is correlated with reduced growth, so income inequality is a problem. Inequality is correlated with reduced growth. A recent study by the International Monetary Fund found that societies with lower inequality are correlated with 'faster and more durable growth'. There are multiple channels by which rising inequality may hurt economic growth: the promotion of credit bubbles, diminished opportunity for the lower and middle class to build human capital, and concentrated wealth exerting undue influence over the political system. Jared Bernstein, an economist at the Center for Budget and Policy says that recent empirical and theoretical work reveals potential linkages between high levels of inequality and the housing bubble, the Great Recession, and its aftermath.

Income inequality reduces economic mobility, so income inequality is a problem. Income inequality reduces economic mobility. With increasing inequality, the US has seen a growing gap in 'enrichment expenditures' - the amount of money parents spend on their kids to build their 'human capital' - between the top and bottom incomes. That is, parents at the bottom of the income distribution are - unsurprisingly - spending much less money on their kids' education than those at the top, and this gap is widening. These growing gaps may make it harder for low-income kids to move into upper income brackets during their lifetime.

Income inequality concentrates political influence in the hands of the elite, so income inequality is a problem. Income inequality concentrates political influence in the hands of the elite. It is no accident that strongly conservative views, views that militate against taxes on the rich, have spread more as the rich get richer compared with the rest of us. In addition to directly buying influence, money can be used to shape public perceptions. As the rich get richer, they can buy a lot of things besides goods and services. Money buys political influence, and if used cleverly, it also buys intellectual influence. This obviously raises the possibility of a self-reinforcing process: as the gap between the rich and the rest of the population grows, economic policy increasingly caters to the interests of the elite. As policy increasingly favors the interests of the rich and neglects the interests of the general population, income disparities grow even wider.

Minimum Wage

The vast majority of economists believe that having a minimum wage costs the economy thousands of jobs, so we should abolish the minimum wage. The vast majority of economists believe the minimum wage law costs the economy thousands of jobs. The most fundamental principle of economics is 'supply and demand'. In the case of labor, this means that the supply of workers goes up as wages go up, and the demand for workers by employers goes down as the wages go up. For example, imagine a janitorial job was advertised for hire. If the wage is $100 per hour, thousands of people would want the job. If the wage was $1 per hour, you probably wouldn't find anyone to do it. Conversely, if the government forced the employer to pay at least $7 per hour, the employer might decide not to hire a janitor at all, instead opting to have other staff pick up the duties. Thus, a job would be lost because of the minimum wage. Another example is restaurant employment. A manager might have $10,000 in her monthly budget to hire bus persons. If the wage is set at $7 per hour, the manager may only be able to hire 10 bus people instead of 15. Setting a mandated wage limit disrupts market forces of supply and demand. Just because there is no minimum wage doesn't mean companies can pay whatever they want. Would you work a dishwashing job that paid 25 cents per hour? Would anyone? If they raised the wage to $4 per hour, they might be able to hire a high school student. Consider some highly skilled jobs such as accountant, lawyer, and engineer. Do these people make $5.15 an hour? Obviously, the answer is no. Market factors of supply and demand determine how many jobs are available and what each job would pay. In summary, as the minimum wage goes up, the number of people employed

goes down. When the minimum wage goes down, the number of people employed goes up. Keep in mind: the minimum wage only applies if someone is employed.

Abolishing the minimum wage will allow businesses to achieve greater efficiency and lower prices, so we should abolish the minimum wage. Abolishing the minimum wage will allow businesses to achieve greater efficiency and lower prices. Anytime you give businesses more flexibility, you will increase efficiency and lower prices. Let me give you some examples. Say a McDonald's franchise has a budget of $70 per hour to pay worker wages (without considering benefits and taxes). If that McDonald's must pay $7 per hour, it can hire 10 workers. If it must only pay $5 per hour, it can hire 14 workers. If you go to get a burger, in which situation are you more likely to get it faster? Consider the same situations for a Wal-Mart. In which case are you most likely to find an employee that can take you to an item or answer questions? Thus, businesses can be more efficient and provide better customer service with a lower wage. Another example: imagine three competing coffee shops. All three need to make a certain profit margin to stay in business and make their effort worthwhile. So they all will lower their prices as much as possible while still covering that necessary profit margin. If one of them tries to charge more, customers will simply go to the competitor shops. Now assume the minimum wage is eliminated and each shop can now reduce labor costs by 25 percent. If each doesn't reduce its coffee prices by a proportional amount, it will lose customers to the other two competitors. So by lowering the minimum wage, the public now has to pay less for their espressos. This is obviously a simplistic example, but the principle applies to all businesses. A company cannot simply charge whatever it wants for a product or service. It must always charge a reasonable multiple of its cost; otherwise, it is heading for bankruptcy.

Non-profit charitable organisations are hurt by the minimum wage, so we should abolish the minimum wage. Non-profit charitable organizations are hurt by the minimum wage. Keep in mind that minimum wage laws apply to more than big businesses, they apply to government and non-profit organizations. Charitable organizations are among those most likely to benefit from the elimination of the minimum wage. Let's take an example. Consider a domestic violence shelter. This type of shelter normally needs workers to clean, collect & organize donations, counsel & assist residents, monitor help-lines, provide legal assistance in such things as obtaining restraining orders, and so on. Volunteers help relieve some of the duties, but it's often tough to find dedicated ongoing volunteers

to do the job. After all, volunteers still have to earn a living, raise a family, etc. However, if the charitable organization were able to pay some amount, even a few dollars an hour, it would better be able to build a more steady set of workers. A non-profit organization may simply not be able to afford a $7 per hour pay rate. Thus, non-profits have only two solutions: dissolve their organizations or hire fewer people to provide the charitable service.

The minimum wage can drive some small companies out of business, so we should abolish the minimum wage. The minimum wage can drive some small companies out of business. Many people believe businesses have endless supplies of cash and can easily withstand minimum wage increases or other cost increases. Unfortunately, that's simply not the case. Over 90 percent of businesses fold within the first few years. Every time there is a recession, thousands of businesses go under. Restaurants, which pay wages at or near the minimum wage level, have the highest rate of failure of any business type. Anytime you increase the costs of businesses, you push them closer to the edge. Let's take an example. Imagine a small neighborhood hardware store. This hardware store isn't going to have the logistics and economy of scale advantages of say, Wal-Mart; thus, it must charge more. It probably makes up the price difference with better service. When you raise the minimum wage, it increases the operating costs for that hardware store even more. Thus, it must raise it's prices to cover costs. Eventually, prices get so high that customers conclude that shopping there isn't worth the additional cost. Slowly, the local hardware store is driven out of business.

