Cognitive and neural mechanisms underlying post-decision processing

Max Rollwage

Wellcome Centre for Human Neuroimaging &
Max Planck UCL Centre for Computational Psychiatry and Ageing
Research
Institute of Neurology
University College London

Dissertation submitted for the degree of
Doctor of Philosophy
of
University College London

August 2020
Declaration

I, Max Rollwage, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

September 18, 2020
Abstract

Contested issues, such as climate change, can generate polarised and rigid views. A prominent source of entrenched beliefs is confirmation bias, where evidence against one’s position is selectively disregarded. Although an extensive literature has documented this altered processing of new information, the underlying cognitive, computational and neuronal mechanisms remain unknown.

In this thesis, I explore the mechanisms underlying this altered processing of new information, its relation to broader societal attitudes, and finally I test an intervention to alleviate this cognitive bias.

In a first set of studies, I combined human magnetoencephalography (MEG) with behavioural and neural modelling to identify the drivers of altered post-decision evidence integration. I show that high confidence in an initial decision leads to a striking modulation of post-decision neural processing, such that integration of confirmatory evidence is amplified while disconfirmatory evidence processing is abolished. This indicates that confidence shapes a selective neural gating for choice-consistent information, reducing the likelihood of changes of mind.

Confirmation bias has received most attention for its potential contribution to societal polarization and entrenchment. Therefore, in a second set of studies, I tested whether cognitive alterations in post-decision evidence integration are related to broader societal attitudes, such as dogmatic and rigid political beliefs. I found that dogmatic participants showed a reduced sensitivity for disconfirming post-decision evidence (i.e. a stronger confirmation bias) and a reduced tendency to actively seek out corrective information.

In a final study, I tested a metacognitive training procedure as a potential intervention to counteract confirmation bias. This training improved participants’ metacognitive ability and through this boosted their processing of post-decision evidence, both on a behavioural and neural level.

These studies provide a novel mechanistic understanding of confirmation bias, exemplify the potential societal implications of altered post-decision processing and enabled an evidence-based intervention to counteract this cognitive bias.
Impact statement

People often ignore information that disconfirms their beliefs, a phenomenon known as confirmation bias which impairs decision-making in a wide variety of situations. While work from social psychology has repeatedly reported this observation, our understanding of confirmation bias has remained largely descriptive.

In my PhD, I set out to deeply understand the cognitive and neural drivers of confirmation bias. I took a novel and unorthodox approach by investigating well-controlled tasks derived from psychophysics that minimize motivational or social influences on my results. Combining these paradigms with a form of computational modelling known as drift-diffusion modelling enabled me to pin down the drivers of altered evidence processing that contribute to confirmation bias. I then created a novel neuroimaging analysis approach, combining MEG data with computational modelling and machine-learning to precisely decode the timecourse of how new evidence is incorporated by the brain on individual trials. This methodology made it possible to measure alterations in neural information processing that underpin confirmation bias. This neuroimaging analysis might also prove valuable for answering other research questions related to (altered) neural evidence accumulation.

Importantly, the gained insights have enabled me to tailor targeted interventions to alleviate the negative effects of this cognitive bias. I designed a cognitive training that increased participants’ metacognitive abilities and through this reduced confirmation bias. Notably, previous “de-biasing” procedures often failed due to a lack of mechanistic understanding. Because confirmation bias has been reported to impair decision-making in a plethora of situations such intervention may prove highly valuable. For instance, de-biasing is a standard tool in human resources training which is used to improve employees’ decision-making. A more robust and validated cognitive training procedure could have an important impact in this area.

In a second strand of studies, I investigated the mechanisms underlying political radicalization. In these studies, I found that people holding dogmatic and radical political beliefs have a reduced ability to detect their own mistakes, display abolished integration of disconfirming evidence (i.e. a stronger confirmation bias) and show a reduced tendency to seek out corrective information. For decades, the field of political psychology has relied on self-reported cognitive features to model variation in political attitudes. However, many cognitive processes are not consciously accessible/reportable (e.g. people are often unaware of having a confirmation bias). By combining behavioural tasks with computational modelling I have begun to uncover the cognitive mechanisms that contribute to political radicalism. These findings might carry wider societal implications as political polarization is a major challenge of our times. Especially in light of the cognitive training that alleviates the negative consequences of these biases, my research might pave the way to counteract polarization in a wide range of societal issues.
In summary, in my work I have utilised methodology from cognitive and computational neuroscience to investigate cognitive biases, an area previously dominated by social science methodology. This new approach has revealed unique insights into the mechanisms underlying these biases and has showcased the merits of adopting a computational approach. These innovations are especially exciting to me as understanding the mechanism of these biases opens up opportunities for developing evidence-based interventions to reduce or abolish cognitive bias.
Acknowledgements

The last 4 years have been an amazingly fun, challenging and exciting ride. Conducting my PhD at an intellectual hub like the FIL has been a huge privilege. I will miss all the colleagues that provided such a wonderful environment which enabled intellectual discussions in meetings, societal and political debates over lunch and amazingly weird banter at the Queen’s Larder.

I am most grateful for the brilliant guidance from my supervisor Steve Fleming. I could not have asked for any better academic supervision both in terms of direct scientific advice as well as his overarching guidance for answering a set of coherent and exciting research questions. Maybe even more importantly, I am thankful for the way he deeply cares on a personal level. My secondary supervisor Ray Dolan has given me broad guidance towards interesting ideas as well as the freedom to explore these novel avenues.

Moreover, I was extremely lucky to have worked with some amazing students who I have supervised. This work has been most enjoyable as well as productive. Lion Schulz has been an outstanding MSc student who repeatedly impressed me with his intellectual abilities, drive, productivity and broad interests. I am sure he will have a bright future ahead. Alisa Loosen has been great help during her MSc project and has been both a very enthusiastic and clever student as well as a good friend. Without Pippa Watson’s practical help with running my EEG-training study I would not have been able to complete this demanding experiment. Working with each of you taught me different and important lessons, but all of these lessons made my PhD much richer.

The MPC has been a place for radical scientific debate as well as great friendship, and often it was impossible to completely disentangle these two. Tobias Hauser and Rani Moran have been kindly providing constructive criticism and feedback to my ideas and have been important collaborators. Mehdi, Magda, Jolanda and Yuki you have been good friends and it was an important counterpart that we never spoke about work or science. Mehdi, you have been an especially dear friend and I have never met a kinder person. Our discussions about the important things in life and how to achieve those have longlastingly influenced me and I am sure we will have many more of these discussions over the coming years.

Marion and Matan you have become important friends to me, your wisdom, radical openness to other opinions as well as your integer behaviour have been inspiring and I will always see you as role models for this. And of course there is Rylan, I am not sure whether you made my life better or easier, but you definitely made it much more exciting. It is impossible to mention all the fun stories that we experienced together, but I will never forget the good times we had. Moreover, all my office mates on the second floor (especially Mikael) have made the time invaluable.

This thesis would probably not exist without Igor Kagan and Caio Moreira who
supervised my first research internship during my undergraduate studies and they have been collaborators since. They have planted the seed for a passion for science and especially the “sharpen the machado” sessions with Caio made me appreciate the sweet sides of scientific work.

Finally, I am very lucky to have such supportive family and partner. My parents and siblings have always been extremely interested in my work and life here in London. It is so nice to know how proud you are about what I am doing and that I can always rely on you even when we are not in physical proximity. Alice you have been the most important person during the final stages of my PhD. Even though it has been hard to finish everything during lockdown, you have always brightened my days. When I hit my limits and everything seemed overwhelming, I could trust on you to pick me up and be a loving voice of support!
Contributions

The work reported in this thesis is entirely my own unless otherwise indicated. All chapters have been directly supervised by my primary supervisor Dr. Steve Fleming, and have benefited from advice from my secondary supervisor, Prof. Ray Dolan. Especially, Dr. Steve Fleming had an important input to all parts of this thesis.

The experiment reported in Chapter 3 was a result of a collaboration between myself, Alisa Loosen, Dr. Tobias Hauser and Dr. Rani Moran, all from the Max Planck UCL Centre for Computational Psychiatry and Ageing Research. I collected the MEG data and led the project, while Alisa Loosen collected the behavioral data as part of her MSc project, which was co-supervised by Dr. Steve Fleming and myself. Dr. Tobias Hauser and Dr. Rani Moran contributed with suggestions for data analysis and interpretation of results, while Dr. Tobias Hauser also gave input to the study design.

The experiment reported in Chapter 5 was based on Lion Schulz’s MSc project, co-supervised by myself and Dr. Steve Fleming. Lion Schulz implemented the paradigm, carried out the behavioural data collection and analysis, all under close supervision of myself. We jointly wrote the paper on which this chapter is based, whereby Lion took the lead on writing the first draft. Therefore, this chapter is a result of a close collaboration between us.

Chapters 7 was a result of a collaboration between myself and Pippa Watson who assisted me with data collection as part of her undergraduate project, which was co-supervised by myself and Dr. Steve Fleming.
Publications during the PhD

Chapter 3 has been published in *Nature Communications* \(\text{(Rollwage et al., 2020a).}\)

Chapter 4 has been published in *Current Biology* \(\text{(Rollwage et al., 2018).}\)

Chapter 5 is currently in revision for *PNAS*.

Chapter 6 has been submitted to a special issue in *Philosophical Transactions B*, and has been uploaded to *bioRxiv* \(\text{(Rollwage & Fleming, 2020).}\)

Part of the discussion in Chapter 8 is drawn from an article that has been published in *Trends in Cognitive Sciences* \(\text{(Rollwage et al., 2019).}\) In addition to the chapters in this thesis, I collaborated on a project with Madeleine Moses-Payne and Prof. Jonathan Roiser from the Institute for Cognitive Neuroscience which was published in *Frontiers in Psychiatry* \(\text{(Moses-Payne et al., 2019).}\) Moreover, I maintained a collaboration with Dr. Igor Kagan and Dr. Arezoo Pooresmaeili from the German Primate Center and European Neuroscience Institute in Goettingen/Germany, and Prof. Gerhard Stemmler from the University of Marburg/Germany, resulting in three papers based on work carried out prior to my PhD \(\text{(Moreira et al., 2018, Rollwage et al., 2020b, 2017).}\)
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration</td>
<td>2</td>
</tr>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Impact statement</td>
<td>4</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>Contributions</td>
<td>8</td>
</tr>
<tr>
<td>Publications during the PhD</td>
<td>9</td>
</tr>
<tr>
<td>Contents</td>
<td>10</td>
</tr>
<tr>
<td>List of Figures</td>
<td>15</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>17</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>18</td>
</tr>
<tr>
<td>1.1 Conceptual overview</td>
<td>18</td>
</tr>
<tr>
<td>1.2 Overarching principles</td>
<td>19</td>
</tr>
<tr>
<td>1.2.1 Confirmation bias</td>
<td>19</td>
</tr>
<tr>
<td>1.2.2 Bayesian models of decision-making and belief updating</td>
<td>21</td>
</tr>
<tr>
<td>1.2.3 Perceptual decision making</td>
<td>22</td>
</tr>
<tr>
<td><strong>2 Literature review</strong></td>
<td>24</td>
</tr>
<tr>
<td>2.1 Overview</td>
<td>24</td>
</tr>
<tr>
<td>2.2 Statistical decision-making models</td>
<td>24</td>
</tr>
<tr>
<td>2.2.1 Bayesian inference</td>
<td>24</td>
</tr>
<tr>
<td>2.2.2 Sequential sampling models</td>
<td>26</td>
</tr>
<tr>
<td>2.2.3 Extension of sequential sampling models to post-decision integration</td>
<td>27</td>
</tr>
<tr>
<td>2.3 Decision confidence</td>
<td>29</td>
</tr>
<tr>
<td>2.3.1 Confidence reflects decision-evidence</td>
<td>29</td>
</tr>
<tr>
<td>2.3.2 Metacognition</td>
<td>30</td>
</tr>
</tbody>
</table>
2.3.3 Metacognitive monitoring .......................... 30
2.3.4 Metacognitive control and the influence of confidence on post-decision processing .................. 32

2.4 Neural basis of perceptual decision-making ................ 33
2.4.1 Non-human primates .............................. 33
2.4.2 Humans ........................................ 34
2.4.3 Neural mechanisms of post-decision evidence integration ...... 35
2.4.4 Multivariate decoding enables the investigation of neural information processing .................. 36

2.5 Open questions ..................................... 38

3 Confidence drives a neural confirmation bias 39
3.1 Introduction ........................................ 39
3.2 Methods .......................................... 40
3.2.1 Participants .................................... 40
3.2.2 Stimuli and experimental design .................... 41
3.2.3 Task procedure .................................. 42
3.2.3.1 Design alterations in behavioural study 2 .......... 43
3.2.3.2 Design alterations in MEG study 3 ................ 43
3.2.4 Scoring and bonus payment ....................... 44
3.2.5 Drift-diffusion modelling ......................... 44
3.2.6 MEG pre-processing ............................ 46
3.2.7 Generalising a pre-decision classifier to the post-decision phase ................................. 46
3.2.8 MEG topography contributing to classification accuracy .... 48
3.2.9 MEG temporal generalization ..................... 48

3.3 Results ........................................... 49
3.3.1 Adaptive use of post-decision evidence ................ 49
3.3.2 Effects of confidence on changes of mind .............. 49
3.3.3 Confidence induces a selective gain for confirmatory evidence .... 50
3.3.4 Neural markers of post-decisional processing ......... 53

3.4 Discussion ....................................... 57

4 Metacognitive failure as a feature of those holding radical beliefs 61
4.1 Introduction ...................................... 61
4.2 Methods .......................................... 63
4.2.1 Participants .................................... 63
4.2.2 Stimuli ........................................ 64
4.2.3 Task and procedure ............................ 64
4.2.3.1 Calibration Phase ......................... 64
4.2.3.2 Confidence task (Task 1) ................. 65
4.2.3.3 Post-decision evidence integration task (Task 2) ... 65
Contents

4.2.4 Data quality and exclusion criteria ........................................... 66
4.2.5 Behavioral Analysis ................................................................. 67
  4.2.5.1 Measurement of metacognitive ability ................................ 67
  4.2.5.2 Measurement of post-decision evidence integration ............ 67
4.2.6 Factor analysis ................................................................. 68
4.2.7 Computational modelling ...................................................... 69
  4.2.7.1 Modelling post-decision evidence integration .................... 69
  4.2.7.2 Model extensions accounting for differences in post-decision evidence integration ............... 71
  4.2.7.3 Model fitting ................................................................. 72
  4.2.7.4 Model comparison ......................................................... 72
  4.2.7.5 Model simulations ......................................................... 72
4.2.8 Statistical analysis ............................................................. 73
4.3 Results .............................................................................. 74
  4.3.1 Factor analytic description of political beliefs ....................... 74
  4.3.2 Radical beliefs are related to reduced metacognitive ability .... 76
  4.3.3 Metacognitive ability predicts post-decision evidence integration 78
  4.3.4 Radical beliefs are characterized by reduced integration of disconfirming post-decision evidence ............... 78
  4.3.5 Computational mechanism underlying radical beliefs .......... 80
4.4 Discussion ....................................................................... 82