Adults who currently work for minimum wage are likely to lose jobs to teenagers who will work for much less, so we should not abolish the minimum wage. Adults who currently work for minimum wage are likely to lose jobs to teenagers who will work for much less. Many adults trying to make a living are forced to work minimum wage jobs. If you take away the government-mandated minimum wage, companies will often be able to hire teenagers for a fraction of the price. A business isn't going to pay $5.15 or $7 to an adult factory worker when it can pay $3.50 to a high school student who likely can do the job just as well. Remember that minimum wage jobs usually require little or no training, so it won't be that hard to replace those workers who are displaced. The end result of a minimum wage abolishment is that teenagers, who often are only looking for supplemental income to pay for cars, parties, etc. take work away from those who are trying to pay the rent or support a family.

Workers need a minimum amount of income from their work to survive and pay the bills, so we should not abolish the minimum wage. Workers need a minimum amount of income from their work to survive and pay the bills. Someone working 40 hours per week at $7.15 an hour will make about $1000 per month after taxes. Rent alone can take almost the whole paycheck, especially in high-cost areas of the country like New York and Los Angeles (some states have higher minimum wages than the federal one specifically for this reason). Then, you add in utilities, food, insurance, car payments, credit cards, and on and on. How can a person possibly survive on less? Businesses can better afford the money than citizens scratching to make ends meet.

Businesses have more power to abuse the labor market without a minimum wage, so we should not abolish the minimum wage. Businesses have more power to abuse the labor market without a minimum wage. History shows that businesses left unchecked will abuse their power. Why do you think labor organizations like the Teamsters, United Auto Workers, AFL-CIO, etc. have come into existence? A tight job market, especially during recessions, give citizens the choice of accepting the terms of business or starving. A minimum wage gives business a reasonable floor that should be paid for the labor of others, whether skilled or unskilled.

The minimum wage forces businesses to share some of the vast wealth with the people that help produce it, so we should not abolish the minimum wage. The minimum wage forces businesses to share some of the vast wealth with the people that help produce it. American businesses take in trillions of dollars every year. Is it too much to ask that they share a pittance of it with the people responsible to bringing it to them? We've all read or heard stories of executives with multi-million dollar bonuses, even with companies that lose money. A few dollars extra per hour for the poorest of the poor shouldn't hurt that much.

Death Penalty

Financial costs to taxpayers of capital punishment are several times that of keeping someone in prison for life, so we should not use capital punishment. Financial costs to taxpayers of capital punishment are several times that of keeping someone in prison for life. Most people don't realize that carrying out one death sentence costs 2-5 times more than keeping that same criminal in prison for the rest of his life. How can this be? It has to do with the endless appeals, additional required procedures, and legal wrangling that

drag the process out. It's not unusual for a prisoner to be on death row for 15-20 years. Judges, attorneys, court reporters, clerks, and court facilities all require a substantial investment by the taxpayers. Do we really have the resources to waste?

Life in prison is a worse punishment than the death penalty, so a more effective deterrent, so we should not use capital punishment. Life in prison is a worse punishment and a more effective deterrent. For those of you who don't feel much sympathy for a murderer, keep in mind that death may be too good for them. With a death sentence, the suffering is over in an instant. With life in prison, the pain goes on for decades. Prisoners are confined to a cage and live in an internal environment of rape and violence where they're treated as animals. And consider terrorists. Do you think they'd rather suffer the humiliation of lifelong prison or be 'martyred' by a death sentence? What would have been a better ending for Osama bin Laden, the bullet that killed him instantly, or a life of humiliation in an American prison (or if he was put through rendition to obtain more information).

The possibility exists that innocent men and women will be put to death, so we should not use capital punishment. The possibility exists that innocent men and women may be put to death. There are several documented cases where DNA testing showed that innocent people were put to death by the government. We have an imperfect justice system where poor defendants are given minimal legal attention by often lesser qualified individuals. Some would blame the court system, not that death penalty itself for the problems, but we can't risk mistakes.

The death penalty creates sympathy for the perpetrators of awful crimes, so we should not use capital punishment. The death penalty creates sympathy for the perpetrators of awful crimes. Criminals usually are looked down upon by society. People are disgusted by the vile, unconscionable acts they commit and feel tremendous sympathy for the victims of murder, rape, etc. However, the death penalty has a way of shifting sympathy away from the victims and to the criminals themselves. An excellent example is the execution a few years ago of former gang leader 'Tookie' Williams. He was one of the original members of the notorious Crips gang, which has a long legacy of robbery, assault, and murder. This is a man who was convicted with overwhelming evidence of the murder of four people, some of whom he shot in the back and then laughed at the sounds they made as they died. This is a man who never even took responsibility for the crimes or

apologized to the victims. These victims had kids and spouses, but instead of sympathy for them, sympathy shifted to Tookie. Candlelight vigils were held for him. Websites like savetookie.org sprang up. Protests and a media circus ensued trying to prevent the execution, which eventually did take place – 26 years after the crime itself! There are many cases like this, which make a mockery of the evil crimes these degenerates commit.

The death penalty gives closure to victim's families who have already suffered so much, so we should use capital punishment. The death penalty gives closure to the victim's families who have suffered so much. Some family members of crime victims may take years or decades to recover from the shock and loss of a loved one. Some may never recover. One of the things that helps hasten this recovery is to achieve some kind of closure. Life in prison just means the criminal is still around to haunt the victim. A death sentence brings finality to a horrible chapter in the lives of these family members. The death penalty provides an effective crime deterrent, so we should use capital punishment. The death penalty provides an effective crime deterrent. Crime would run rampant as never before if there wasn't some way to deter people from committing the acts. Prison time is an effective deterrent, but with some people, more is needed. Prosecutors should have the option of using a variety of punishments in order to minimize crime. It also provides a deterrent for prisoners already serving a life sentence. What about people already sentenced to life in prison? What's to stop them from murdering people constantly while in prison? What are they going to do–extend their sentences? Sure, they can take away some prison privileges, but is this enough of a deterrent to stop the killing? What about a person sentenced to life who happens to escape? What's to stop him from killing anyone who might try to bring him in or curb his crime spree?

Prisoner parole or escapes can give criminals another chance to kill, so we should use capital punishment. Prisoner parole or escapes can give criminals another chance to kill. Perhaps the biggest reason to keep the death penalty is to prevent the crime from happening again. The parole system nowadays is a joke. Does it make sense to anyone outside the legal system to have multiple 'life' sentences + 20 years? Even if a criminal is sentenced to life without possibility of parole, he still has a chance to kill while in prison, or even worse, escape and go on a crime/murder spree. The death penalty helps solve the problem of overpopulation in the prison system, so we should use capital punishment. The death penalty helps with the problem of overpopulation in the prison system. Prisons across the country face the problem of too many prisoners

and not enough space & resources. Each additional prisoner requires a portion of a cell, food, clothing, extra guard time, and so on. When you eliminate the death penalty as an option, it means that prisoner must be housed for life. Thus, abolishing the death penalty only adds to the problem of an overcrowded prison system.

Gun Control

Restricting gun ownership will likely reduce the number of violent crimes, so gun control should be increased. Restricting gun ownership will likely reduce the number of violent crimes. Most violent crimes are committed with guns - in 2008, 67% of all murders in the US were committed with firearms. The presence of a gun makes it much easier for a person to kill, makes the killing more instantaneous, more detached - the killer doesn't have to think much about what he is doing. Since guns make it much easier for people to commit violent crimes, and the majority of violent crimes are in fact committed with guns, restricting gun ownership seems highly likely to drastically reduce the number of such crimes.

Suicides and crimes of passion are higher with gun availability, so gun control should be increased. Suicides and crimes of passion are higher with gun availability. It's much easier to act immediately on your impulses when a gun is available. Research finds that residents of homes where a gun is present are 5 times more likely to experience a suicide than residents of homes without guns. Although the reader may or may not disagree with the morality behind suicide being illegal, the fact remains that a gun makes it easier to commit suicide or crimes in a fit of rage, depression, or under the influence of drugs or alcohol. Furthermore, there is conflicting evidence as to whether any kind of substitution occurs - that is, it seems that many of these suicides and crimes wouldn't occur without guns.

Legalized gun ownership means guns have a greater chance of falling into the hands of kids, so gun control should be increased. If guns are legal, it's much more likely that they will accidentally fall into the hands of children. This could lead to some deadly accidents. In one community, Fayetteville, N.C., in 2013, two children - one four years old, and one two - got hold of loaded guns and killed themselves. According to a report by two Boston doctors presented last month at a conference of the American Academy of Pediatrics, about 500 children and teenagers die each year from gunshot wounds and

another 7,500 are injured. None of these tragic deaths would occur if the US had stricter gun control.

Terrorism, school shootings, and other modern circumstances make guns more deadly, so gun control should be increased. Terrorism, school shootings and other modern circumstances make guns more deadly. It's no longer the case that guns are only responsible for the odd sole shooting - in recent times, the availability of guns has led to some horrific events. Just two years ago, in Newtown, Connecticut, 20-year-old Adam Lanza killed his mother, himself, and 26 people at Sandy Hook Elementary School. Lanza had four guns with him, which he apparently had no problem getting hold of. Everyone will agree that it's vital we find some way to stop these horrific crimes - and preventing young people like Lanza from easily accessing guns seems like a necessary first step towards that.

If gun control is increased, law-abiding citizens will be left without any weapons to use in defense, so gun control should not be increased. If gun control is increased, law-abiding citizens will be left without any weapons to use in defense. By one government estimate, Americans use guns to defend themselves or thwart crimes hundreds of times a day. Since criminals will always find some way to obtain their guns, increasing gun control is actually likely to cause a greater imbalance between the number of guns possessed by the general population and criminals. This imbalance is more likely to increase crime and homicide rates than it is to decrease it.

Guns in the possession of citizens are an added protection against government tyranny, so gun control should not be increased. Guns in the possession of citizens are an added protection against government tyranny. Allowing the government access to guns - but not citizens - creates a situation where a country's citizens are essentially at the mercy of the government. Most Americans do not trust their government, or more properly, the people who hold the highest positions in it. Pro-gun citizens consider their guns the same protection. They arm themselves for the possibility of government agents taking away their rights one by one until they live in a police state in which the government is able to do anything it wants because the civilian populace is unarmed and cannot resist. It's hard to predict what will happen in the future, and there is always a small risk of government tyranny - an outcome most would agree would be catastrophic for the United States. Gun rights for citizens reduces the risk of this catastrophic outcome.

Banning guns would take away yet another piece of our liberty, so gun control should not be increased. Banning guns would take away another piece of our liberty, which is one more step to socialism and totalitarianism. When the State starts deciding what is and isnt good for you, when it begins acting upon paternalistic grounds, then we ought to start worrying. This right to keep and bear arms is codified in the Second Amendment to the United States Constitution itself, which reads: A well regulated militia, being necessary to the security of a free state, the right of the people to keep and bear arms, shall not be infringed. To deny civilians this right would be an infringement on liberty.

Banning guns would create another potentially large source of organized criminal revenue as a black market develops, so gun control should not be increased. Banning guns would create another potentially large source of organized criminal revenue as a black market develops. Many drugs are illegal in the United States, and yet a large number of citizens are on everything from pot to cocaine, creating a million-dollar business for drug lords and dealers. This business is growing all the time, and despite the drug laws, drugs are easily available for anyone who wants to buy and promote a life of crime for the sellers. If we increase gun control, this will very likely lead to a black market making multi-millionaires out of illegal gun dealers. This will also lead to a much higher concentration of guns in the hands of killers and criminals - who are perfectly willing to seek out black market opportunities to get their hands on firearms - relative to civilians, who are less willing.

D.2 Experiments 2 and 3

In the next two studies, the arguments used were the same as in the first study. Instead of being shown sentence summaries of the arguments, however, subjects were simply asked "what type of argument they would like to read", and for each topic given the choice between "an argument against issue", or "an argument in favour of issue".

For example, in the case of the minimum wage, the question was presented as follows:

What type of argument would you like to read?

- I'd like to read an argument in support of the minimum wage
- I'd like to read an argument against the minimum wage

The materials used to measure attitude strength and position were also improved, based on learning from the first experiment and exploring related literature in more depth.

We used the attitude strength measures used by Taber and Lodge (2006) - four items measured on a sliding scale from 1 to 100, combined to form a composite measure:

- How much do you personally care about issue? ("I don't care at all" to "I care about it a great deal")
- Compared to how you feel about other public issues, how strong are your feelings regarding issue? ("Not strong at all" to "incredibly strong")
- Some people report that they are very certain of their feelings on issue. Others say they are not certain at all. How certain are you of your feelings on issue? ("Not certain at all" to "incredibly certain")
- People have told us they have thought a lot about some issues and haven't thought at all about some other issues. How would you rate the amount of thinking you have done about issue? ("Very little" to "A great deal")

We also asked two questions to measure attitude opinion, and asked subjects to indicate agreement on a scale from 1 to 9 (rather than Yes/No/Not Sure) to make this measure both more robust and more discriminatory. Subjects therefore answered the following questions to assess their opinions:

Income Inequality

- Differences in income in [*subjects country - UK or US*] are too large
- Ordinary working people do not get their fair share of the nation's wealth

Minimum Wage

- It is important the the government require businesses to pay workers a minimum wage
- Raising the minimum wage would simply make it harder for low-skill workers to find employment, so would be harmful

Death Penalty

- The death penalty is necessary
- It is immoral for society to take a life regardless of the crime the individual has committed

Gun Control

- It is not the government's job to pick and choose the types of weapons it finds acceptable for citizens to own.
- Guns, like cars, should only be used by responsible citizens. Gun control laws just insure that responsible people are using guns in a responsible manner.

D.3 Experiments 4-6

Subsequent studies (the Taber & Lodge replication, and following two studies looking at the information source manipulation) used the exact same materials as Taber and Lodge (2006), which can be found online at this link, but we also reproduce below.