5 Dogmatism manifests in lowered information search under uncertainty 85
  5.1 Introduction ................................................................... 85
  5.2 Methods ........................................................................ 87
    5.2.1 Participants ................................................................. 87
    5.2.2 Factor Analysis ............................................................. 88
    5.2.3 Experimental design .................................................... 88
      5.2.3.1 Stimuli ................................................................. 88
      5.2.3.2 Task and procedure ............................................... 88
      5.2.3.3 Information-seeking task ........................................ 88
    5.2.4 Statistical Analysis ...................................................... 90
      5.2.4.1 Task behavior ......................................................... 90
      5.2.4.2 Factor scores and behavioral variables ..................... 90
      5.2.4.3 Computational modeling ......................................... 91
  5.3 Results ........................................................................ 92
    5.3.1 Measuring Dogmatism .................................................. 92
    5.3.2 Interrelation between factor scores ................................ 93
    5.3.3 Measuring Information Search ....................................... 94
    5.3.4 Information Search and Dogmatism ................................... 94
8 General discussion

8.1 Overview .................................................. 137

8.2 Confidence induces biased post-decision information processing instead of Bayes optimal weighting ................................. 138

8.3 Mechanisms underlying confirmation bias ........................................ 139

8.4 Altered information processing and dogmatic beliefs ............................. 140

8.5 Metacognitive training enhances the sensitivity for new information ...... 142

8.6 Conclusion ........................................................................ 144

Bibliography ............................................................................ 145
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Extension of a classical DDM to model post-decision evidence accumulation</td>
</tr>
<tr>
<td>2.2</td>
<td>Illustration of the theoretical distinction between metacognitive sensitivity and confidence bias</td>
</tr>
<tr>
<td>2.3</td>
<td>Illustration of univariate versus multivariate neuroimaging analysis</td>
</tr>
<tr>
<td>3.1</td>
<td>Task timeline of the post-decision evidence task</td>
</tr>
<tr>
<td>3.2</td>
<td>Influence of post-decision evidence on the final decision</td>
</tr>
<tr>
<td>3.3</td>
<td>Influence of confidence on post-decision evidence integration</td>
</tr>
<tr>
<td>3.4</td>
<td>Drift-diffusion modelling fits to the second decision</td>
</tr>
<tr>
<td>3.5</td>
<td>Outline of MEG analysis for quantifying accumulation of post-decision evidence at a neural level</td>
</tr>
<tr>
<td>3.6</td>
<td>MEG analysis investigating the influence of confidence on post-decision evidence processing</td>
</tr>
<tr>
<td>4.1</td>
<td>The left and right extremes of the political spectrum are associated with intolerant and dogmatic world views</td>
</tr>
<tr>
<td>4.2</td>
<td>Behavioral tasks measuring metacognitive sensitivity and post-decision evidence integration</td>
</tr>
<tr>
<td>4.3</td>
<td>Impaired metacognitive sensitivity and reduced disconfirmatory evidence integration predict facets of radicalism</td>
</tr>
<tr>
<td>4.4</td>
<td>Individual differences in radicalism are captured by a choice bias model</td>
</tr>
<tr>
<td>5.1</td>
<td>Information seeking task</td>
</tr>
<tr>
<td>5.2</td>
<td>Liberal and conservative extremes of the political spectrum predict levels of dogmatism and belief superiority</td>
</tr>
<tr>
<td>5.3</td>
<td>Participants adaptively seek additional information</td>
</tr>
<tr>
<td>5.4</td>
<td>Dogmatism is characterized by a reduction in information search, leading to less veridical judgements.</td>
</tr>
<tr>
<td>5.5</td>
<td>Individual differences in confidence-driven information search as captured by a computational model</td>
</tr>
<tr>
<td>6.1</td>
<td>Comparison of agents’ performance with different biases in information processing</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>6.2</td>
<td>Performance of a confidence-weighted confirmation bias as function of metacognitive ability.</td>
</tr>
<tr>
<td>7.1</td>
<td>Training hypothesis and task design</td>
</tr>
<tr>
<td>7.2</td>
<td>Behavioural training effects</td>
</tr>
<tr>
<td>7.3</td>
<td>Training effects on neural post-decision evidence processing</td>
</tr>
<tr>
<td>7.4</td>
<td>Association of metacognitive ability and belief updating in real-world beliefs.</td>
</tr>
</tbody>
</table>
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>BOLD</td>
<td>blood-oxygen-level-dependent signal</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CPP</td>
<td>centroparietal positivity</td>
</tr>
<tr>
<td>DDM</td>
<td>Drift-diffusion model</td>
</tr>
<tr>
<td>DIC</td>
<td>Decision information criterion</td>
</tr>
<tr>
<td>DV</td>
<td>Decision variable</td>
</tr>
<tr>
<td>EEG</td>
<td>electroencephalography</td>
</tr>
<tr>
<td>ERP</td>
<td>evoked response potentials</td>
</tr>
<tr>
<td>fMRI</td>
<td>function magnet resonance imaging</td>
</tr>
<tr>
<td>GLM</td>
<td>generalized linear model</td>
</tr>
<tr>
<td>LIP</td>
<td>lateral intraparietal area</td>
</tr>
<tr>
<td>MEG</td>
<td>magnetoencephalography</td>
</tr>
<tr>
<td>RDK</td>
<td>Random dot motion kinetogram</td>
</tr>
<tr>
<td>SEM</td>
<td>standard error fo the mean</td>
</tr>
<tr>
<td>SVM</td>
<td>support-vector machine</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Conceptual overview

A hallmark of cognitive flexibility is the capacity to change one’s mind in light of new information. In order to revise previous beliefs, it is crucial to constantly incorporate new evidence, even after having committed to a decision. The importance of this capacity is clearly exemplified by the Covid-19 pandemic. Scientific knowledge on this issue is quickly evolving and both policymakers as well as citizens have to consider these new facts in order to make the right decisions. This might mean that previously held beliefs need to be corrected as new information becomes available. As an example, while the UK government initially did not recommend to wear face-masks in public, new evidence became available (Chu et al., 2020), which changed the governmental recommendations¹ making it essential for citizens to adapt their beliefs (and accompanied behaviours) in order to follow the recommended guidelines.

Although this kind of post-decision evidence integration is clearly important for the formation of accurate beliefs, it has been repeatedly reported that humans show biases in the processing of new information. Specifically, people process new information tainted by their existing beliefs, favouring information that supports their prior views. This selective information consideration is known as confirmation bias (Nickerson, 1998).

Most research on this topic has been conducted in the area of social psychology (Klayman, 1995; Nickerson, 1998). While this work has led to a number of studies reporting this phenomenon, our understanding of confirmation bias has remained largely descriptive. One reason might be that cognitive biases were often studied in scenarios involving complex real-world beliefs such as political attitudes (Nyhan & Reifler, 2010; Taber et al., 2009; Taber & Lodge, 2012). The complexity of such higher-order beliefs makes it difficult to disentangle the various contributing factors.

In this thesis, I set out to understand the cognitive and neural drivers of altered post-decision evidence processing. Specifically, I will focus on a role of confidence

¹see government guidelines from the 6 June 2020
on post-decision processing and investigate how people’s confidence alignment (i.e. metacognitive ability) might be a crucial determinant for the veridical integration of new information.

In the first part of the thesis (chapter 3), I examine the computational and neural mechanisms underlying biased evidence incorporation. Hereby, I combined computational modelling and magnetoencephalography (MEG) recordings, showing that a selective neural gating for choice-consistent information drives a behavioural confirmation bias. Moreover, I identified a special relevance of people’s confidence for the processing of new information, as a neural confirmation bias was enhanced after high confidence decisions.

Since it has been argued that selective information intake might play a crucial role for societal polarization and entrenchment of beliefs (del Vicario et al., 2017; Lilienfeld et al., 2009), in the second part of this thesis (chapter 4 and 5) I investigated the relation of this cognitive bias with broader societal issues such as radical and dogmatic (political) beliefs. Here I show, that individual differences in the proneness for biased information processing are related to rigid and entrenched societal beliefs, indicating that these alterations in information processing bear relevance for real-world attitudes.

In the final part of this thesis (chapter 6 & 7), I use these mechanistic insights, gained from the earlier chapters, to test potential strategies for alleviating the negative behavioural consequences of selective post-decision evidence accumulation. In chapter 6, I use simulation-based modelling to theoretically show that people’s metacognitive ability is a crucial factor determining the negative consequences of selective information processing. In the final chapter 7, I test a metacognitive training procedure as potential intervention to counteract confirmation bias and increase participants sensitivity for new information. I confirmed that the cognitive training increased participants’ metacognitive ability and through this boosted their post-decision evidence processing.

1.2 Overarching principles

1.2.1 Confirmation bias

Confirmation bias has been reported in psychological research for decades and has been shown in a vast of studies (Klayman 1995, Nickerson 1998, Wason 1960). This selective information intake has been reported in a variety of tasks, including the formation of clinical diagnosis (Groopman 2008), inference about people’s character (Snyder & Swann 1978), investment decisions (Park et al. 2010), witness reports in court (Kassin et al. 2013), views about societal issues such as capital punishment (Lord et al. 1979) and climate change (Sunstein et al. 2016), and even scientific investigations themselves (Koehler 1993, Mynatt et al. 1977). Maybe
most prominently, confirmation bias has been reported with respect to politically charged beliefs, as people are generally prone to process information tainted by their political convictions (Kaplan et al., 2016; Nyhan & Reifler, 2010; Redlawsk, 2002; Taber et al., 2009; Taber & Lodge, 2012).

All these studies robustly find that people are unlikely to revise their initial views in light of new evidence. However, the cognitive processes that drive this resistance for new information might differ widely between studies, whereby at least two generally different mechanisms can be distinguished (Klayman, 1995): a) active search for confirming evidence (Klayman & Ha, 1987; Wason, 1960; Sharot & Sunstein, 2020); b) reluctance to incorporate disconfirming information (Lord et al., 1979). Even this distinction is very crude and likely to include different cognitive processes, as for instance the reluctance to integrate disconfirming information could be driven by reduced credibility assignment to contradictory information (Lord et al., 1979) or a differential sensitivity for confirming and disconfirming information (Hendry et al., 1989). In most cases, the underlying cognitive processes driving confirmation bias have not been assessed and manipulated precisely enough to disentangle these different mechanisms. Nevertheless, a few recent studies from the field of cognitive neuroscience have identified a selective gain for choice-consistent information as one of the major mechanisms driving confirmation bias (Cheadle et al., 2014; Palminteri et al., 2017; Talluri et al., 2018, 2020). Therefore, in this thesis I will mainly focus on the integration of confirming and disconfirming information as measure of confirmation bias (in chapter 3, 4, 6 & 7), whereas in chapter 5, I will specifically investigated the active search for new information.

Besides the different cognitive processes that potentially contribute to the behavioural manifestation of confirmation bias, there has been debate about the cognitive or motivational nature of selective information intake. In a nutshell, the questions is whether people want to receive information that confirms their beliefs (Kunda, 1990; Redlawsk, 2002) or whether a cognitive limitation is driving this selective information intake (Doherty & Mynatt, 1986). It is therefore important to dissociate confirmation bias from a desirability bias (Sharot, 2011; Sharot et al., 2007; Weinstein, 1989). Confirmation bias describes the selective processing of information that confirms one’s existing beliefs, whereas desirability bias describes the preferential processing of positively valenced information. In everyday life, the confirmation of an existing belief is also often desirable (e.g. a stock trader will only make money if her previous investment decisions are confirmed by the market moving in the predicted direction) and thus a confirmation bias will often co-occur with a desirability bias. However, these two processes describe subtly different mechanism and they can be disentangled in scientific studies (Sunstein et al., 2016; Tappin et al., 2017).

Here I focus on confirmation bias, and more specifically the cognitive aspects that
drive confirmation bias. Therefore, I aimed to minimize the influence of motivational factors by using simple perceptual-decision making tasks (for a justification of this choice see section 1.2.3). While motivational factors surely play an important role for confirmation bias (Nickerson [1998] Sunstein et al. [2016]), this additional complexity might makes it hard to isolate and probe specific contributions of cognitive mechanisms. Therefore, I aimed to eliminate these additional factors.

1.2.2 Bayesian models of decision-making and belief updating

Confirmation bias directly includes the word “bias”, and thus carries the notion of maladaptive behaviour. While it seems intuitive that a selective consideration of evidence should lead to erroneous behaviour and detrimental task performance, this view has been challenged, as selective information integration might be adaptive under some circumstances (Klayman & Ha [1987]). Crucially, in order to identify whether a specific behaviour has negative or positive consequences, one needs to know what the optimal behaviour (and information processing strategy) in a scenario would be. Thus, a benchmark strategy is needed, against which other behavioural or information processing strategies can be evaluated. Bayesian models of decision-making and belief-updating provide a normative framework of how evidence should inform decisions (Green et al. [1966]), and moreover how existing knowledge should be combined with new information in order to update beliefs (Mathys et al. [2011]; O’Reilly et al. [2013]; Meyniel et al. [2015a]).

The general assumptions of these models is that decisions are made by arbitrating the evidence in favour for each alternative. The alternative that is supported by most evidence will be chosen (Beck et al. [2008]; Gold & Shadlen [2002, 2007]; Knill & Pouget [2004]). This evaluation of evidence in favour of each choice option directly carries a notion of uncertainty, giving a graded measure of the confidence with which a decision can be made (Fleming & Daw [2017]; Kepecs & Mainen [2012]; Pouget et al. [2016]). When combining new information with existing knowledge, this notion of certainty/confidence becomes crucial as a belief that is held with higher certainty (i.e. a belief that received much support in the past) requires more evidence in order to be revised (Meyniel et al. [2015a]; Meyniel & Dehaene [2017]). Bayesian belief updating exactly describes the degree to which a belief should be revised based on the new information (see a mathematical formulation in section 2.2.1), whereby a prior belief and new information are combined weighted by their respective certainty.

Importantly, in order to judge whether participants behave in line with principles of Bayesian decision-making and belief updating, we as experimenters need to know how strong the supporting evidence for either alternative was. However, the exact quantification of evidence strength is a non-trivial task in most situations, po-
tentially explaining why it has been challenging for many studies to explicitly test whether a confirmation bias leads to deviations from optimality (Klayman, 1995; Klayman & Ha, 1987). Therefore, one prerequisite for judging optimality (or deviation from optimality) in decision-making and information processing is that we as experimenters are able to model a normative belief updating strategy.

1.2.3 Perceptual decision making

Perceptual decision-making tasks allow tight control over the presented stimuli, and thus the evidence which is received by participants (Gold & Shadlen, 2007). For instance, a random dot motion kinetogram (RDK) is often used in perceptual decision-making (Fleming et al., 2018; Gold & Shadlen, 2007; Kiani & Shadlen, 2009), and consists of a cloud of dots that either move to the left or the right-side of the screen while participants are tasked to indicate the motion direction. Most of the dots move randomly while a subset of dots is coherently moving in the target direction. This subset of coherently moving dots can be experimentally controlled, making it easy to precisely quantify the evidence strength. Moreover, this precise control over the presented evidence makes it trivial to increase or decrease the stimulus strength. As participants might differ in their perceptual sensitivity (Green et al., 1966), an adjustment of difficulty enables us to equate the internal/subjective evidence strength for each participant by matching their performance (García-Pérez, 1998; Levitt, 1971). Such individual adjustment of task difficulty is not easily achieved with other forms of information (e.g. verbal statements about societal issues) that are widely used for the study of confirmation bias (Lord et al., 1979). Since the evidence strength is known and quantifiable (in form of a single number) for perceptual stimuli, we can make precise predictions about how much this information should influence participants’ beliefs, enabling us to potentially detect divergences from normative evidence integration.