In the last two experiments (looking at information source), the only change from the original method and materials was in how the choices of argument were presented to people.

When asked, Which argument would you like to read?, those in the group names condition were given the same choice as those in Taber & Lodges original - arguments from any of the following groups (with descriptions):

Affirmative Action:

- NAACP: The National Association for the Advancement of Colored People is the oldest and largest Civil Rights Organization in the United States. The NAACP supports affirmative action programs.
- Committee to End Preferences: A citizen group devoted to ending racial and gender preferences, quotes and set-asides. The CtEP opposes affirmative action programs.

- Democratic Party: We provide a sampling of statements made by Democratic politicians on the issue. Historically, the Democratic Party has supported affirmative action programs.
- Republican Party: We provide a sampling of statements made by Republican politicians on the issue. Historically, the Republican party has opposed affirmative action programs.

Gun control

- NRA: The largest organization of gun owners and advocates in the United States. The NRA opposes gun control.
- Citizens Against Handguns: A Maryland-based group devoted to the elimination of handgun sales in the U.S. Citizens Against Handguns favors gun control.
- Democratic Party: We provide a sampling of statements made by Democratic politicians on the issue. Historically, many Democrats have favored gun control.
- Republican Party: We provide a sampling of statements made by Republican politicians on the issue. Historically, many Republicans have opposed gun control.

Those in the treatment ('pro/con') condition were given a more abstract choice, in line with the earlier experiments, between:

Gun control:

- Id like to read an argument in support of gun control
- Id like to read an argument against gun control

Affirmative action:

- Id like to read an argument in support of affirmative action
- Id like to read an argument against affirmative action

D.3.1 Attitude measures

Extremity/Position (9 pt. Likert type agree-disagree response options)

Affirmative Action

- Equal opportunity for African-Americans is very important but it's not really the government's job to guarantee it.
- Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.
- It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.
- Over the past few years blacks have gotten less than they deserve.
- Affirmative action helps to level the playing field, giving blacks an equal chance.
- Blacks do not help themselves by pushing in where they're not wanted.

Gun Control

- Curbing gun violence is very important, but limiting the right to bear arms is not really an effective way to do this.
- Everyone's rights and freedoms are important, but sometimes, as with gun control, it is necessary to limit freedom for the greater public good.
- Guns, like cars, should only be used by responsible citizens. Gun control laws just ensure that responsible people are using guns in a responsible manner.
- Over the past few years our right to bear arms has been eroding. This encroachment on our rights must be stopped.
- There should be no limits on the number of guns someone can own.
- It is not the government's job to pick and choose the types of weapons it finds acceptable for citizens to own.

Attitude Strength (continuous sliding response scale)

- How much do you personally care about issue?
- Compared to how you feel about other public issues, how strong are your feelings regarding issue?
- Some people report that they are very certain of their feelings on issue. Others say they are not certain at all. How certain are you of your feelings on issue?
- People have told us they have thought a lot about some issues and haven't thought at all about some other issues. How would you rate the amount of thinking you have done aboutissue?

D.3.2 Arguments used

Affirmative Action (Pro):

Some whites claim to be victims of affirmative action programs. Nonsense! White Americans have long benefited from a society biased toward white interests, so any current preferences for minorities are only fair. There are no innocent victims of affirmative action. Therefore, we should all support affirmative action programs.

The largest group of Americans to benefit from affirmative action thus far are women. Before 1964, women were excluded from many higher paying occupations and professions based on stereotype, custom and law. There were virtually no women police officers, lawyers, or doctors, for example. Progress has been made, but women still need affirmative action programs.

Nothing in the Constitution prohibits affirmative action. In fact, the Supreme Court upheld affirmative action programs in education in a landmark case. In this case, the Court explicitly stated that "affirmative action is consistent with the Constitution." When a company with a history of past discrimination passes over a white man and hires a qualified minority or woman instead, that isn't 'reverse discrimination.' When black professional athletes were first hired, breaking the 'color barrier' in sports, some white ballplayers lost job opportunities. But that was not 'reverse discrimination,' it was a first step toward ending discrimination.

In the historic words of one African-American leader, "America has given the Negro people a bad check marked insufficient funds." It is about time that America makes

good on its promise of opportunity for all. Affirmative action programs are a necessary first step toward racial equality in America.

In 1990, the average black male worker earned just $731 for every $1,000 earned by a white male worker in a comparable position. Moreover, though white males make up only 43% of the workforce, they occupy 97% of America's top executive positions. After decades of discrimination, only tough affirmative action programs can level the playing field.

Affirmative action programs are very effective. A study from the Clinton administration shows that the percentage of blacks entering the fields of law and medicine has increased from less than 2% to over 10% in the past 20 years. Affirmative action is working.

Who says racism is dead in America? Far from it. Surveys show that a majority of white Americans still believe that African- and Latino Americans are less intelligent, less hard working and less patriotic than whites. Affirmative action programs are an important step toward changing these racist attitudes.

Affirmative Action (Con):

Affirmative action plans treat people based on race, not past or present circumstances. Middle class blacks are given preferences while lower class whites are not! This is unfair reverse discrimination and is itself a form of racism. Affirmative action programs must stop.

Many of the victims of affirmative action are Asian-Americans who have been excluded from top schools due to racial quotas. But they had no role at all in the country's history of discrimination against blacks and they are truly innocent victims! Affirmative action programs are doing more harm than good.

According to a prominent African-American economist, under affirmative action, blacks often get admitted into schools and programs even though they have worse credentials than most white applicants. As a result, their dropout rate is higher. Affirmative action plans harm both blacks and whites and should be stopped.

The Constitution absolutely prohibits racial discrimination, including affirmative action. As one landmark case declared, "our Constitution is color-blind, and neither knows nor

tolerates classes among citizens." Therefore, affirmative action plans are unconstitutional.

The preeminent African-American leader of all time put it best: "Men should be judged by the content of their character, not the color of their skin." Clearly this statement recognizes the injustice of any form of racial preferences. In other words, even the most famous black leader in American history opposed affirmative action!

Merit has always been the most important factor determining success in this country. People of all races and classes can get ahead if they are willing to work. Unfortunately, some Americans expect to be handed a free lunch. Opportunities exist for all, but you have to be willing to pull your weight. Affirmative action violates the merit principle and should be ended.

In a recent national poll, 50% of Americans said they oppose affirmative action. It seems that most of our laws these days favor minorities, and Americans are getting fed up. If a majority of American citizens believe that affirmative action programs are unfair, then why have these laws not been repealed? End affirmative action now!

Affirmative action programs at American universities 'stigmatize' African Americans and other minority students who are assumed to be incompetent because they were admitted based on color, not on merit. Individuals, whether black or white, are far more likely to be successful if they prove their abilities in equal competition rather than receiving unfair and unearned advantages. Affirmative action works to the disadvantage of minorities.

Gun Control (Pro):

A study in a prominent medical journal found that you or a member of your family are 43 times more likely to be killed by your own gun than by an intruder's. Guns aren't the protection many people think they are. We need stricter gun control.

Self-defense arguments for the need of guns are silly: guns only become necessary for self-defense because there are so many guns out there. Thus, guns should be outlawed outright – then we won't need to worry about self-defense.

The United States has the highest murder rate of all industrialized nations. It is also the only industrialized country that has lenient gun laws. We therefore say: bring down the number of guns, bring down the murder rate.

Several recent school tragedies highlight the fact that guns have become a menace to our children. It's very simple: our schoolyards should not be battlefields. We need to reduce access to guns; we need stricter gun control.

In one poll of imprisoned felons, only 27% report buying guns on the black market; the rest got their weapons through legal channels. Obviously, tougher gun controls are needed to keep these legal' guns out of criminal hands.

Recent trials against gun manufacturers have consistently found them guilty, and have forced the gun industry to pay out huge sums of money. If the courts can find good reason to rein in the gun industry, then it is high time for Congress to follow suit.

A study of 743 gunshot deaths reports that 398 occurred in a home where a gun was kept. Only 9 of the 743 were deemed to be justified by the police. It follows that gun owners are not as responsible as they claim to be.

A gun should only be fired if one's life is in danger and all other options have been exhausted. Most 'self-defense' shootings do not meet these criteria. Thus use of guns in self-defense only contributes to the crime rate.

Gun Control (Con):

A main reason why our murder rate is so high is that most crime victims do not resist. These victims are twice as likely to be injured compared to those who defend themselves. Carrying a gun is thus one's ultimate protection against violent crime.

The liberal media distorts gun issues: they only talk about tragedies involving guns. Yet guns were used defensively 2.5 million times last year. The real tragedy would be to outlaw guns – crime would spiral out of control.

The Bill of Rights guarantees the right of all citizens to bear arms. Quite simply, gun control measures are unconstitutional infringements on a basic right of citizenship.

Most privately-owned guns in American are owned by sportsmen and are used for completely peaceful purposes. These guns pose no risk to society, but they are unfairly targeted by gun control legislation.

Stricter gun control laws have not passed Congress, reflecting serious misgivings the American people have about gun control. However, the courts have repeatedly ignored the will of the people, finding gun manufacturers in the wrong. We need to limit the power of the courts in gun control cases.

A national council reported in 1991 that handgun accidents killed less than 15 children under the age of 6. This number is minuscule when compared to the total number of accidental deaths of young children. It simply is not worth outlawing guns to save just a handful of lives.

Laws that require guns to be locked up defeat the purpose of gun ownership: how can I protect my family if I must first retrieve my gun from its locker? We thus need to repeal laws regulating guns in private homes.

Gun control legislation can only regulate guns sold through legal outlets. But these days, many criminals buy their guns illegally. Gun control legislation therefore cannot regulate the most dangerous guns in society.

Bibliography

Adams, J. S. and Stacy, J. (1961). Reduction of cognitive dissonance by seeking consonant information. *The Journal of Abnormal and Social Psychology*, 62(1):74–78.

Adler, J. (2004). Reconciling open-mindedness and belief. *Theory and Research in Education*, 2(2):143–159.

Adorno, T., Frenkel-Brunswik, E., Levinson, D. J., and Sanford, R. N. (1950). The authoritarian personality, studies in prejudice series. In *Studies in prejudice*, pages 1–27.

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgement under uncertainty: heuristics and biases*, pages 294–305.

Anderson, C. a., Lepper, M. R., and Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39(6):1037–1049.

Ansburg, P. I. and Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences*, 34(7):1141–1152.

Arkes, H. R., Wortmann, R. L., Saville, P. D., and Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *The Journal of Applied Psychology*, 66(2):252–254.

Austerweil, J. and Griffiths, T. (2008). A Rational Analysis of Confirmation with Deterministic Hypotheses. *Proceedings of the Cognitive Science Society*, 30(30).

Bacon, F. (1620). *Novum organum*.

Baehr, J. (2011). The Structure of Open-Mindedness. *Canadian Journal of Philosophy*, 41(2):191–213.

Baron, J. (1985). *Rationality and Intelligence*. Cambridge University Press.

Baron, J. (1991). Beliefs about Thinking. In Voss, J. F., Perkins, D. N., and Segal, J. W., editors, *Informal reasoning and education*, pages 169–186.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning*, 7:221–235.

Baron, J. (1996). Actively Open-minded Thinking. *Almanac*, 42(24).

Baron, J. (2000). *Thinking and Deciding*. Cambridge University Press.

Baron, J. (2012). The point of normative models in judgment and decision making. *Frontiers in Psychology*, 3.

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, 32(1):176–184.

Bazerman, M. H. and Moore, D. A. (2013). *Judgment in managerial decision making*, volume 2nd. John Wiley & Sons, 8 edition.

Ben-David, I., Graham, J. R., and Harvey, C. R. (2010). Managerial Miscalibration. *Quarterly Journal of Economics*.

Biesanz, J. C., Neuberg, S. L., Smith, D. M., Asher, T., and Judice, T. N. (2001). When accuracy-motivated perceivers fail: Limited attentional resources and the reemerging self-fulfilling prophecy. *Personality and Social Psychology Bulletin*, 27(5):621–629.

Brannon, L. A., Tagler, M. J., and Eagly, A. H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology*, 43(4):611–617.

Brechan, I. (2002). *Selective exposure and selective attention: The moderating effect of confidence in attitudes and the knowledge basis for these attitudes*. Unpublished master's thesis, University of Florida.

Buchak, L. (2010). Instrumental rationality, epistemic rationality, and evidence-gathering. *Philosophical Perspectives*, 24.

Bullard, S. (1996). *Teaching Tolerance: Raising Open-Minded, Empathetic Children*. Doubleday.

Cacioppo, J. T. and Petty, R. E. (1982). The need for cognition. *Journal of Personality & Social Psychology*, 42(1):116–131.

Chang, W., Chen, E., Mellers, B., and Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5):509–526.

Chater, N. and Loewenstein, G. (2015). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*.

Clarke, P. and James, J. (1967). The Effects of Situation, Attitude Intensity and Personality on Information-Seeking. *Sociometry*, 30(3):235–245.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Political Polarization on Twitter. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media*.

Corner, A., Harris, A. J. L., and Hahn, U. (2010). Conservatism in belief revision and participant skepticism. *Proceedings of the 32nd annual conference of the Cognitive Science Society*, pages 1625–1630.

Cotton, J. L. (1985). Cognitive Dissonance in Selective Exposure. In Zillmann, D., editor, *Selective Exposure To Communication*.

Cotton, J. L. and Hieser, R. A. (1980). Selective exposure to information and cognitive dissonance. *Journal of Research in Personality*, 14(4):518–527.

Crupi, V., Tentori, K., and Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, 116(4):971–985.