In contrast, many studies have investigated biases in information processing with respect to real-world attitudes such as political beliefs (Nyhan & Reifler, 2010; Taber et al., 2009; Taber & Lodge, 2012) and societal issues (Lord et al., 1979; Sunstein et al., 2016). While this approach has the appeal of high ecological validity, these situations are influenced by many factors. For instance, our political beliefs might have a tight link to our self-concept (Federico & Ekstrom, 2018) and are strongly influenced by our social network (Cohen, 2003; Marks et al., 2019; Van Bavel & Pereira, 2018). Investigating information processing in these kind of issues makes it likely that participants are influenced by a multitude of factors. While a mixture of these factors probably influences confirmation bias in real-world settings, this complexity makes it hard to investigate and isolate the contribution of specific cognitive processes.

Perceptual tasks have the advantage of removing these potentially confounding
influences. Participants usually have no prior experiences with the perceptual tasks used in our laboratories and they do not have any prior investments in their decision. For instance, a participant is unlikely to care about whether a cloud of dots is moving to the left or to the right side of the screen. On the other hand, the participant likely cares about whether her political convictions are supported by evidence or not.

Moreover, in perceptual decision-making tasks a ground truth exists (e.g. the dots are moving to left or the right) and is accessible the experimenter. Therefore, we can quantify the accuracy of a belief and pay participants based on their performance. This should further incentivize participants to strive for accurate decisions and minimize other influencing factors. In comparison, in the real-world, people might not always have the goal of forming the most accurate view as other factors might be more important (e.g. securing social status or social belonging) which could make it rational to not change one’s mind in light of evidence [Van Bavel & Pereira, 2018]. Using simple tasks removes a layer of complexity and makes motivational influences less likely.

Finally, the neural mechanisms involved in decision formation are relatively well understood for perceptual tasks [Albright, 1993; Britten & van Wezel, 1998; Donner et al., 2009; Gold & Shadlen, 2007; Kelly & O’Connell, 2015], making it possible to investigate the neural mechanisms associated with (altered) post-decision evidence processing.

Based on these advantages, I used perceptual decision-making tasks as a model system to investigate post-decision evidence processing throughout my thesis. However, due to the simple nature of these tasks, people might question the generalizability to broader societal issues. Thus, I specifically investigate in chapter 4 and 5 whether evidence processing in these simple perceptual tasks generalizes to broader societal issues.
Chapter 2

Literature review

2.1 Overview

This thesis investigates (altered) post-decision evidence integration, using perceptual decision-making tasks. In this research area, a plethora of theoretical frameworks exist for modelling evidence accumulation. In this chapter, I will review the existing literature. First, I will introduce the formalism of Bayesian inference which can be used as normative model of information usage and decision-making. Secondly, I will focus on sequential sampling models \cite{Ratcliff1998}, that have successfully been used to study evidence accumulation in simple decisions such as perceptual tasks. Further, I will review the extension of these models to study post-decision evidence accumulation.

Since this thesis is specifically concerned about the influence of confidence on post-decision evidence processing, in the second part of this chapter, I will review models of subjective confidence. A specific emphasis will be on the link between confidence and decision evidence. Afterwards, I will review existing literature on the accuracy of people’s confidence judgments, a self-reflective process known as metacognition. As confidence has been suggested as an internal control signal, I will further focus on existing literature suggesting a role of confidence in the processing of new information.

In the final part of this literature review, I will present existing findings regarding the neural mechanism underlying the accumulation of pre- and post-decision evidence.

2.2 Statistical decision-making models

2.2.1 Bayesian inference

Our brains are safely hidden in our skulls, only having access to the outside world through our senses. Therefore, the brain has the challenging tasks to constantly
infer what is happening in the environment based on a stream of noisy incoming information (Dayan et al. 1995, Friston 2010). In order to do this, our brains have to select one hypothesis about the (unknown) state of the world out of a set of competing alternatives, based on the information provided from the sensory system (Gold & Shadlen 2007). This process is far from trivial, especially since the available information is often imperfect at any given moment in time (Beck et al. 2008, Knill & Pouget 2004). It has been argued that the best way of mastering this task is to combine prior knowledge with incoming evidence in order to infer the state of the world (Dayan et al. 1995, Doya et al. 2007, Friston 2010). Bayes theorem provides the statistical rules with which an optimal combination can be achieved, providing a normative model of evidence usage (MacKay & MacKay 2003). The core idea is that the probability of a state of the world given the data (i.e. the posterior probability of a hypothesis, which is the quantity the brain is trying to infer) can be inferred from the likelihood of the data under this hypothesis and the prior probability of the hypothesis. Bayes theorem provides an equation describing how to derive these conditional probabilities from each other:

$$P(hypothesis|data) = \frac{P(data|hypothesis) \times P(hypothesis)}{P(data)} \quad (2.1)$$

Hereby, $P(hypothesis)$ represents the prior probability of the hypothesis and $P(data)$ represents the probability of the data combined over all potential hypotheses. In a simple situation where participants do not have any prior experience with a task, the equation reduces to:

$$P(hypothesis|data) = \frac{P(data|hypothesis)}{P(data)} \quad (2.2)$$

$$= \frac{P(data|hypothesis)}{P(data|hypothesis) + P(data|all alternative hypotheses)} \quad (2.3)$$

This equation simply means that the evidence in favour for one hypothesis should be compared to the evidence for all alternative hypotheses, whereby the most favoured alternative should be chosen. Importantly, this equation directly quantifies how certain one should be in a hypothesis given the data (see also section 2.3.1).

However, in this thesis I am mainly interested in a situation of continued evidence integration. Hereby, prior knowledge plays a key role and Bayes theorem describes how the new information ($P(data \mid hypothesis)$) should be weighted with the prior knowledge ($P(hypothesis)$). Since the probability of the data (given all possible hypothesis) is invariant to the evaluated hypothesis, the denominator becomes a constant, leading the posterior belief to be only driven by the prior and the new information:

$$Posterior \propto prior \times likelihood \quad (2.4)$$
Importantly, this updating can be performed iteratively whenever new information becomes available. In this case, the posterior belief of the current inference step becomes the prior of the next inference step. Bayes theorem therefore provides a mathematical framework for how to infer causes in the world given sensory information and also provides a way of combining this sensory evidence with existing knowledge about the world (Mathys et al., 2011; O’Reilly et al., 2013).

2.2.2 Sequential sampling models

Since sensory information is often uncertain, a continued accumulation of evidence over time can increase the signal to noise ratio and help to make better decisions. Sequential sampling models follow this intuition and assume that our brains accumulate evidence over time (Gold & Shadlen, 2001, 2007; Ratcliff & Rouder, 1998). This class of models is a dynamic extension of the statistical decision-making models explained in the previous section (Gold & Shadlen, 2002; Wald & Wolfowitz, 1948). Sequential sampling models take the evolution of evidence into account and model the dynamics that lead up to a decision. Therefore, these models do not only account for choices, but can also explain reaction times. The assumption is that at every single time step a noisy information sample \( x \) is drawn from a distribution (whereby the mean of this distribution compared to its variance indicates the signal strength for this hypothesis) and these noisy samples are summed over time to form the evidence in favour for each hypothesis. Once enough evidence in favour for one of the options is accumulated a decision threshold is hit and a decision is initiated.

Given a simplified situation where only two options are available (\( h_1 \) and \( h_2 \)), a single variable can track the balance of evidence for both options and determine the decision processes. In such situation, a decision variable (DV) can track the balance of evidence in favour for each choice-option (\( h_1 \) and \( h_2 \)) by calculating a (logarithmic) likelihood-ratio (log-odds):

\[
DV_{1,2|x} = \log \frac{P(x_1, x_2, \ldots, x_n|h_1)}{P(x_1, x_2, \ldots, x_n|h_2)}
\] (2.5)

\[
= \sum_{i=1}^{n} \log \frac{P(x_i|h_1)}{P(x_i|h_2)}
\] (2.6)

This decision variable increases or decreases in response to additional evidence in favour for \( h_1 \) or \( h_2 \). The logarithmic transformation allows additivity of evidence samples over time (instead of a multiplication). This sequential probability ratio test (Wald & Wolfowitz, 1948) is a dynamic extension of Bayesian decision models in the sense that instead of having one evidence sample that determines the likelihood, there will multiple evidence samples integrated in this likelihood.

However, for calculating the exact probability ratio the distribution of \( x \)’s needs to be known, which might not always be the case. The drift-diffusion model (DDM)
Chapter 2. Literature review

is a popular version of sequential sampling models (Ratcliff & Rouder, 1998), which does not rely on the knowledge about the underlying distribution of $x$’s. In this sense the DDM is a non-Bayesian process model that however approximates the sequential probability ratio test under some conditions (Bogacz et al., 2006), making it a practically useful tool for fitting reaction time and choice data (Ratcliff & Rouder, 1998). Moreover, the DDM disentangles starting point and drift-rate of an evidence accumulation processes. The starting point describes the support for a hypothesis at the beginning of the accumulation process. This support for a hypothesis before evidence is integrated is equivalent to a prior in Bayesian terms (Gold & Shadlen, 2007). On the other hand, the drift-rate corresponds to the average rate with which new information is accumulated and thus represents the processing of new information in a narrower sense. Therefore, the DDM can help to separate prior evidence from the rate with which new evidence is integrated over time.

2.2.3 Extension of sequential sampling models to post-decision integration

Sequential sampling models were classically used to model optimal stopping criteria that determine when enough evidence for a hypothesis exists and no further evidence is needed (Gold & Shadlen, 2001, 2002). Therefore, it was often assumed that crossing a decision bound would end the evidence accumulation process. However, empirical findings have challenged this view and suggest that the ongoing evidence accumulation, after a decision was made, might account for the occurrence of spontaneous changes of mind (Murphy et al., 2015; Resulaj et al., 2009; Van Den Berg et al., 2016). Hereby, it is assumed that the DV continues to track evidence for both choice options and if the post-decision evidence favours the unchosen option sufficiently, participants will change their mind (Yeung & Summerfield, 2012). While these studies mainly focussed on continued accumulation of evidence that is still “in the pipeline”, an extension of these sequential sampling models to situations where new information is explicitly presented might proof fruitful for studying post-decision evidence processing.

In a situation where participants are presented with post-decision evidence evidence after having made an initial decision, the end point of the pre-decision accumulation process should be the starting point of the post-decision process (see Figure 2.1). As evidence has already been accumulated during the pre-decision phase this should be incorporated in the final decision, a processes being equivalent to incorporating prior knowledge when deriving a posterior belief in Bayesian inference. The way to do this is by simply adding the log-odds of the pre- ($x_{pre}$) and post-decision ($x_{post}$) evidence:

$$DV_{1,2|x_{pre},x_{post}} = \log \frac{P(x_{pre}|h_1)}{P(x_{pre}|h_2)} + \log \frac{P(x_{post}|h_1)}{P(x_{post}|h_2)} \tag{2.7}$$
It has been shown that participants incorporate pre- and post-decision evidence in this way, as predicted by Bayesian principles of evidence integration \cite{Fleming2018}.

**Figure 2.1:** Extension of a classical DDM to model post-decision evidence accumulation. A DV tracks the balance of evidence for choice A versus choice B. The state of the DV at the time of the initial decision determines which option is chosen. However, after this initial decision, more evidence is presented (post-decision evidence). The final decision should incorporate both the pre- and post-decision evidence. In order to combine pre- and post-decision evidence, the post-decision accumulation process should start at the point where the pre-decision accumulation process ended. If the balance of evidence favours one of the choices more strongly, this decision will be made with more confidence (as illustrated with different grey shadings). Confirming post-decision evidence will lead to an increase of confidence in the decision, whereas disconfirming post-decision evidence (indicated by the red line) might lead to a change of mind if this post-decision evidence is sufficiently favouring the previously unchosen option.

Importantly, an optimal evidence integration strategy should take the accumulated evidence from the pre-decision phase into account when forming a final decision, however the initial decision (and initial belief) should not alter the sensitivity for new information. In terms of the DDM, this would mean the evidence accumulated during the initial phase should determine the starting point of the post-decision accumulation (see Figure 2.1), but the pre-decision evidence should not influence the post-decisional drift-rate.

However, empirical evidence has shown that making a decision can alter the way new information is processed, reducing the sensitivity for new evidence \cite{Bronfman2015}. Talluri et al. \cite{2018} did not only find an overall reduced sensitivity for new information after choice-commitment, but found a selective bias in post-decision evidence incorporation. While sensitivity for disconfirming evidence was reduced, a selective gain for confirming evidence was observed. These results show that the
act of making a decision alters the processing of new information and induces a confirmation bias. This further indicates that extending models of evidence accumulations to the post-decision period might provide a mechanistic understanding regarding the cognitive processes driving confirmation bias.

2.3 Decision confidence

2.3.1 Confidence reflects decision-evidence

Decisions are usually accompanied by a graded feeling of confidence. Such subjective confidence is often formalized as the probability of having made the correct decision (Fleming & Lau, 2014; Kepecs & Mainen, 2012).

Bayesian decision-making models directly quantify the degree of confidence that should accompany a decision, as the evidence in favour for the chosen option directly translates into confidence (Meyniel et al., 2015b; Pouget et al., 2016):

\[ P(\text{correct}) = \frac{P(\text{choice} | \text{data})}{P(\text{all alternatives} | \text{data})} \]  

In this formulation, confidence and decision evidence are equivalent. This idea is supported by findings from non-primate single-cell recordings, which show that the same neurons in the lateral intraparietal area (LIP) track both decision-evidence as well as confidence (Kiani & Shadlen, 2009).

In the same spirit, in the framework of the DDM it has been argued the DV determines both the choice as well as the accompanying confidence (Van Den Berg et al., 2016). However, since an optimal stopping rule should determine the amount of evidence that ends the accumulation process (Wald & Wolfowitz, 1948), each choice that is made based on a boundary crossing should result in the same confidence judgment (Moran et al., 2015; Pleskac & Busemeyer, 2010). Therefore, extensions of existing sequential sampling models have been suggested in order to explain trial-by-trial variability of confidence ratings. For instance, it has been suggested that evidence which was still “in the pipeline” keeps accumulating after the decision, but before the confidence rating, explaining variability of confidence judgments (Moran et al., 2015; Pleskac & Busemeyer, 2010; Yu et al., 2015; Murphy et al., 2015). Other researchers have suggested that confidence represents a noise-corrupted read-out of evidence (Calder-Travis et al., 2020), that partially different aspects of the information might be used for decision and confidence reports (Zylberberg et al., 2012) or that confidence judgments incorporate additional information on top of the decision-evidence (Fleming & Daw, 2017; Kiani et al., 2014). While the debate about the exact information incorporated in confidence ratings is still on-going, it seems warranted to assume that confidence reflects decision-evidence to a large degree (see Figure 2.1), but that both quantities can also be separated as confidence might rely on partially different information (see next section).
2.3.2 Metacognition

More evidence in favour of the chosen option will increase the likelihood of a correct decision and should also be accompanied by higher confidence (Kepecs & Mainen, 2012). Thus, confidence can be used as indicator for choice accuracy, and it has been argued that it is useful to compute this quantity in order to monitor and guide behaviour (Meyniel et al., 2015b). Treating confidence as a monitoring or control signal of other cognitive processes (i.e. the decision formation), makes it a metacognitive processes. Metacognition is defined as “cognition about cognition” or “thinking about thinking” (Fleming et al., 2012; Koriat, 2007). Therefore, confidence and its relation to task performance has been at the core of the scientific study of metacognition (Fleming et al., 2012).

Metacognition can be divided into two main processes, metacognitive monitoring and metacognitive control (Nelson & Narens, 1990). Metacognitive monitoring describes the tracking and monitoring of other cognitive processes, whereas metacognitive control describes the usage of this monitoring signal in order to guide future behaviour. In terms of confidence, metacognitive monitoring translates into the accuracy with which confidence tracks task performance (Fleming & Lau, 2014), whereas metacognitive control refers to the usage of this confidence signal in order to regulate or guide future behaviour (Desender et al., 2018; Risko & Gilbert, 2016).

2.3.3 Metacognitive monitoring

Metacognitive monitoring has received the majority of scientific attention, with research investigating how to measure the quality with which confidence tracks choice accuracy (Fleming & Lau, 2014; Maniscalco & Lau, 2012), the underlying neural mechanisms enabling metacognitive monitoring (Fleming et al., 2010; McCurdy et al., 2013; Roumis et al., 2010) and how individual differences in metacognitive monitoring relate to real-world issues such as mental-health symptoms (Rouault et al., 2018).

Metacognitive ability measure the correlation between choice accuracy and confidence. The stronger the correlation, the better confidence indicates choice accuracy and the more useful it is as monitoring signal (Fleming & Lau, 2014). High metacognitive ability is therefore characterized by strong trial-by-trial correspondence between confidence ratings and choice accuracy.

Two aspects of this correspondence can be distinguished: bias (or calibration) and sensitivity (or resolution) (Baranski & Petrusic, 1994). Bias describes the correspondence between the confidence rating and the actual probability of being correct, which captures a general tendency to be over- or underconfident (Moore & Healy, 2008). Sensitivity, on the other hand, describes the degree to which confidence discriminates between correct and incorrect decisions. These two measures
are distinguishable (see Figure 2.2), as for instance a person might be overall very confident (i.e. high confidence bias) but still be able to distinguish between her correct and incorrect decision (i.e. being systematically more confident when being correct) (Fleming and Lau, 2014).

![Figure 2.2: Illustration of the theoretical distinction between metacognitive sensitivity and confidence bias. Reproduced with permission from Fleming and Lau, 2014. Each graph shows a probability distribution of confidence ratings for correct (blue) and incorrect (red) trials, where confidence is presented on the x-axis. Confidence bias describes the overall level of confidence, independent of whether the trial is correct or incorrect, and thus captures a general tendency to be over- (or under-) confident. Metacognitive sensitivity represents the separation between the distributions, i.e. the extent to which participants are able to discriminate between their correct and incorrect decisions based on their confidence ratings.]

Signal-detection theory naturally handles this distinction and enables to tease apart these two quantities (Green et al., 1966). Importantly, signal-detection theory can be easily applied to confidence ratings as a function of correct and incorrect decisions (Clarke et al., 1959). Applying signal-detection theory models to confidence data separates the tendency to report high confidence (bias; derived by fitting confidence criteria) from the ability to separate correct from incorrect decisions (sensitivity; derived from the overlap of the confidence distributions for correct and incorrect decisions).

Maniscalco & Lau (2012) developed such a signal-detection based measure of metacognitive ability, called meta – d’. This metric measures metacognitive sen-
sensitivity independently of a confidence bias. Since confidence is assumed to reflect a read-out of decision-evidence, there is an inherent relation between the evidence available for the decision and the evidence available for a confidence rating \cite{Maniscalco2012}. Thus, when measuring metacognitive sensitivity, this quantity will be strongly influenced by perceptual performance, with better metacognitive ability when task performance is higher. Hence, when making specific inferences about metacognitive abilities, it is crucial to experimentally equate performance \cite{Fleming2010} or statistically control for it \cite{Fleming2014}. The measure of $meta - d'$ is in the same units as the perceptual sensitivity measure $d'$.

It has been suggested that forming the ratio of $meta - d'/d'$ is a pure measure of metacognitive efficiency \cite{Maniscalco2012} which should be used for answering specific questions about metacognitive ability \cite{Fleming2017}.\footnote{Ratios are sensitive to the value in the denominator, potentially resulting in extreme values. When $d'$ values are small, statistical control of $d'$ is preferable to using the ratio $(meta - d'/d')$.
}

As explained, confidence has been seen as a noisy read-out of decision evidence, whereby higher levels of noise corruption will result in less insightful confidence ratings, i.e. lower metacognitive ability. Therefore, it has been assumed that $meta - d'$ should not exceed $d'$ as the best achievable outcome would be a noise free read-out, resulting in $meta - d' = d'$ \cite{Maniscalco2012}. However, empirically it has been repeatedly observed that the $meta - d'/d'$ ratio can exceed values of 1 \cite{Fleming2017, Moreira2018}. A potential explanation might be that confidence depends on partially separable information to the decision itself \cite{Fleming2017, Kiani2014, Zylberberg2012, Miyoshi2020}, or that continued processing of evidence after the decision makes more evidence available at the time point of the confidence ratings \cite{Bronfman2015, Moreira2018, Murphy2015, Pleskac2010, Yu2015}.

Independent of the exact origin, it has been shown that individual differences in metacognitive abilities are related to a variety of real-world issues such as mental-health \cite{David2012, Rouault2018}, educational success \cite{Hacker2009} or the formation of societal attitudes \cite{Fischer2020, Fischer2019}. For instance, in an educational setting metacognitive monitoring can help students to identify topics in which they have insufficient knowledge, thus indicating that they need to study more in this domain.

### 2.3.4 Metacognitive control and the influence of confidence on post-decision processing

Confidence might be an especially useful signal when it is used for guiding future information processing and behaviour (although see for instance \cite{Frith2012} arguing for other, (social) reasons for why confidence representations could be important). In particular, confidence is thought to play a role in cognitive offloading
(Gilbert et al., 2019; Risko & Gilbert, 2016), as an internal learning signal in the absence of external feedback (Guggenmos et al., 2016) as well as in guiding information seeking (Desender et al., 2018, 2019b).

Most importantly however, it has been suggested that confidence influences the incorporation of new evidence as implied by Bayes theorem (see section 2.2.1). As suggested by Bayes theory, it has been shown that high confidence reduces belief updating based on new information (Meyniel et al., 2015a; Meyniel & Dehaene, 2017). This confidence weighting describes a normative view of the influence of confidence on post-decision evidence integration. Theoretical models have explained the neural implementation of this confidence weighting as an amplification of evidence processing when confidence is low and dampened processing of evidence when confidence is high (Atiya et al., 2019). These predictions manifest in human neuroimaging data, showing that confidence (or confidence related neural signatures) modulates the neural processing of post-decision evidence (Murphy et al., 2015) by making the brain less sensitive to surprising information, i.e. reducing neural gain (Meyniel, 2020).

While such confidence-modulated information processing is in line with a Bayesian account, it does not necessarilly mean that this modulation implements a perfectly Bayesian belief updating. For instance, research in the social sciences has reported that confirmation bias is strongest when people are highly confident (Park et al., 2010; Pomerantz et al., 1995). This suggests that high confidence might induce biased information processing. Importantly, these two alternatives can be empirically distinguished. Bayesian belief updating predicts a symmetric reduction in sensitivity for confirming and disconfirming information when confidence is high, whereas a confidence induced confirmation bias would induce a selective bias against disconfirming information.

2.4 Neural basis of perceptual decision-making

2.4.1 Non-human primates

The neural basis of perceptual decision-making has been studied in detail in non-human primates, often using RDKs. In line with ideas of sequential sampling models, it has been assumed that activity in decision relevant brain regions should track the evolution of a DV, ramping up over time until a decision threshold is reached and a choice is elicited (Gold & Shadlen, 2007). Brain areas that are involved in tracking such DV should show specific characteristics including: a) a build-up of activity in response to evidence whereby the build-up rate should be proportional to the evidence strength, b) a characteristic build-up to a threshold which elicits a response, a relation between build-up rate and c) response times, d) task accuracy, and e) confidence (Kelly & O’Connell, 2015).
Different brain areas have been shown to fulfil some of these criteria, including the superior colliculus (Horwitz & Newsome 1999), the dorsolateral prefrontal cortex (Kim & Shadlen 1999) and the frontal eye field (Thompson & Schall 2000), whereby the exactly involved brain networks seem to depend to some degree on the presented stimuli (Romo et al. 2002) and the required effector (Romo et al. 2004), e.g. whether a hand movement or saccade is required for indicating a decision. Maybe most prominently, the LIP has been suggested as brain area in which a DV is formed and tracked (Gold & Shadlen 2007; Sugrue et al. 2004). Activity in LIP fulfils all benchmark characteristics for tracking an internal DV, as it builds up over time in response to presented evidence, until a stereotypical decision threshold is reached, predicting decision initiation. In addition, the build-up rate scales with presented evidence strength and predicts choice accuracy, reaction times and confidence (Hanks et al. 2006; Kiani & Shadlen 2009; Roitman & Shadlen 2002; Shadlen & Newsome 2001).

2.4.2 Humans

Different non-invasive neuroimaging techniques have been employed to study equivalent neural dynamics that support decision formation in humans. Studies using functional magnet resonance imaging (fMRI) have identified a variety of brain regions involved in the formation of perceptual decisions, including the dorsolateral prefrontal cortex, the anterior and posterior cingulate cortex, the inferior parietal lobe, the lateral occipital complex and the fusifom/parahippocampal gyrus (Cortese et al. 2016; Filimon et al. 2013; Heekeren et al. 2006; Philiaestides et al. 2011; Philiaestides & Sajda 2007). While it has been questioned to which degree the involved neural areas might be dependent on stimulus characteristics and the required motor responses (Filimon et al. 2013), the more substantial problem with fMRI studies is the slow response of the blood-oxygen-level-dependent (BOLD) signal, which makes it impossible to track a rapid build-up of neural activity as needed for the study of the neural underpinnings of evidence accumulation (Kelly & O’Connell 2015).

Therefore, more conclusive results regarding the neural correlates of evidence integration in humans has been provided by studies using neuroimaging techniques that allow a highly temporal resolved evaluation of neural activity, namely electroencephalography (EEG) and magnetoencephalography (MEG).

In addition to the study of event-related potentials (ERP) that fulfil some mentioned criteria for evidence accumulation signals (Philiaestides & Sajda 2006; Ratcliff et al. 2009), there have been two main neural signatures suggested as candidates for tracking a DV. First, a lateralized build-up of beta and gamma band power (12-36Hz and 64-100Hz, respectively) over the motor cortex has been suggested to track a DV by integrating input from sensory areas (de Lange et al. 2013; Donner et al. 2009). This signal has been shown to be effector selective and thus depends
on the required action (Twomey et al., 2016).

In contrast, an ERP over centroparietal regions, known as the centroparietal positivity (CPP), has been identified as supramodal, effector-independent build-up signal, and thus has been suggested as domain-general evidence accumulation signature in the brain (Kelly & O'Connell, 2013; Kelly & O'Connell, 2015; O'Connell et al., 2012). This neural signature shows a temporal build-up over time, to a stereotypical threshold which triggers a decision (Twomey et al., 2015). Moreover, this signal scales with the presented evidence strength and predicts reaction times (Kelly & O'Connell, 2013; O'Connell et al., 2012), thus fulfilling benchmark characteristics for a neural analogous of a drift-rate (Philiastides et al., 2014). Additionally, this neural signature has been shown to correlate with decision confidence and accuracy of error detection, as would be predicted from a neural signal tracking a DV (Gorman & Philiastides, 2015; Murphy et al., 2015). However, the CPP seems not to be choice-specific, i.e. it does not differentiate between left and right decisions in a RDK (Kelly & O'Connell, 2015). Therefore, this neural signature might not be available for tracking choice-specific neural evidence.

Nonetheless, this signature tracked evidence accumulation over a range of different tasks and independent of required action (Kelly & O'Connell, 2013; Murphy et al., 2015; O'Connell et al., 2012). Overall, this indicates that time-resolved ERP measures can be used in order to derive neural metrics of evidence integration.

2.4.3 Neural mechanisms of post-decision evidence integration

While many studies have investigated the neural mechanisms underlying evidence accumulation, there has been less focus on extending these mechanisms to post-decision evidence processing. Murphy et al. (2015) directly extended existing frameworks to post-decision evidence accumulation. They showed that a post-decisional CPP build-up represented a post-decisional evidence accumulation process that was used for error detection. This CPP build-up received input from medio-frontal areas, indicating a top-down influence on the post-decisional evidence integration. However, in this study, post-decision evidence integration was investigated with respect to endogenous processing of evidence that was still “in the pipeline” since no additional information was presented after the decision.

Interestingly, studies that used probabilistic learning tasks also identified a posterior P300 as relevant neural signature of surprise and belief updating (Meyniel, 2020; Nassar et al., 2019). Based on the functional similarity between P300 and CPP (Twomey et al., 2015) this might indicate a role of similar neural mechanisms underlying pre- and post-decision information integration.

Fleming et al. (2018) investigated which brain regions are relevant for the processing of exogenously presented post-decision evidence. To this end, participants
performed a random dot motion task and received additional information (i.e. they saw another RDK) after their initial decision. When the new information confirmed the initial decision, this should increase participants’ internal evidence for their decision, while disconfirming post-decision evidence should decrease the evidence in favour for the initial decision. In this set-up, the posterior medial prefrontal cortex coded for the likelihood that the initial decision was (in)correct. This indicates that the brain might track a signal regarding whether or not a previous decision needs to be revised, i.e. post-decision evidence seems to be processed in a reference frame with respect to a previous decision [Fleming et al., 2018; O’Connell & Murphy, 2018].

In line with these findings, a recent study [Kappes et al., 2020] has investigated the neural processes underlying the incorporation of social post-decision information. They similarly showed that the posterior medial prefrontal cortex tracked the strength with which this additional information indicated the correctness of an initial decision. However, interestingly this was only true for confirming but not disconfirming information, indicating a lack of neural processing for disconfirming evidence. These findings are consistent with a neural process underlying confirmation bias.

Taken together, this work suggests that pre-frontal brain activity might be especially relevant for post-decision evidence processing with respect to a previous decisions, and also indicates alterations in this neural process might support a confirmation bias. Summing up, both centroparietal as well as frontal brain activity might be relevant for tracking post-decision evidence.

2.4.4 Multivariate decoding enables the investigation of neural information processing

As shown in the previous section, there might be various brain areas relevant for the incorporation of post-decision evidence. A multitude of potential candidate regions that could track the computation of interest (in this case post-decision evidence integration) makes it hard to know \textit{a priori} which brain signal will be most critical. Moreover, when being interested in the information processing of the brain as a whole (instead of activity in specific sub-regions) it becomes a non-trivial task to combine signals from different regions. In a situation where the key research questions are related to neural representation of mental states rather than the underlying functional anatomy, neural decoding techniques can be a powerful tool [Haynes & Rees, 2006].

Instead of investigating whether activity at every location is independently related to the mental state of interest (e.g. evidence integration), multivariate decoding analyses test whether brain activity distributed over the scalp (often the whole brain) differentiates between the mental states of interest. Hereby, the decoding
algorithm finds (complex) topographical patterns of activity that best differentiated between the mental states of interest and therefore enables us as researchers to infer the mental state based on neural activity (Haynes & Rees, 2006). This approach solves the outlined problems by taking whole brain data into account and directly providing a weighting of different neural signals.

Multivariate analysis further has the advantage of enhanced statistical power by combining information from many different regions (Sidtis et al., 2003). This enables us to decode mental features that were not decodable from univariate analysis (see Figure 2.3), as the information contained in each region might be weak (Haynes & Rees, 2005). This increase in statistical power then enables more sophisticated analysis, for instance of trial-by-trial data (Kurth-Nelson et al., 2016; Liu et al., 2019).