Daniel, K. D., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor Psychology and Security Market Under- and Overreactions. *Journal of Finance*, 53(6):1839–1886.

Davies, M. F. (1985). Cognitive-style differences in belief persistence after evidential discrediting. *Personality and Individual Differences*, 6(3):341–346.

Davies, M. F. (1993). Dogmatism and the persistence of discredited beliefs. *Personality and Social Psychology Bulletin*, 19(6):692–699.

De Finetti, B. (1964). Foresight: its logical laws in subjective sources.

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41(1):417–440.

Ditto, P. H. and Lopez, D. F. (1992). Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions. *Journal of Personality and Social Psychology*, 63(4):568–584.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., and Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, 43(2):111–121.

Dugas, M. J., Freeston, M. H., and Ladouceur, R. (1997). Intolerance of uncertainty and problem orientation in worry. *Cognitive Therapy and Research*, 21(6):593–606.

Edwards, W. (1965). Optimal Strategies for Seeking information: Models for Statistics, Choice Reaction Times, and Human Information Processing. *Journal of Mathematical Psychology*, 2:312–329.

Edwards, W. (1982). Conservatism in human information processing. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgement under uncertainty: heuristics and biases*.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Elqayam, S. and Evans, J. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processe. *Psychological Review*, 101(3):519–527.

Evans, J. S. B. T. (1993). Bias and rationality. In Over, D. E., editor, *Rationality: Psychological and philosophical perspectives*, pages 6–30. Taylor & Frances/Routledge.

Feather, N. T. (1969). Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology*, 13(2):129–144.

Feeney, A., Evans, J., and Venn, S. (2008). Rarity, pseudodiagnosticity and Bayesian reasoning. *Thinking & Reasoning*, 14(3):209–230.

Feeney, A., Evans, J. S. B. T., and Clibbens, J. (2000). Background beliefs and evidence interpretation. *Thinking & Reasoning*, 6(2):97–124.

Feldman, L., Stroud, N. J., Bimber, B., and Wojcieszak, M. (2013). Assessing Selective Exposure in Experiments: The Implications of Different Methodological Choices. *Communication Methods and Measures*, 7(3-4):172–194.

Festinger, L. (1957). A theory of cognitive dissonance.

Fischer, P., Jonas, E., Frey, D., and Schulz-Hardt, S. (2005). Selective exposure to information: the impact of information limits. *European Journal of Social Psychology*, 35(4):469–492.

Fischoff, B. and Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90(3):239–260.

Forster, M. R. (2009). Notice: No-Free-Lunches for Anyone, Bayesians Included.

Freedman, J. L. (1965). Preference for dissonant information. *Journal of Personality and Social Psychology*, 2(2):287–289.

Frenkel-Brunswik, E. (1949). Intolerance of Ambiguity As an Emotional and Perceptual Personality Variable. *Journal of Personality*, 18(1):108–143.

Frey, D. (1986). Recent Research on Selective Exposure to Information. *Advances in Experimental Social Psychology*, 19:41–80.

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena.

Gardner, P. (1993). Should we teach children to be open-minded? Or, is the Pope open-minded about the existence of God? *Journal of Philosophy of Education*, 27(I):39–43.

Gardner, P. (1996). Four anxieties and a reassurance: Hare and McLaughlin on being open-minded - British Education Index - ProQuest. *Journal of Philosophy of Education Vol.30,no.2*, 30(2):271–276.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285.

Garrett, R. K., Carnahan, D., and Lynch, E. K. (2011). A turn toward avoidance? Selective exposure to online political information, 2004-2008.

Gelman, A. (2016a). More on replication crisis. *Statistical Modeling, Causal Inference, and Social Science*.

Gelman, A. (2016b). Replication crisis crisis: Why I continue my "pessimistic conclusions about reproducibility". *Statistical Modeling, Causal Inference, and Social Science*.

Gelman, A. and Loken, E. (2013). The garden of forking paths : Why multiple comparisons can be a problem , even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time Multiple comparisons doesn ' t have to feel like fishing. *Department of Statistics, Columbia University*.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650–69.

Gil de Zúñiga, H., Correa, T., and Valenzuela, S. (2012). Selective Exposure to Cable News and Immigration in the U.S.: The Relationship Between FOX News, CNN, and Attitudes Toward Mexican Immigrants. *Journal of Broadcasting & Electronic Media*, 56(4):597–615.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on Estimating the reproducibility of psychological science. *Science*, 351(6277).

Gilbert, D. T., Krull, D. S., and Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4):601–613.

Gilovich, T. and Thomas (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, 44(6):1110–1126.

Griffin, D. and Brenner, L. (2004). Perspectives on probability judgement calibration. In *Blackwell Handbook of Judgement and Decision Making*, pages 177–199.

Gurcay-Morris, B. (2016). *The Use of Alternative Reasons in Probabilistic Judgement*. Degree of doctor of philosophy, University of Pennsylvania.

Hahn, U. and Harris, A. J. L. (2014). What Does It Mean to be Biased: Motivated Reasoning and Rationality. *Psychology of Learning and Motivation*, 61:41–102.

Hahn, U., Harris, A. J. L., and Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, 29(4):337–367.

Harding, P. and Hare, W. (2000). Portraying science accurately in classrooms: emphasizing open-mindedness rather than relativism. *Journal of Research in Science Teaching*, 37(3):225–236.

Hare, W. (1985). *In Defence of Open-Mindedness*. McGill-Queen's University Press.

Hare, W. (1993). *Open-Mindedness and Education*. McGill-Queen's Press.

Hare, W. (2003). The ideal of open-mindedness and its place in education. *Journal of Thought*, 38(2):3–10.

Hare, W. (2006). Why Open-Mindedness Matters. *Think*, 5(13).

Hare, W. and McLaughlin, T. H. (1994). Open-mindedness, Commitment and Peter Gardner. *Journal of Philosophy of Education*, 28(2):239–244.

Hare, W. and McLaughlin, T. H. (1998). Four Anxieties about Open-Mindedness: Reassuring Peter Gardner. *Journal of Philosophy of Education*, 32(2):284–292.

Harris, A. J. L. and Hahn, U. (2009). Bayesian Rationality in Evaluating Multiple Testimonies : Incorporating the Role of Coherence. *Journal of Experimental Psychology*, 35(5):1366–1373.

Harris, A. J. L., Hsu, A. S., and Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Thinking & Reasoning*, 18(2):311–343.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555–588.

Hearst, E. and Wolff, W. T. (1989). Addition versus deletion as a signal. *Animal Learning & Behavior*, 17(2):120–133.

Henrion, M. and Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics*, 54(791).

Hillis, J. W. and Crano, W. D. (1973). Additive effects of utility and attitudinal supportiveness in the selection of information. *The Journal of Social Psychology*, 89:257–269.

Himelboim, I., Smith, M., and Shneiderman, B. (2013). Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter. *Communication Methods and Measures*, 7:3–4.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26.