Figure 2.3: Illustration of univariate versus multivariate neuroimaging analysis. The illustration shows a situation in which a participant’s choice for left or right motion is inferred from brain activity. For illustration purposes I consider a simplified situation in which brain activity from only 3 channels is recorded. Each channel independently does not allow a reliable distinction between brain activity for left versus right choices. However, considering the brain activity of these three channels in combination can distinguish between these choice options. In this example, a simple difference between activity at channel A and activity at channel C shows a clear contrast in brain activity between these two types of trials.

In line with this, multivariate decoding has successfully been used in perceptual decision-making tasks to decode upcoming decisions (Cortese et al., 2016; Peters et al., 2017). Hereby, complex activity patterns distributed across the scalp seem to contribute to the decoding of perceptual choices (Peters et al., 2017).

Since the neural questions in this thesis are mainly related to the neural processing of evidence per se, rather than about precise anatomical hypotheses, I will employ multivariate decoding techniques for analysis of neuroimaging data throughout my thesis.
2.5 Open questions

1. Confidence has been suggested as an internal control signal that guides future information processing and behaviour. Hereby, Bayesian belief updating makes clear predictions about how confidence in a belief should influence the weighting of new information ([Meyniel et al., 2015a]). While most empirical findings appear in line with Bayesian principles ([Meyniel, 2020; Meyniel et al., 2015a; Meyniel & Dehaene, 2017]), it has not been explicitly tested whether confidence indeed implements a Bayesian weighting of new information or whether this confidence weighting instead implements a biased processing of new information ([Park et al., 2010; Pomerantz et al., 1995]). In chapter 3, I investigate how confidence in an initial decision influences post-decision evidence processing, both in terms of computational as well as neural mechanisms.

2. Altered post-decision processing has been suggested as a cognitive feature influencing a wide range of societal issues ([Klayman, 1995; Nickerson, 1998]). For instance, societal polarization and entrenchments of beliefs have been suggested as potential consequences of biased information processing ([Lilienfeld et al., 2009; Zmigrod, 2020]). However, this link has rarely been tested explicitly. In chapter 4 and 5, I use behavioural tasks in conjunction with computational models in order to investigate the underlying computational mechanisms influencing societal issues such as belief radicalization. In particular, I will relate computational alterations in information processing to real-world (political) beliefs. In chapter 4, I investigate a relationship between metacognitive ability, the processing of post-decision evidence and radical political beliefs. In chapter 5, I extend this research to investigate the relationship between dogmatism and active information search.

3. Findings from chapter 3 indicate that confidence modulates post-decision evidence processing in a manner that induces a neural confirmation bias. This suggests that well aligned confidence (i.e. high metacognitive ability) might be crucial for the accuracy with which new information is processed. In chapter 6, I use simulation-based modelling to theoretically explore the effects of metacognition on post-decision evidence processing. In chapter 7, I use these insights to test whether a metacognitive training procedure can successfully boost participants processing of new information.
Chapter 3

Confidence drives a neural confirmation bias

3.1 Introduction

The philosopher Bertrand Russell opined “The most savage controversies are about matters as to which there is no good evidence either way”. While this view applies in some situations, even more troubling are instances where polarization and entrenchment of opinion persists in the face of contrary evidence, exemplified by debates on climate change and vaccinations. This polarization is most evident when opposing parties are highly confident in their positions (Park et al. 2010, Pomerantz et al. 1995). A psychological-level explanation for such entrenchment is the idea that people selectively incorporate evidence in line with their beliefs, known as confirmation bias (Nickerson 1998). Although an extensive literature has documented this bias in behaviour (Lord et al. 1979, Klayman 1995, Nickerson 1998), the underlying cognitive, computational and neuronal mechanisms are not understood.

So far, an investigation of confirmation bias has been restricted largely to scenarios involving complex real-world beliefs such as political attitudes (Kaplan et al. 2016, Lord et al. 1979, Nyhan & Reifler 2010). However, the complexity of such higher-order beliefs makes it difficult to disentangle the various contributors to biased information processing. For instance, people may have a strong personal investment in their political opinions, leading to a significant motivation to discount new information that goes against their beliefs. Intriguingly, confirmation biases have recently been demonstrated in low-level perceptual tasks that are unlikely to evoke such motivated reasoning (Braun et al. 2018, Talluri et al. 2018, Urai et al. 2019). These studies indicate a source of confirmation bias may be a generic shift in the way the brain incorporates new information.

Here I adopt such a task to study the computational and neural basis of post-decisional shifts in sensitivity to choice-consistent information. Perceptual decision-making is well-described using sequential sampling models which assume the brain
accumulates noisy evidence for each choice option to a decision bound (Gold & Shadlen, 2007). This accumulation process is thought to be supported by neuronal populations in parietal and prefrontal cortex (Kiani & Shadlen, 2009; Kelly & O’Connell, 2015). Importantly, while perceptual tasks allow tight control over the processes involved, they also permit generalisation to more complex decisions (Hauser et al., 2017; Rollwage et al., 2018; Rouault et al., 2018), and similar principles appear to underlie choice and confidence formation in both simple and more complex tasks (Pleskac & Busemeyer, 2010; Sanders et al., 2016). However while the processes underlying perceptual decision-making have been studied in detail, little is known about the mechanisms governing accumulation of evidence after a choice has been made, or how such processing is shaped by pre-existing beliefs and confidence (Bronfman et al., 2015; Desender et al., 2018; Fleming et al., 2018; Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; Talluri et al., 2018; Van Den Berg et al., 2016).

Here I combine theoretical models and neural metrics to identify alterations in post-decisional processing that may contribute to the phenomenon of confirmation bias. In total, three experiments were conducted to answer the different facets of this question. The first experiment was set up to show a direct influence of confidence on post-decision processing and changes of mind. The second experiment was optimized for fitting DDMs in order to investigate the computational alterations in post-decision evidence accumulation induced by confidence. In the third experiment, I recorded MEG activity to investigate neural changes related to post-decision processing. Across all experiments, participants were presented with a sample of moving dots (pre-decision evidence) before indicating their initial decision (motion to the left or right) and confidence in their choice (see Figure 3.1). They were then presented with a second sample of moving dots (post-decision evidence) before making a final choice and providing a confidence estimate. Importantly, pre-and post-decision evidence always indicated the same direction of motion such that post-decision evidence was helpful. Accordingly, an ideal Bayesian observer should use post-decision evidence to change its mind after initial mistakes whereas a confirmation bias would blunt this belief flexibility (Fleming et al., 2018; Rollwage et al., 2018).

3.2 Methods

3.2.1 Participants

Each study contained a different group of participants. I analysed data from 28 participants in study 1 (Mean age = 23.8; SD age = 6.3; 16 female) and 23 participants in study 2 (Mean age = 25.7; SD age = 7; 12 female). Participants were excluded based on the following set of pre-defined criteria: using the same initial confidence rating
more than 90% of time (N=3 in study 1; N=2 in study 2), performance below 55% or above 87.5% correct decisions in one of the pre-decision evidence conditions (see explanation of the experimental conditions below) indicating non-convergence of the staircase procedure (N=3 in study 1; N=2 in study 2). For the MEG study 3, participants conducted an initial behavioural training session before being screened according to the same criteria reported above. MEG data of a final sample of 25 subjects was analysed (Mean_age= 24.6; SD_age= 4.1; 16 female). Data of 4 subjects could not be analysed due to technical problems with recording triggers. As I applied machine learning classification algorithms to the neural data in order to decode decisions (left versus right) and confidence (high versus low) it was important that participants showed relatively balanced responses for these two categories. 2 subjects were excluded because they chose one response more than 80% of the time for either the decision or confidence. In addition to a basic payment (£10 for behaviour and £20 for MEG) participants received a performance-based bonus (up to £5 for behaviour and £8 for MEG). All studies were approved by the Research Ethics Committee of University College London (8231-001) and all subjects gave written informed consent.

3.2.2 Stimuli and experimental design

The psychophysical task was an adaptation of the task used by Fleming et al. (2018) and programmed in MATLAB 2012a (Mathworks Inc., USA) using Psychtoolbox-3.0.14. Stimuli were RDKs, viewed at a distance of approximately 45 cm. The RDKs were clouds of white dots (0.12° diameter) within a white circular aperture with a radius of 7° on a grey background that lasted for 350 ms. The direction of motion was rightward or leftward along the horizontal meridian. The speed of movement was 5° per second and the density of dots in the whole experiment was set to 60 dots per degree. Each set was replotted three apertures later in which a
subset of dots, determined by the percent coherence, was offset from their previous
location towards the target movement direction, and another subset was offset in
the opposite direction, whereas the rest was replotted randomly.

Unlike in a classical RDK stimulus, dots moved coherently in both the target
direction and the opposite direction. The remaining dots moved randomly (percent-
ages described below). I used a psychophysical manipulation of positive evidence to
dissociate subjective confidence from objective task performance (Zylberberg et al.
2012). In the high positive evidence the proportion of dots moving in the incorrect
direction was set to 15% and the proportion moving in the correct direction was a
higher percentage, staircased to ensure the targeted performance level (see below).
In the low positive evidence condition the motion coherence of dots moving in the
incorrect direction was set to 5%, whereas the dots moving in the correct direction
was also staircased to ensure the same performance as in the high positive evidence
condition. The rationale for this manipulation was that accuracy and confidence
are usually highly correlated, hindering specific claims about the unique role of
confidence. The positive evidence manipulation enabled us to selectively increase
confidence while keeping performance constant, thus making it possible to determine
direct effects of changes in confidence on post-decision processing.

All experiments adapted a full 2 (pre-decision positive evidence level) by 2 (post-
decision evidence strength) factorial design yielding a total of 4 experimental con-
ditions each corresponding to 90 trials. high positive evidence and low positive
evidence stimuli were each followed by one of two post-decision evidence conditions
(weak or strong). For the post-decision evidence a constant level of evidence in the
incorrect direction was employed (i.e. I did not manipulate the overall amount of
positive evidence in the post-decision phase). The post-decision coherence level in
the incorrect direction was derived from the averaged staircased pre-decision val-
ues as [incorrect coherence low positive evidence + incorrect coherence high pos-
itive evidence]/2. Weak post-decision evidence stimuli were created by specifying
correct-direction coherence as [staircased correct coherence low positive evidence +
staircased correct coherence high positive evidence]/2. Strong post-decision evidence
stimuli were then derived by multiplying this coherence level by a factor of 1.3.

3.2.3 Task procedure

In every study, participants first performed 180 trials of a calibration phase before
performing the main task which consisted of 360 trials (behavioural studies) or 352
trials (MEG study). In the calibration phase subjects judged whether the dots were
moving to the left or to the right side of the screen, without rating their confidence
or seeing additional post-decision evidence. The response had to be given within
1.5 sec after stimulus offset. Low and high positive evidence stimuli were randomly
interleaved. As described above, the coherence of the target direction was adapted
with a staircase procedure to obtain a performance of 60% correct in study 1 and 71% correct in studies 2 and 3 (García-Pérez, 1998).

The main task had the same core structure for all studies with slight variations, explained below, to optimize each study for the specific research question and planned analysis. Participants were first presented with a moving dot stimulus before they indicated their initial decision (left or right) together with a confidence rating. In behavioural studies 1 and 2 the decision was indicated by pressing the left or right arrow key on the keyboard and was directly combined with a graded confidence rating (7-point sliding scale between 50% and 100%), where pressing the (same) arrow key again moved a slider along the confidence scale. In the MEG study, subjects first made a left versus right decision, before giving a binary high/low confidence rating. After this initial decision, participants received a second sample of moving dots (i.e. post-decision evidence) which was always in the same (correct) direction as the pre-decision evidence presentation, but of variable strength. Subjects were instructed that this evidence was bonus information that could be used to inform their final decision and confidence. After the post-decision evidence, participants were again asked to judge the motion direction and indicate their confidence.

### 3.2.3.1 Design alterations in behavioural study 2

In study 2, I optimized the experimental design to allow drift-diffusion modelling of the second/final decision. While in study 1 subjects had to withhold their final response for 300ms after the offset of the post-decision evidence (i.e. responding was only possible after this delay), in study 2 participants were able to make their final response freely as soon as they had decided. This allowed us to use response times as a proxy for crossing a decision threshold, which would not have been possible if the response was delayed.

### 3.2.3.2 Design alterations in MEG study 3

In the MEG study, participants indicated their responses by pressing up or down button on a keypad with their right thumb. I disentangled the participant’s decision (left/right and high/low confidence) from the motor response they had to perform (pressing the up or down key on a key pad), by randomising the mapping between decision options and key presses. Specifically, on any given trial leftward motion could be indicated by pressing the up key and on another trial by pressing the down key. Similarly, high confidence could be indicated in one trial by pressing the up key and in a different trial by pressing the down key. The mapping between decisions and motor responses was revealed once responding was possible, by presenting the letters L or R (and H or L for confidence ratings) above/below the horizontal plane. This approach ensured that decoding of motion direction was not trivially confounded by motor preparation signals. Additionally, I introduced
delays of 500ms after the presentation of each stimulus but before participants were informed about the response mappings to allow decoding analysis to be applied in a time window when subjects could form an abstract decision about motion direction but were not yet able to prepare a response.

3.2.4 Scoring and bonus payment

Participants were instructed to rate their confidence as a subjective probability of being correct and were rewarded according to the correspondence between their confidence and task accuracy. An incentive-compatible Quadratic Scoring Rule (Brier, 1950) was applied equally to both the initial and final decisions:

\[
\text{points} = 100 \times \left[ 1 - (\text{correct} - \text{confidence})^2 \right]
\]

where correct is equal to 1 on trial \(i\) if the choice was correct and 0 otherwise, and the subject’s confidence rating on this trial can range between 0 and 1. The Quadratic Scoring Rule is a proper scoring rule in that maximum earnings are obtained by jointly maximizing the accuracy of both choices and confidence ratings. This scoring rule also ensures that confidence is orthogonal to the reward the subject expects to receive for each trial: maximal reward is obtained both when one is maximally confident and right, and minimally confident and wrong. The points gained on each trial were summed and participants were given a £1 bonus payment for every 15000 points earned. After each block participants were informed of their current total number of points. This was the only performance feedback that was given and subjects did not receive specific information regarding the correctness of their motion direction decisions.

3.2.5 Drift-diffusion modelling

Drift-diffusion modelling was conducted in Python 2 using Jupyter Notebook (5.50). The model was fit using accuracy coding such that decision boundaries and reaction time distributions corresponded to those for correct and incorrect responses. However, by design, initially correct decisions led to confirming post-decision evidence (because the motion direction was always the same in the pre and post-decision periods) and initially incorrect decisions always led to disconfirming post-decision evidence.

Within the DDM there are two natural ways to account for biases in a decision process: by shifting the starting point towards one of the decision boundaries, or by altering the drift rate to induce a bias in the processing of information. I also considered the possibility that other factors (e.g. decision bound) could be altered, but in initial simulations such changes were unable to explain the observed behavioural patterns. Since it has been reported that confidence might affect boundary separation (Van den Berg et al., 2016), I included a dependency of the boundary
separation on confidence in each of the models (note however that a symmetrical influence on boundary separation cannot explain any choice-dependent effects on changes of mind).

A hierarchical Bayesian variant of the DDM (hDDM) enabled us to investigate the dependencies of the model parameters on the initial decision and confidence on a trial-by-trial basis \cite{Wiecki2013}. The hDDM simultaneously estimates individual parameters drawn from a group distribution using Markov-Chain Monte-Carlo methods. This procedure not only estimates the most likely value of the model parameters but also uncertainty in the estimate. The hDDM toolbox \cite{Wiecki2013} was used to compare 10 hDDMs. The best-fitting model was identified by comparing Deviance Information Criterion (DIC) scores and ensuring that the winning model adequately fitted the qualitative data patterns. A regression analysis was used to investigate the dependency of the starting point and drift-rate parameters on the initial decision (1= correct decision leading to confirmatory post-decision evidence, -1= incorrect decision leading to disconfirmatory post-decision evidence), initial confidence (parametrically ranging from -1 to 1) or their interaction.

In all models the drift-rate, starting point, non-decision time and boundary separation were fitted hierarchically with individual parameter estimates for each participant, whereas dependencies of starting point and drift-rate on experimental factors were estimated as fixed group-level effects. In all model fits I incorporated an influence of post-decision evidence strength on the drift-rate. First a baseline model was estimated where none of the parameters depended on confidence or an initial decision. Subsequently, I created three model families that had dependencies of starting point and/or drift-rate on (i) initial confidence, (ii) initial decision or (iii) the interaction of initial confidence $\times$ initial decision (i.e. confidence was allowed to amplify or attenuate the influence of the initial decision on the starting point and/or drift-rate). Within each model family I created three different models with dependencies of these variables on starting point, drift-rate or both.

**Baseline model (Model 1):**

starting point $\sim 1$

drift-rate $\sim 1+ \text{ post-decision evidence strength}$

boundary separation $\sim 1+ \text{ confidence}$

**Confidence dependency (Model 4):**

starting point $\sim 1+ \text{ confidence}$

drift-rate $\sim 1+ \text{ post-decision evidence strength } +\text{confidence}$

boundary separation $\sim 1+ \text{ confidence}$

**Initial decision dependency (Model 7):**

starting point $\sim 1+ \text{ initial decision}$
drift-rate $\sim 1 + \text{post-decision evidence strength} + \text{initial decision}$
boundary separation $\sim 1 + \text{confidence}$

**Full model (Model 10):**

starting point $\sim 1 + \text{confidence} + \text{initial decision} + \text{confidence} \times \text{initial decision}$

drift-rate $\sim 1 + \text{post-decision evidence strength} + \text{confidence} + \text{initial decision} + \text{confidence} \times \text{initial decision}$

boundary separation $\sim 1 + \text{confidence}$

Reaction times faster than 200ms were discarded from the model fits and the outlier probability was set to 0.05, as recommended in previous literature (Wiecki et al., 2013). The models were estimated with a Markov chain of 100,000 samples with 50,000 burn-in samples (i.e. discarding the first 50,000 iterations), and a thinning factor of 25, resulting in 2,500 posterior samples. To ensure convergence, the posterior traces and their autocorrelation were inspected and the Gelman–Rubin statistic was calculated for each parameter. The posterior distributions of the best-fitting model were interrogated to retrieve parameter estimates.

### 3.2.6 MEG pre-processing

MEG was recorded continuously at 600 samples/second using a whole-head 273-channel axial gradiometer system (CTF Omega, VSM MedTech), while participants sat upright inside the scanner. Data was segmented into 8200 ms segments from 200 ms before to 8000 ms after trial onset, where each segment encompassed one trial. Each epoch was aligned to the onset of the trial or, for analysis of the post-decisional phase, was realigned to the onset of post-decision evidence (to minimize any presentation delays that may have occurred during the trial). The data were resampled from 600 Hz to 100 Hz to conserve processing time and improve signal to noise ratio, resulting in data samples spaced every 10 ms. All data were then high-pass filtered at 0.5 Hz to remove slow drift. All analyses were performed directly on the filtered, cleaned MEG signal, consisting of a 273 channel $\times$ 821 sample matrix for each trial, in units of femtotesla.

### 3.2.7 Generalising a pre-decision classifier to the post-decision phase

I built a machine-learning classification algorithm to predict participants’ decisions on each trial (leftward vs. rightward motion) at each timepoint during the decision phase. Having trained such an algorithm I could then apply it to a distinct set of trials and use the probabilistic prediction of the classifier as a neural DV for leftward versus rightward motion (Cortese et al., 2016; Peters et al., 2017). Specifically,
I used a support-vector machine (SVM) classifier trained on sensor-level whole-brain activity (normalized amplitude of all MEG channels). The classifier labels were the trial-by-trial choices made by participants (left or right) while the features encompassed a matrix of activity at each MEG sensor (z-scored for each time point) at a given time point (average activity over 100ms window, shifted in steps of 10ms). The classifier was trained on MEG activity in the pre-decision time phase (e.g. 250 ms after the onset of pre-decision evidence) and then reapplied to the corresponding time point in the post-decision phase (e.g. 250 ms after the onset of post-decision evidence). I computed the predictions of the classifier across an 850 ms time window, starting with post-decision stimulus onset and ending with the presentation of response options (i.e. when the mapping between choices and motor responses was revealed).

I used linear kernels and a default regularization parameter of C=1 within the svmtrain/svmpredict routines of libsvm [Chang & Lin 2011]. A leave-one-out procedure was used, training the classifier on all trials except one (using pre-decision data only) and testing it on the left-out trial (using post-decision data). Training the SVM results in a hyperplane that best separates the two classes of trials (see Figure rreffig:NeuralIntegrationA) in a high-dimensional space. If a trial is far away from this hyperplane it is unlikely to be a misclassification, while trials that are close to the hyperplane might easily be misclassified. Thus, the distance to the hyperplane represents the decodable evidence for a decision and can thus be used as a graded measure of the neural DV ([Cortese et al.] 2016; [Peters et al.] 2017).

After reapplying the classifier to every trial and time point during the post-decision phase, I obtained a timeseries of neural evidence accumulation within each trial (see Figure 3.5A, right panel). I focussed on the time from the onset of the post-decision stimulus to the timepoint of peak decodability at which the pre-decision classifier best generalized to the post-decision phase. The accumulation process can be summarized by fitting a linear regression to the time series (see Figure 3.5A, right panel) on each trial, where the slope is analogous to the drift rate in a drift-diffusion model, and the intercept analogous to the starting point. A positive slope corresponds to a change of the neural DV towards predicting rightward motion decisions while a negative slope corresponds to a change towards leftward motion (see Figure 3.5B). By taking the absolute value of these slope values (i.e. reversing the sign on trials in which leftward motion was presented), I could derive a general index for the sensitivity of the neural DV to the motion direction presented on the screen (see Figure 3.5A & B). Based on our behavioural findings I expected that both the slope and the intercept would be influenced by the interaction of initial decision (confirmatory post-decision evidence =1; disconfirmatory post-decision evidence=-1) × confidence (low confidence= -1; high confidence =1). Thus, I entered the initial decision, confidence and their interaction as simultaneous predictors in a hierarchical
regression model.

### 3.2.8 MEG topography contributing to classification accuracy

To explore which brain areas carried the information about evidence for a left versus a right decision, I trained a SVM classifier for each participant at the time point of highest decodability using subsets of 30 randomly selected sensors and repeated this procedure 2500 times. The contribution of each sensor $s$ was taken to be the mean of all prediction accuracies achieved using an ensemble of 30 sensors that included $s$ [Kurth-Nelson et al., 2015; Liu et al., 2019].

### 3.2.9 MEG temporal generalization

The extent to which a classifier trained on neural data obtained from one time point generalizes to other time points can provide insight how mental representations change over time [King & Dehaene, 2014]. I utilized this temporal generalization method to formally test whether the same processing steps (leading up to a decision) occur at similar times in the pre- and post-decision phases. Most critically, I also investigated whether this processing cascade was altered by participants’ confidence in their choice.

For the temporal generalization analysis I trained our classifier on every time-point in the pre-decision phase and tested it on every timepoint in the post-decision phase yielding a 2D matrix of decoding accuracy (see Figure 3.5A top-left panel). A 4-fold stratified cross-validation was implemented for each subject and repeated 100 times to account for potential random biases in assigning trials to folds. Through this stratification I obtained a balanced number of trials within each condition in each fold (left/right decision, high/low confidence, change/no change of mind, and all combinations of these factors). Classifiers were trained on 3 out of 4 folds and tested on the left-out fold. Decoding accuracy was determined by the area under a Receiver Operator Curve that sought to predict the decision based on the continuous decision variable outputted by the classifier. Decoding accuracy was calculated separately for the 4 different conditions (low confidence and change of mind; high confidence and change of mind; low confidence and no change of mind; high confidence and no change of mind). Importantly, classification accuracy was based on how well the initial decision (rather than the final decision) could be predicted based on neural data. Since we are dealing with a two-class decoding problem one can directly infer the decoding accuracy of the alternative decision from the classification accuracy of the initial decision.

I estimated the main effect of confidence on decoding accuracy to isolate confidence-induced changes in temporal generalisation from the pre- to post-decision
Chapter 3. Confidence drives a neural confirmation bias

phase. I used a cluster-based permutation test \cite{MarisOostenveld2007, Nichols2002} to determine statistical significance ($p < .05$, corrected for multiple comparisons). I calculated the contrast of high>low confidence averaging over change/no change of mind trials $[[\text{high confidence and no change of mind} - \text{low confidence and no change of mind}] + [\text{high confidence and change of mind} - \text{low confidence and change of mind}]]$. I identified adjacent timepoints all individually exceeding t-values corresponding to $p < .05$ uncorrected, and stored the sum of t-values for each cluster. I then applied a sign-flip permutation test (randomly switching the contrast direction for a subset of subjects of the sample, i.e. low-high instead of high-low) and repeated this procedure 1000 times. The distribution of summed t-values over all permutations built the null distribution for our statistical test. If the observed sum of t-values within a cluster exceeded the 5% quantile of this distribution (separately calculated for negative and positive values) I labelled this cluster as showing a significant main effect of confidence in this portion of the temporal generalisation matrix.

3.3 Results

3.3.1 Adaptive use of post-decision evidence

First, I established that using post-decision evidence is indeed adaptive in this task. Subjects improved their choice accuracy from the initial to the final decision ($t(27) = 7.45, p < 10^{-7}$, see Figure 3.2A), showing that the new evidence was beneficial. Specifically, this increase in accuracy was due to participants changing their mind more often when they were initially incorrect compared to when they were correct ($t(27) = 13.39, p < 10^{-12}$, see Figure 3.2C), indicating that participants changed their mind to reverse mistakes. This improvement in accuracy was accompanied by a decrease in reaction times of the final decision compared to those of initial decisions ($t(27) = 3.21, p = .003$, see Figure 3.2B). I also examined participants’ final confidence ratings (see Figure 3.2D, a confidence rating below 50% indicates that participants changed their mind) as a function of post-decision evidence strength, showing that participants reduced their confidence in initially erroneous decisions ($t(27) = 7.38, p < 10^{-7}$) and that this effect was most pronounced when post-decision evidence was strong ($t(27) = 3.03, p = .005$).

3.3.2 Effects of confidence on changes of mind

Having established that the usage of post-decision evidence was adaptive, I hypothesised that a confirmation bias would occur more often when people are highly confident in their original choice \cite{Atiyaetal2019, Desenderetal2018, Qiuetal2019}. In order to dissociate subjective confidence from objective performance I
used a psychophysical manipulation ("positive evidence" (Zylberberg et al., 2012)) to selectively boost participants’ confidence (mean difference=.024, CI=[.008, .04], Cohen’s d = .21, t(27)=3.0, p=.005; see Figure 3.3B) while leaving performance (mean difference=.006, CI=[-.022, .034], Cohen’s d = .02; Bayesian t-test indicating equality: BF01=4.61; see Figure 3.3A) and reaction times (mean difference=.005, CI=[-.029, .018], Cohen’s d = -.04; Bayesian t-test indicating equality: BF01=4.51) unaffected. Moreover, there were no significant differences between left versus right motion choices (mean difference=-.02, Cohen’s d= .11, CI=[-.045, .005], t(27)=-1.68, p=.11).

I next set out to test whether this boost in confidence influenced changes of mind. There were notable individual differences in the degree to which the manipulation boosted participants’ confidence (see Figure 3.3B & C). Importantly, subjects who experienced a stronger confidence boost through the positive evidence manipulation also showed a stronger reduction in changes of mind (r=-.69, p <.0001.; see Figure 3.3C), an effect not explained by an impact of positive evidence on accuracy or reaction time (p=.005 when controlling for these effects). This supports a notion that confidence drives reductions in changes of mind.

### 3.3.3 Confidence induces a selective gain for confirmatory evidence

I next reasoned that confidence may reduce changes of mind by promoting a bias towards processing of confirmatory post-decision evidence. I sought to test this hypothesis by revealing the process through which confidence affects accumulation of post-decision evidence, applying a combination of drift-diffusion modelling and
Figure 3.3: Influence of confidence on post-decision evidence integration (behavioural study 1: n=28 participants). A-B A psychophysical manipulation of positive evidence selectively increased confidence of the first decision (B) while keeping accuracy constant (A). This increase in confidence was replicated across all three studies. Data are presented as mean values ± SEM; grey dots represent individual participant data. Paired t-test (two-tailed): **p=.005; n.s.=non-significant; LPE = low positive evidence condition; HPE = high positive evidence condition. C Between-subject relationship between the degree to which positive evidence increased confidence (x-axis: confidence in the high positive evidence condition - confidence in the low positive evidence condition) and its effect on changes of mind (y-axis: changes of mind in the high positive evidence condition - changes of mind in the low positive evidence condition). This correlation was replicated in all three studies. Orange data points represent subjects showing the opposite of the intended effect of the manipulation on confidence (higher confidence in the low positive evidence condition). Pearson correlation (two-tailed): ***p=.0001.

I considered two potential mechanisms through which confidence might reduce changes of mind. First, confidence might reflect a shift in the starting point of post-decision accumulation to be closer to the bound associated with an initial decision, consistent with a continuation of pre-decisional evidence accumulation (influence on starting point; Figure 3.4A upper panel). Second, confidence may induce selective accumulation of evidence in line with an initial decision (influence on drift rate; Figure 3.4A lower panel) – a clear instance of confirmation bias.

Critically, these two mechanisms make different predictions in terms of the distributions of response times for the final decision (Braun et al., 2018; Urai et al., 2019). I compared 10 DDMs that embodied these different predictions (see Figure 3.4B). I employed accuracy coding such that the bounds correspond to a correct versus an incorrect decision, such that a positive drift-rate represents stronger integration of the presented (correct) motion direction. Note, by design, confirmatory post-decision evidence was received when an initial decision was correct, and disconfirmatory evidence when an initial decision was incorrect. In addition, in light of suggestions that confidence might also affect the separation of decision bounds, and thus the trade-off between speed and accuracy of subsequent decisions (Desender et al., 2019a; Van Den Berg et al., 2016) I allowed for a dependency of boundary recordings of post-decisional fluctuations in a neural decision variable using MEG.
separation on initial confidence in all models.