Hsu, J. (2009). People Choose News That Fits Their Views. *Live Science*.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8).

Iyengar, S. and Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, 59:19–39.

James, O. B. (1985). Statistical decision theory and bayesian analysis. *Springer Series in Statistics, ISBN-10: 0-387-96098-8 and-13*, pages 978–0387.

Jern, A., Chang, K.-M. K., and Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review*, 121(2):206–24.

Johnson, H. M. and Seifert, C. M. (1998). Updating accounts following a correction of misinformation. *Journal of experimental psychology. Learning, memory, and cognition*, 24(6):1483–1494.

Johnson-Laird, P. N., Legrenzi, P., and Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, 63(3):395.

Jonas, E., Greenberg, J., and Frey, D. (2003). Connecting terror management and dissonance theory: Evidence that mortality salience increases the preference for supporting information after decisions. *Personality and social psychology bulletin*, 29(9):1181–1189.

Kahneman, D. (2011). *Thinking , Fast and Slow*. Farrar, Straus and Giroux.

Kelman, H. C. (1967). The Problem Of Deception In Social Psychology Experiments. *Psychological Bulletin*, 67(1):1–11.

Kern, L. and Doherty, M. E. (1982). Pseudodiagnosticity in an idealized medical environment.

Kida, T. E. (2006). *Don't Believe Everything You Think: The 6 Basic Mistakes We Make in Thinking*. Prometheus Books.

Klayman, J. (1995). Varieties of Confirmation Bias. *Psychology of Learning and Motivation*, 32:385–418.

Klayman, J. and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228.

Klayman, J., Soll, J. B., González-Vallejo, C., and Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247.

Konnikova, M. (2016). The psychological research that helps explain the election. *The New Yorker*.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., and Carnot, C. G. (1993). Attitude strength: One construct or many related constructs?

Kruglanski, A. W. (2013). *The psychology of closed-mindedness*. Psychology Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96(4):674–689.

Kuhn, D. and Lao, J. (1996). Effects of Evidence on Attitudes: Is Polarization the Norm? *Source: Psychological Science*, 7(2):115–120.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kuhn, T. S. (1963). The function of dogma in scientific research. In Crombie, A., editor, *Scientific change*, pages 347–369. Heinemann, London.

Kuhn, T. S. (1979). The Essential Tension: Selected Studies in Scientific Tradition and Change.

Kull, S., Ramsay, C., and Lewis, E. (2003). Misperceptions, the Media, and the Iraq War. *Political Science Quarterly*, 118(4):569–598.

Kunda, Z. P. U. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, 108(3):480–498.

Kwong, J. M. C. (2016). Open-Mindedness as Engagement. *Southern Journal of Philosophy*, 54(1):70–86.

Lambie, J. A. (2014). *How to be Critically Open-Minded: A Psychological and Historical Analysis*. Houndmills: Palgrave Macmillan.

Langnickel, F. and Zeisberger, S. (2016). Do we measure overconfidence? A closer look at the interval production task. *Journal of Economic Behavior & Organization*, 128:121–133.

Larrick, R. (2004). Debiasing. In *Blackwell Handbook of Judgement and Decision Making*, pages 316–337.

Lavine, H., Lodge, M., and Freitas, K. (2005). Threat, authoritarianism, and selective exposure to information. *Political Psychology*, 26(2):219–244.

Le Mens, G. and Denrell, J. (2011). Rational learning and information sampling: on the "naivety" assumption in sampling explanations of judgment biases. *Psychological review*, 118(2):379–392.

Lee, M. D., Zhang, S., Munro, M., and Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, 12(2):164–174.

Leitgeb, H. and Pettigrew, R. (2010a). An Objective Justification of Bayesianism I: Measuring Inaccuracy*. *Philosophy of Science*, 77(2):236–272.

Leitgeb, H. and Pettigrew, R. (2010b). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, 77(2):236–272.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Lian, B. (2017). Open your mind to let happiness in. *The Telegraph*.

Liberman, A. and Chaiken, S. (1992). Defensive Processing of Personally Relevant Health Messages. *Personality and Social Psychology Bulletin*, 18(6):669–679.

Lichtenstein, S. and Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26(2):149–171.

Lilienfeld, S. O., Ammirati, R., and Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, 4(4):390–398.

Lodge, M. and Taber, C. S. (2000). Three steps toward a theory of motivated political reasoning. In *Elements of reason: Cognition, choice, and the bounds of rationality*, pages 183–213.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, 47(6):1231–1243.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Lord, C. G. and Taylor, C. a. (2009). Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence. *Social and Personality Psychology Compass*, 3:827–841.

Lundgren, S. T. A. U. and Prislin, R. S. D. S. U. (1998). Motivated cognitive processing and attitude change. *Personality and social psychology bulletin*, 24(7):715–726.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press.

McCrae, R. R. and Costa, P. T. (1997). Conceptions and Correlates of Openness to Experience. In *Handbook of Personality Psychology*, number May, pages 825–847. Elsevier.

McFarland, C., Cheam, A., and Buehler, R. (2007). The perseverance effect in the debriefing paradigm: replication and extension. *Journal of Experimental Social Psychology*, 43:233–240.

McFarland, S. G. and Warren, J. C. (1992). Religious orientations and selective exposure among fundamentalist Christians. *Journal for the Scientific Study of Religion*, 31:163–174.

Mckenzie, C. R. M. (2004). Framing Effects in Inference Tasks And Why They're Normatively Defensible. *Memory & Cognition*, 32(6):874–885.

Meehl, P. E. (1967). Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science*, 34(2):103–115.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., and Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2:191–215.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115.

Mercier, H. and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34:57–111.

Messing, S. and Westwood, S. J. (2012). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, 41(8):1042–1063.

Miri, B., David, B. C., and Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in Science Education*, 37(4):353–369.

Mitchell, T. M. (1980). The Need for Biases in Learning Generalizations.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2014). Managing Self-Confidence. *National Bureau of Economic Research*.

Moore, D. and Healy, P. (2008). The Trouble With Overconfidence. *Psychological Review*, 115(2):502–517.

Moore, D. A., Tenney, E. R., and Haran Ben-Gurion, U. (2015). Overprecision in Judgment. In *Blackwell Handbook of Judgement and Decision Making*.

Mullen, B., Brown, R., and Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, 22(2):103–122.

Mynatt, C. R., Doherty, M. E., and Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology Section A*, 46(4):759–778.

Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. *Psychological Review*, 112(4):979–999.

Neuberg, S. L. and Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, 65(1):113–131.

Newman, J., Wolff, W. T., and Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of experimental psychology. Human learning and memory*, 6(5):630–650.

Nickerson, R. S. (1998). Confirmation Bias : A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2):175–220.

Nisbett, R. E. and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*.

Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian models. *Frontiers in Psychology*, 5.

Oaksford, M. and Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*, 101(4):608–631.

Oaksford, M. and Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford University Press.

Oatley, K. G. (1996). Emotions, rationality and informal reasoning. *Mental models in cognitive science: Essays in honour of Phil Johnson-Laird*, pages 175–196.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Paulhus, D. L. (1991). Measurement and control of response bias. (April):17–59.

Perfors, A. F. and Navarro, D. J. (2009). Confirmation bias is rational when hypotheses are sparse. *Cognitive Science Society.*

Perkins, D., Bushey, B., and Faraday, M. (1986). Learning to Reason. Technical report, Harvard Graduate School of Education.

Peterson, C. and Seligman, M. E. P. (2004). *Character Strengths and Virtues: A Handbook and Classification, Volume 1.* American Psychological Association.

Peterson, C. R. and Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70(1):117–121.

Pham, M. T. (2007). Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of general psychology*, 11(2):155.

Plous, S. (1993). *The psychology of judgement and decision making.* McGraw-Hill , Inc.

Pornpitakpan, C. (2004). The Persuasiveness of Source Credibility : A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology*, 34(2):243–281.

Pyszczynski, T. and Greenberg, J. (1987). Toward an Integration of Cognitive and Motivational Perspectives on Social Inference: A Biased Hypothesis-Testing Model. *Advances in Experimental Social Psychology*, 20:297–340.

Ramsey, F. P. (1926). Truth and probability. *The foundations of mathematics and other logical essays*, pages 156–198.

Redlawsk, D. P. (2002). Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *The Journal of Politics*, 64(04):1021–1044.

Riggs, W. (2010). Open-mindedness. *Metaphilosophy*, 41(1-2):172–188.

Risen, J. and Gilovich, T. (2007). Informal logical fallacies. *Critical Thinking in Psychology*, pages 110–130.

Rokeach, M. (1960). *The Open and Closed Mind: investigations into the nature of belief systems and personality systems.* Basic Books.

Rosenbaum, L. L. and McGinnies, E. (1973). Selective Exposure: An Addendum. *The Journal of Psychology*, 83(2):329–331.

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm. *Journal of Personality and Social Psychology*, 32(5):880–892.

Savolainen, R. (2014). Emotions as motivators for information seeking: A conceptual analysis. *Library & Information Science Research*, 36(1):59–65.

Schulman, G. (1971). Who Will Listen to the Other Side? Primary and Secondary Group Support and Selective Exposure. *Social Problems*, 18:404–415.

Schwarz, B. and Sharpe, K. (2006). Practical Wisdom: Aristotle meets Positive Psychology. *Journal of Happiness Studies*, 7:377–395.

Schwarz, N., Frey, D., and Kumpf, M. (1980). Interactive effects of writing and reading a persuasive essay on attitude change and selective exposure. *Journal of Experimental Social Psychology*, 16(1):1–17.

Sears, D. O., Freedman, J. L., and Festinger, L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31(2):194–213.

Seligman, M. E. P., Steen, T. A., Park, N., and Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *The American psychologist*, 60(5):410–21.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

Simon, H. A. (2000). Bounded rationality in social science: today and tomorrow. *Mind and Society*, 1:25–39.

Skov, R. B. and Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, 22(2):93–121.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., and Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory Cognition*, 20(4):392–405.

Smith, S. M., Fabrigar, L. R., Powell, D. M., and Estrada, M.-J. (2007). The Role of Information-Processing Capacity and Goals in Attitude-Congruent Selective Exposure Effects. *Personality and Social Psychology Bulletin*, 33(7):948–960.

Snyder, M. and Campbell, B. (1980). Testing Hypotheses about Other People: The role of the hypothesis. *Personality and Social Psychology Bulletin*, 6(3).

Snyder, M. and Swann, W. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social . . .*, 36(11):1202–1212.

Snyder, M. and White, P. (1981). Testing hypotheses about other people: strategies of verification and falsification. *Personality and social psychology bulletin*, 7(1):39–43.

Sorrentino, R. M. and Short, J.-A. C. (1986). Uncertainty orientation, motivation, and cognition. In *Handbook of motivation and cognition: Foundations of social behavior*, pages 379–403.

Spiegel, J. S. (2012). Open-mindedness and intellectual humility. *Theory and Research in Education*, 10(1):27–38.

Stangor, C. and Walinga, J. (2010). *Introduction to Psychology*. Flatworld Knowledge.

Stanovich, K. (2011). *Rationality and the Reflective Mind*.

Stanovich, K. E. (2009). Distinguishing the reflective , algorithmic , and autonomous minds: Is it time for a tri-process theory ? In *In two minds: Dual processes and beyond*, pages 55–88.

Stanovich, K. E., Toplak, M. E., and West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior*, 36(September 2015):251–285.

Stanovich, K. E. and West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89(2):342–357.

Stanovich, K. E. and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665.

Stroud, N. J. (2007). Media Effects, Selective Exposure, and Fahrenheit 9/11. *Political Communication*, 24(4):415–432.

Taber, C. S., Cann, D., and Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, 31(2):137–155.

Taber, C. S. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.

Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction.* Random House.

Thompson, D. (2012). How confirmation bias shapes the debate over income inequality. *The Atlantic.*

Todd, P. M. ., Fiddick, L., and Krauss, S. (2000). Ecological rationality and its contents. *Thinking & Reasoning*, 6(4):375–384.

Todd, P. M. and Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3):167–171.

Tooby, J. and Cosmides, L. (1992). The psychological foundations of culture. In Barkow, J., Cosmides, L., and Tooby, J., editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 19–136. Oxford University Press, New York.

Toplak, M. E. and Stanovich, K. E. (2003). Associations Between Myside Bias on an Informal Reasoning Task and Amount of Post-Secondary Education. *Hofer & Pintrich*, 17:851–860.

Tourangeau, R., Rasinski, K. A., Bradburn, N., and D'Andrade, R. (1989). Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology*, 25:401–421.

Trope, Y. and Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, 19(6):560–576.

Trope, Y. and Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology*, 23(6):445–459.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Tweney, R. D., Doherty, M. E., and Kleiter, G. D. (2010). The pseudodiagnosticity trap: Should participants consider alternative hypotheses? *Thinking & Reasoning*, 16(4):332–345.

Valentino, N. A., Banks, A. J., Hutchings, V. L., and Davis, A. K. (2009). Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility. *Political Psychology*, 30(4).

Villarica, H. (2012). Confirmation bias shapes how we read online. *The Atlantic*.

Vineberg, S. (2011). Dutch book arguments.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, 12(3):129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281.

Wason, P. C. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23(1):63–71.

Webster, D. M. and Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6):1049–1062.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans Use Directed and Random Exploration to Solve the Explore Exploit Dilemma. *Journal of experimental psychology: General*, 143(6):2074–2081.

Wolfe, C. R. and Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, 14(1):1–27.

Wolfers, J. (2014). How Confirmation Bias Can Lead to a Spinning of Wheels. *The New York Times*.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wright, R. (2012). How "confirmation bias" can lead to war. *The Atlantic*.