The models differed as to whether the starting point and/or drift-rate were affected by confidence (models 2-4), accuracy of the initial decision (models 5-7; i.e. correct=1 and incorrect=-1, capturing a general confirmation bias) and their interaction (models 8-10; i.e. capturing a confirmation bias that depends on confidence). The winning model (Model 10) as indicated by the DIC (ΔDIC=151.6 relative to the best model without an interaction term, see Figure 3.4B) incorporated dependencies of starting point and drift-rate on all factors (confidence, initial decision and the interaction) and provided a good fit to the data (Figure 3.4C & D):

\[
\text{starting point} \sim 1 + \text{confidence} + \text{initial decision} + \text{confidence} \times \text{initial decision} \\
\text{drift-rate} \sim 1 + \text{confidence} + \text{initial decision} + \text{confidence} \times \text{initial decision} \\
\text{boundary separation} \sim 1 + \text{confidence}
\]

After accounting for main effects, I observed a dependency of the starting point on the interaction between confidence and initial decision (95% equal-tailed interval=.08-.18; Figure 3.4E righthand panel), indicating participants started the accumulation process closer to the bound of the initial decision when highly confident in their choice. Even more striking was the discovery of a similar interaction effect on drift rate (95% equal-tailed interval=.11-.26; Figure 3.4E righthand panel) indicating participants selectively accumulated evidence supporting their initial choice, and were more likely to do so when they were more confident. While a confidence-related shift in starting point might reflect normative usage of pre-decision evidence (because high confidence in an initial decision might reflect greater pre-decision evidence accumulation, and thus be closer to a post-decisional bound; see Figure 2.1), an influence of confidence on the drift rate is a clear instance of confirmation bias.

I next quantified the relative contribution of both effects to the observed data patterns. To this end, I started with the best-fitting model (model 10) and eliminated either the dependencies of the starting point or the drift-rate. Eliminating starting point dependencies reduced the model fit only slightly (ΔDIC=48.1), whereas elimination of drift-rate dependencies reduced the model fit severely (ΔDIC=425.6). Accordingly, when comparing these two models directly (which are matched for complexity), the model with drift-rate dependencies explained the data better than the model with starting point dependencies (ΔDIC=377.5). This indicates that a confirmation bias, and a boost of confirmation bias through confidence, is better explained by selective accumulation of choice consistent information rather than a shift in starting point.

Such a confirmation bias led to a boost in accumulation of the veridical motion direction following high-confidence correct decisions (as such information served to confirm the original choice), whereas it led to a reduction in evidence accumulation (manifest as a lowered drift rate) following high-confidence errors (as new informa-
tion served to disconfirm an originally wrong decision).

Figure 3.4: Drift-diffusion modelling fits to the second decision (behavioural study 2, n=23 participants). A Illustration of how confidence may reduce changes of mind through either a shift in starting point towards the decision bound of the initial decision (upper panel) and/or a selective increase of drift-rate for evidence supporting the initial decision (lower panel). B Model comparison of the 10 alternative DDMs. Models were compared based on the DIC where lower values indicate better fit. I present ΔDIC representing the difference of each model’s DIC against the best fitting model in the set (the best fitting model 10 has a ΔDIC of zero). Model 1 is a baseline model with no dependencies of drift-rate or starting point. Models 2-4 represent a model family in which the drift-rate and/or starting point are affected by initial confidence. Models 5-7 represent a model family in which the drift-rate and/or starting point are affected by the initial decision (i.e. a simple confirmation bias). Models 8-10 represent a model family in which the drift-rate and/or starting point are affected by initial confidence, initial decision and their interaction (i.e. a confirmation bias boosted or attenuated by confidence). C-D Model simulations (of the best fitting model) reproduce behavioural patterns of accuracy and reaction times of the second decision when plotted as a function of the initial decision and initial confidence. Due to the task structure participants received confirming post-decision evidence when they were initially correct and disconfirming post-decision evidence after initial mistakes. Model simulations are shown as dotted lines, behavioural data as solid lines. Data are presented as mean values +/- 95% confidence intervals. The top panel of D plots the full distribution of response times and model predictions for the different trial types (high confidence and no change of mind, low confidence and no change of mind, high confidence and change of mind, low confidence and change of mind). E Posterior distribution of model parameters of the best-fitting model. The dependencies of the drift rate (purple lines) and starting point (green lines) on initial confidence (left panel), initial decision (middle panel) and the interaction between confidence x initial decision (right panel) are presented. The dotted vertical lines represent an effect of zero/no effect. Note that these dependencies are simultaneously fitted, controlling for mutual influences. Markov-Chain Monte-Carlo sampling of posterior parameter distribution: ***P(parameter>0)>.999. Sec=seconds.

3.3.4 Neural markers of post-decisional processing

While our drift-diffusion model fits support a distinct influence of initial choice and confidence on post-decisional processing, they allow only indirect inference on how
confidence affects evidence accumulation. To quantify this process more directly I used MEG to obtain a time-resolved neural metric of post-decision accumulation. Specifically, I trained a SVM classifier on brain activity (normalized amplitude of all MEG channels) at each time point (10ms timebins) in the pre-decision time window (lasting 850 ms from stimulus onset to the presentation of choice options; note that the trial timeline for the MEG study differed slightly to the timeline presented in Figure 3.1) to predict which choice (left or right) was made on each trial. I then applied the trained classifier to brain activity at the corresponding time point in the post-decision time window, enabling us to derive a probabilistic prediction of neural evidence favouring a leftward versus rightward decision (see Figure 3.5A left panel). Positive values indicate prediction of a rightward decision and negative values indicate prediction of a leftward decision (see Figure 3.5B). I next fitted a linear regression to the time series of classifier predictions within each trial (see Figure 3.5A right panel) to obtain a trial-by-trial neural measure of the starting point (intercept) and drift rate (slope). These measures of neural evidence accumulation (slope) should be highly responsive to the presented motion direction during the post-decision period, and I show this was indeed the case (hierarchical regression: $\beta=.07$, $t(8550)=6.89$, $p<10^{-11}$, Figure 3.5B).

The slopes extracted from this analysis are signed, such that positive values indicate evidence for a rightward choice and negative values evidence for a leftward choice. In order to obtain an unsigned metric of evidence accumulation strength, I flipped the sign of slopes extracted from trials in which leftward motion was presented (I conducted the same flip for the intercept to obtain an unsigned metric of the starting point). This unsigned metric quantifies a propensity to correctly integrate the presented information, analogous to a drift rate in the accuracy coded drift-diffusion model employed in Figure 3.4.

A neural analogue of the drift-rate (or change in internal decision variable) should be related to characteristic features of the observer’s decision. Specifically, stronger internal evidence accumulation should be related to a higher likelihood of having made a correct decision, faster response times and higher confidence (Kelly & O’Connell, 2015; Gold & Shadlen, 2007; Kiani & Shadlen, 2009). In order to check whether our classifier predictions satisfied these criteria for metrics of internal evidence accumulation, I entered both the trial-by-trial slope and intercept of the post-decision accumulation process as simultaneous predictors in a hierarchical regression model to predict a) reaction times, b) choice accuracy and c) confidence of the final decision. Steeper slopes predicted faster reaction times ($\beta=-0.007$, $t(8549)=-2.83$, $p=.005$, see Figure 3.5D), a higher likelihood of a correct decision ($\beta=.16$, $t(8549)=3.05$, $p=.002$, see Figure 3.5E) and higher confidence ($\beta=.14$, $t(8549)=3.53$, $p=.0004$, see Figure 3.5F). I also observed significant effects of the intercept on accuracy ($\beta=.1$, $t(8549)=2.0$, $p=.045$, see Figure 3.5E) and confidence
Chapter 3. Confidence drives a neural confirmation bias

(\(\beta = .12, t(8549) = 3.07, p = .002\), see Figure 3.5F) which is to be expected if participants maintain a representation of the evidence obtained in the pre-decision phase, and if the strength of this pre-decisional accumulation predicts the likelihood of being both correct and confident.

I next asked whether specific sensor clusters drive the classifier performance. Previous studies using EEG have identified a centro-parietal event-related potential (or CPP) as a neural marker of internal evidence accumulation \([\text{Kelly & O’Connell, 2015; O’Connell et al., 2012; Kelly & O’Connell, 2013; Tagliabue et al., 2019}].\) Accordingly, when identifying the features that contributed most strongly to classifier decoding accuracy (Figure 3.5C) I also found that centro-parietal sensors make a disproportionate contribution to an ability to differentiate between left and right decisions.

Having identified a neural metric of evidence accumulation, I next turned to our central question of whether confidence induces a selective accumulation for choice-consistent information as measured using MEG. As hypothesized, I found that after high confidence (vs. low confidence) decisions, accumulation of neural evidence was facilitated if it was confirmatory, but largely abolished if it was disconfirmatory (Figure 3.6A & B). In other words, our MEG analysis reveals that high confidence leads to post-decision accumulation becoming “blind” to disconfirmatory evidence. To formally quantify this effect, I entered the slope and starting point of neural evidence accumulation on each trial into hierarchical regression models with initial decision, high vs. low initial confidence and their interaction as predictors. I obtained a significant effect of initial decision (\(\beta = .042, t(8547) = 2.96, p = .003\)) and its interaction with confidence (\(\beta = .038, t(8547) = 2.64, p = .008\), see Figure 3.6C) on slope in the absence of effects on starting point (\(p > .05\)). Consistent with our drift-diffusion model fits, these results indicate that a confidence-induced confirmation bias is predominantly driven by a selective accumulation of choice-consistent information.

I further reasoned that this approach may remain blind to changes in the starting point of post-decision evidence accumulation because of an asymmetry in evidence availability at the start of the pre- and post-decision phases. In other words, simply reapplying the (non-predictive) classifier weights obtained at the beginning of the pre-decision phase to the same time point in the post-decision phase could render the analysis pipeline blind to starting point offsets. To address this concern, I evaluated the extent to which the entire timecourse of classifier predictions obtained in the pre-decision phase generalised to the post-decision phase, without making assumptions about their relative timing \([\text{King & Dehaene, 2014}].\) This analysis provides insight into how putative processing stages identified in the pre-decision phase are reinstated in the post-decision phase, and crucially how this timecourse is affected by confidence. I found a cluster of time points in which a representation of the initial
Figure 3.5: Outline of MEG analysis for quantifying accumulation of post-decision evidence at a neural level (MEG study 3, n=25 participants). **A** I trained a machine-learning classification algorithm on the pre-decision phase using MEG activity to predict left vs. right choices, and reapplied this classifier to the corresponding time point during the post-decision phase. The distance of each trial to the separating hyperplane provides a graded measure of neural evidence for a left or right decision, with changes in the classifier prediction within each trial providing a neural metric of evidence accumulation (see righthand panel). The inset shows the temporal generalization of decoding accuracy from the pre- to post-decision phases, indicating that the pre-decision classifier generalises to the post-decision phase along the major diagonal (i.e. corresponding time-points). AUC=area under the curve, DV = decision variable. **B** Grand average of the left/right classifier prediction in response to post-decision evidence. The light gray line shows the change in neural representation when rightward motion is presented and the black line shows the change in neural representation when leftward motion is presented. Regression lines show fits to the group-averaged data for visualisation purposes. Note that positive classifier values indicate evidence for a rightward decision and negative values evidence for a leftward decision. **C** Contributions of sensors to decoding left versus right decisions. The group average of contributions for each sensor is presented. In line with previous research on the neural correlates of evidence accumulation, sensors in centro-parietal regions made the highest contributions to decodability of (abstract) left versus right decisions. **D-F** Validation of neural metrics of post-decision evidence accumulation. Neural measures of the slope and starting point (intercept) of evidence accumulation extracted from the post-decision phase were entered as simultaneous predictors of D) reaction times E) accuracy and F) confidence of the final decision. Fixed effects from a hierarchical regression model are presented ± SEM. Hierarchical regression (two-tailed): D) **p=.005; E) *p=.045, **p=.002; F) **p=.002, ***p=.0004.
decision was activated earlier in the post- compared to the pre-decision phase when confidence was high (p=.01, corrected for multiple comparisons; Figure 3.6D). Such early reinstatement of a later processing stage is consistent with confidence enhancing a representation of the initial decision (i.e. shifting a starting point towards the bound of the initial decision) or inducing an expectation for evidence supporting an initial decision at the beginning of the post-decision period. Together these results indicate that confidence changes both the neural representation of evidence for an initial decision at the beginning of the post-decision phase (analogous to a change in starting point) as well as enhancing the processing of evidence supporting an initial decision (analogous to a change in drift rate).

3.4 Discussion

By combining behavioural and neural modelling I provide experimental evidence that holding high confidence in a decision leads to a striking modulation of post-decision processing and the emergence of a behavioural confirmation bias. These findings are consistent with a neural representation of confidence acting as a top-down controller (Atiya et al., 2019) that selectively amplifies processing of choice-consistent information.

A confirmation bias in the current experiment was observed in low-level perceptual decisions with limited emotional or cognitive content, suggesting that choice-induced biases in evidence accumulation represent a core principle of neural information processing (Luu & Stocker, 2018; Urai et al., 2019). In most real-world decisions, additional motivational (Taber et al., 2009) and social (Van Bavel & Pereira, 2018) influences (e.g. not revising a decision in order to appear self-consistent) are presumably also in play. These additional influences may amplify, or add to, effects of confidence on post-decisional processing in complex ways. An advantage of starting with an investigation of confirmation biases within lower-level tasks is that the potential for such interactions can be minimized, allowing a focused investigation of the processes that drive post-decisional shifts in evidence accumulation.

Computational modelling of the evidence accumulation process enabled further arbitration between apparently optimal information usage and a confirmation bias, by separating the influence of confidence on post-decisional starting point and drift rate. A shift in starting point is potentially normative as it may reflect the contribution of stronger pre-decision evidence to higher confidence, indicating that participants incorporate both pre- and post-decision evidence when reaching a final decision (see Figure 2.1). In contrast, the influence confidence on drift-rate represents a distortion in the integration of new evidence and thus a classic instance of confirmation bias.

I focussed on effects of the initial decision and initial confidence on the starting
Figure 3.6: MEG analysis investigating the influence of confidence on post-decision evidence processing (MEG study 3, n=25 participants). A-B Neural metrics of post-decision accumulation separated into confirming (consistent with initial decision) and disconfirming (inconsistent with initial decision) post-decision evidence and as a function of A) high and B) low initial confidence. More positive values on the y-axis indicate stronger (more veridical) representation of the presented motion. Weighted group averages (grand average) are presented and regression lines are fits to this averaged data. C Effects of initial decision and confidence on the slope of neural evidence accumulation in response to post-decision evidence (slope). The righthand panel shows weighted mean values ± SEM for the strength of neural evidence integration (slope) within each condition. Grey dots represent individual participants’ data. The left-hand bar shows the fixed effect ± SEM for the initial decision × confidence interaction effect from a hierarchical regression (two-tailed): **p=.008. D Effect of confidence on temporal generalization of decoding accuracy from the pre- to the post-decision phase. Higher confidence is associated with higher decodability of the initial decision (i.e. stronger representation of the initial decision, yellow colours). A stronger representation of the initial decision was seen at the beginning of the post-decision period when confidence was high, consistent with confidence shifting a starting point towards the bound of the initial decision. The contoured area represents a cluster of timepoints with a significant main effect of confidence (permutation test, p <.05 corrected for multiple comparisons). The time window starts with stimulus presentation (0ms) and ends when the response options are presented (850ms). Dotted lines indicate the offset of the stimulus (pre- or post-decision stimulus respectively).
point and the drift-rate of the post-decision accumulation process. An alternative interpretation of the results could be that instead of changing the weighting of new information, people might opt to not accumulate new information all together when highly confident. Such an effect would manifest as a generally reduced drift-rate after high confidence decisions, i.e. a main effect of confidence on the drift-rate. This effect was incorporated in all the tested models (see Figure 3.4E, left panel) and did not show any significant results, suggesting this alternative is less likely.

Previous literature has suggested that similar behavioural effects to those of a confirmation bias could be modelled by a time-dependent collapse of the decision bound \cite{Hanks2011}. Importantly, all of our models incorporated a dependency of the decision bound on initial confidence which rules out a simple, symmetric adjustment accounting for the behavioural effects of a confidence-induced confirmation bias. However, considering more sophisticated and asymmetric dependencies of a time-dependent collapse of decision bounds on initial confidence could in principle account for some of the data features. If the decision bound of the chosen option collapsed more quickly when participants are highly confident (i.e. confidence inducing an urgency signal for choosing the same option again), this would be mathematically equivalent to a choice-dependent increase in drift-rate based on confidence \cite{Hanks2011}. Therefore, behavioural modelling along cannot disentangle these different views. However, the neural data complements the behavioural analysis and can shed light on this issue. If confidence caused a collapsing bound (rather than an amplification of choice-consistent evidence), one would assume that neural evidence for confirming information would be lower after high confidence (compared to low confidence) decisions. However, the neural data (see Figure 3.6A) shows the opposite pattern, with enhanced neural representations of confirming evidence after high confidence decisions, indicating that confidence induces a selective weighting for choice-consistent information rather than a time dependent adjustment of the decision bound.

In turn, the usage of MEG recordings in combination with machine learning classification revealed a neural marker of these shifts in post-decision evidence accumulation. This measure complemented my behavioural modelling results and yielded direct support for a hypothesis that confidence alters the way in which the brain accumulates new information, consistent with a selective gating of choice-consistent information. In the current task, where new evidence is always helpful, this bias against incorporating conflicting post-decision evidence is normatively maladaptive. In other scenarios, however, where new evidence may be distracting and/or actively misleading, a confirmation bias might prove helpful. For instance, previous attempts to explain the value of selective evidence accumulation focused on its role in directing attention towards aspects of the environment with the highest potential for information gain \cite{Cheadle2014, Parr2018}, or in increasing
the robustness of decisions against the influence of noise (Qiu et al. 2019, Tsetsos et al. 2016). However, the fact that confidence increases choice-consistent information processing goes against the idea that confirmation bias is itself driven by a need for certainty (Nickerson 1998, Skov & Sherman 1986). Instead, I observed the strongest confirmation bias when people were already confident in their decisions.

The study of cognitive biases has remained largely distinct from parallel efforts to understand the processes governing evidence accumulation in simple decisions. I suggest that extending models of evidence accumulation to post-decisional processing enables a unique window onto biases in higher-order cognition (Talluri et al. 2018). For instance, I will show in chapter 4 that alterations in post-decision processing are predictive of higher-level attitudes such as beliefs about political issues (Rollwage et al. 2018), suggesting that insights gained from the study of confirmation bias in simple decisions can be applied to understand the drivers of polarization and entrenchment across a range of societal issues. For instance, a central role for confidence in shaping the fidelity of evidence accumulation indicates that metacognitive interventions may be one route towards ameliorating this pervasive cognitive, as I will show in chapter 7.

Since all results in this chapter were obtained in the context of perceptual decision-making tasks, in the next chapters I will focus on the link between altered post-decision processing in such simple tasks and real-world beliefs. Specifically, I will focus on radical and dogmatic political beliefs and their link to cognitive changes such as confirmation bias.
Chapter 4

Metacognitive failure as a feature of those holding radical beliefs

In the previous chapter I investigated the cognitive, computational and neural mechanisms driving confirmation bias in a perceptual task. However, confirmation bias has received most attention regarding its potential influence on societal attitudes (Sunstein et al., 2016; Nickerson, 1998; Lord et al., 1979). For instance, this cognitive bias has been suggested as a critical driver of societal polarization and entrenchment of belief (Lilienfeld et al., 2009). Supporting this, it has been shown that a general resistance to new information is associated with extreme and dogmatic beliefs (Zmigrod et al., 2019a,b,c; Zmigrod, 2020). However, the underlying computational mechanisms driving these rigid beliefs have not been explored. In this and the next chapter, behavioral tasks in conjunction with computational modelling will be used to investigate the cognitive underpinnings of radical and dogmatic beliefs. In chapter 4, I focus on metacognitive abilities and the integration of disconfirming post-decision evidence (i.e. confirmation bias) in relation to radical beliefs. In chapter 5, I extend this research agenda to investigate a link between active information search and dogmatism.

4.1 Introduction

Widening polarization about political, religious and scientific issues threatens open societies, leading to entrenchment of beliefs, reduced mutual understanding and a pervasive negativity surrounding the very idea of consensus (Kohut et al., 2012; Iyengar & Westwood, 2015). An unjustified certainty in one’s beliefs is a characteristic common to those espousing radical beliefs (Ortoleva & Snowberg, 2015; Toner et al., 2013; Brandt et al., 2015; Greenberg & Jonas, 2003), and such overconfidence is observed for both political and non-political issues (Ortoleva & Snowberg, 2015; Toner et al., 2013; Brandt et al., 2015), implying a general cognitive bias in radicals. However, the underpinnings of radicals’ distorted confidence estimates remain unknown.
In particular, one-shot measures of the discrepancy between performance and confidence are unable to disentangle the contributions of confidence bias (changes in an overall belief about performance, which may be affected by optimism [Ais et al., 2016] and mood [Rouault et al., 2018]) from changes in metacognitive sensitivity (an ability to distinguish accurate from inaccurate performance; [Fleming & Lau, 2014]).

This distinction may be particularly important, as changes in metacognitive sensitivity may account for radicals’ reluctance to change their mind in the face of new evidence. Decision neuroscience has highlighted that metacognitive sensitivity depends on mechanisms that facilitate monitoring and revision of confidence in previous choices [Van Den Berg et al., 2016; Fleming et al., 2018]. This ability relies on specific neural circuitry [Fleming et al., 2012] in the prefrontal cortex that promotes reflection on one’s performance and, even in the absence of explicit feedback, a realization that mistakes have been made [Rabbitt, 1966; Yeung & Summerfield, 2012]. It has generally been assumed that a resistance of radicals to change their beliefs is due to social and motivational factors, such as the desire to maintain a positive self-image [Nyhan & Reifler, 2010; Kaplan et al., 2016; Taber et al., 2009; Redlawsk, 2002], whereas the role of metacognitive capacities has received less attention. However, changes of mind depend not only on a motivation to change but also on a (metacognitive) capacity to realize that one’s beliefs are wrong.

By employing simple perceptual discrimination tasks, it is possible to precisely quantify metacognitive sensitivity – the extent to which people’s confidence judgments are sensitive to task performance – and to disentangle metacognitive sensitivity from overconfidence bias [Fleming & Lau, 2014]. Such tasks provide an objectively correct answer (which is rarely the case for direct assays of political attitudes where the ground truth is often unknown or unavailable), thus enabling a precise, quantitative and objective measure of metacognitive ability as well as a normative prediction for changes of confidence in light of new evidence. Moreover, the usage of a perceptual task makes it unlikely that participants have a priori vested interests in a particular decision outcome, thus diminishing any strong link to participants’ self-concept and providing an assay of the relationship between domain-general metacognitive abilities and radicalism. Here I test a hypothesis that limitations in metacognitive sensitivity lead to a resistance to belief change, even when motivational factors are minimized, and that such metacognitive limitations are associated with the entrenched beliefs that are exemplified by radicals.

To typify a spectrum of radical views I first conducted a separate online survey of 344 US participants (Study 1) who completed questionnaires about political issues [Toner et al., 2013; Everett, 2013; Funke, 2005; Van Hiel et al., 2006; Altemeyer, 2002]. I included standard questionnaires about political orientation, voting behaviour, attitudes towards specific political issues, intolerance of opposing po-
political attitudes, belief rigidity and (left- and right-wing) authoritarianism. These questionnaires were selected based on prior models of political radicalism as stemming from a combination of intolerance to others’ viewpoints, dogmatic/rigid beliefs, and authoritarianism, which represents adherence to ingroup authorities and conventions, and aggression in relation to deviance from these norms (Rokeach 1960; Van Hiel 2012; Eysenck 1968). However, I stress that radicalism is likely to reflect a general cognitive style that transcends the political domain – as exemplified by links between religious fundamentalism and increased dogmatism/authoritarianism (Altemeyer 2002; Altemeyer & Hunsberger 1992) – and instead refers to how one’s beliefs are held and acted upon (Snow & Cross 2011).

In Studies 2 and 3, I then go on and investigate a link between these radical beliefs and cognitive characteristics related to information processing.

4.2 Methods

4.2.1 Participants

All three studies were conducted online and recruited subjects from the US via the online labour market Amazon Mechanical Turk (Buhrmester et al. 2016). Mechanical Turk has been shown to be more representative of the population than typical college student samples or other online recruitment systems (Huff & Tingley 2015; Paolacci et al. 2010), and produces high quality data (Buhrmester et al. 2016) with good internal and external validity (Horton et al. 2011; Rand 2012), even when using complex behavioural tasks (Crump et al. 2013). All data were collected in the year 2017. Subjects gave informed consent and the study was approved by the Research Ethics Committee of University College London (1260-003).

In Study 1 subjects completed questionnaires about political issues (see below). In Studies 2 and 3, participants filled out the same questionnaires as in Study 1 and additionally completed two perceptual decision-making tasks (a confidence task and a post-decision evidence integration task; see below).

In Study 1 I analysed data from 344 subjects (46% women, mean age 34.9 years, range 19-73 years). In Study 2 a sample of 381 subjects were included for analysis of questionnaires and behavioural tasks (51.4% women, mean age 36.0 years, range 19-70 years). In Study 3 a sample of 417 subjects was included for analysis of questionnaires and behavioural tasks (47% women, mean age 35.8 years, range 18-71 years). The sample size of Study 3 (N=417) was defined by an a priori power analysis based on effect sizes from Study 2 for associations between radicalism and meta-d’ (power=77%) or disconfirmatory evidence integration (power=89%).

Self-reported political orientation (“very liberal”=0 to “very conservative”=100) in both samples was somewhat skewed to the liberal end of the spectrum (Study 1: mean = 38.1; Study 2: mean =38.0; Study 3: mean =41.9) which is in line with
previous reports of Amazon Mechanical Turk workers being more liberal than the general US population [Huff & Tingley, 2015]. However, there remained substantial variability in political orientation as measured via factor analysis.

4.2.2 Stimuli

Perceptual discrimination experiments were programmed in JavaScript using JsPsych (version 5.0.3) and hosted on the online research platform Gorilla (https://gorilla.sc/). The experiment was accessed via a web browser and participants were required to use full-screen mode to complete the task. Stimuli consisted of two black squares (each 250 × 250 pixels) centrally positioned on the screen, one square to the left and the other square to the right of centre. These squares were subdivided into grids of 625 cells, randomly filled with white dots. One square always contained 313 cells filled with dots and the other square contained a greater number of filled cells (the exact number of additional dots was adjusted for each individual using a staircase procedure). The difference in dot number between the two squares determined the judgment difficulty. Five such configurations were presented per trial, each for 150ms, creating the impression of flickering dots. The exact location of dots per configuration was random, but within one trial, the difference in number of dots and also the side which contained more dots remained constant.

4.2.3 Task and procedure

For Studies 2 and 3, the experiment lasted around 1 hour. After receiving general information and instructions, participants began the behavioural experiment which was divided into 3 parts. First, participants completed 120 trials of a calibration phase (which was used to individually adjust the task difficulty, see below), in which they performed the perceptual judgement by reporting whether the left or the right square contained more dots. Second, there were 60 trials of the same perceptual judgement followed by a confidence rating (“confidence task”, Task 1). Finally, there were 120 trials of a “post-decision evidence integration task” (Task 2), in which subjects performed the perceptual judgement, received additional post-decision evidence and then rated their confidence. After completing the behavioural tasks, participants filled out questionnaires regarding their political orientation and radicalism.

4.2.3.1 Calibration Phase

Before performing the main task, each participant performed a calibration phase comprising 120 trials judging whether the left or the right square contained more dots without confidence ratings, using their computer keyboard (using the “W” and “E” keys to indicate left and right, respectively). Responses were unspeeded
and possible only after stimulus offset. During the calibration phase (but not the experimental phase), visual feedback was delivered to indicate whether the judgment was correct (a green frame around the chosen square) or incorrect (a red frame around the chosen square). The calibration phase was used to find a stimulus strength (dot difference between left and right) for each participant that elicited approximately 71% correct performance (actual performance: mean = 73.2%, sd = 6.3%) in the discrimination task using a 2-down-1-up staircase procedure (García-Pérez, 1998) operating on the logarithm of dot difference. Participants completed 70 trials of the staircase and the average of the last 25 trials was stored and used as the individual stimulus strength throughout the rest of the experiment. For the post-decision evidence integration task I also presented stronger evidence than the staircased stimulus strength on a subset of trials. This stronger level of evidence was generated by multiplying the logarithm of the staircased dot difference by a factor of 1.3. To quantify the performance level induced by this stimulus strength for each individual, participants performed 50 additional perceptual judgements at this higher stimulus strength interleaved with the staircased trials (after 20 initial “burn-in” trials to allow the staircase to converge) and yoked to the current staircase value. In the group as a whole, the higher evidence strength evoked performance levels of approximately 80% correct (sd = 3.8%).

4.2.3.2 Confidence task (Task 1)

The confidence task comprised a total of 60 trials, all at the same (lower) stimulus strength determined in the calibration phase. On each trial, participants judged which side contained more dots, before rating confidence in their decision (a detailed trial timeline is displayed in Figure 4.2A). Participants were instructed to report their confidence as a subjective probability that their decision was correct, rated on a 9-point sliding scale. The scale midpoint was labelled with 50%, the lowest category with 0% and the highest category with 100%. Confidence ratings were incentivized using the quadratic scoring rule (Brier, 1950).

4.2.3.3 Post-decision evidence integration task (Task 2)

The post-decision evidence integration task consisted of 120 trials, split into 60 trials with low post-decision evidence strength and 60 trials with high post-decision evidence strength, pseudo-randomly interleaved. Within each trial, participants first judged which side contained more dots as described under “Confidence task” above. After this initial decision, they received additional evidence. The location of higher dot density in the post-decision evidence presentation was always of the same (correct) sign as the pre-decision evidence presentation, but of variable strength. Subjects were instructed that this evidence was “bonus” information that could be used to inform their confidence in their initial response. The post-decision evidence
could either have the same strength as the pre-decision evidence (low post-decision evidence) or have a higher evidence strength (high post-decision evidence). After the presentation of post-decision evidence, participants were asked to indicate their confidence in their initial decision.

4.2.4 Data quality and exclusion criteria

In Study 1 I analysed questionnaire data from 344 subjects. Six subjects were excluded from the original sample (N=350) because they failed to answer correctly at least one of two catch questions interspersed within the questionnaires (“If you have read this question please choose Agree Completely” and “Please choose Disagree completely if you read this question”). In Study 2 a sample of 381 subjects were included for analysis of questionnaires and behavioural tasks. To ensure data quality I excluded 123 subjects (original sample N=504) based on a range of pre-defined exclusion criteria. First, 5 subjects were excluded from questionnaire analysis due to answering at least one of the two catch questions incorrectly, as described above. An additional 77 subjects were excluded due to performance in the perceptual decision-task being above 85% or below 60% correct, indicating that the staircase procedure was insufficient to produce threshold performance. Five subjects were excluded because they chose a single confidence rating more than 90% of the time and an additional 10 subjects were excluded due to median confidence response times of below 850 ms, indicating that they rated their confidence very quickly and possibly without care. Finally, 26 subjects were excluded due to a large proportion of missed trials (>5%).

In Study 3 a sample of 417 subjects was included for analysis of questionnaires and behavioural tasks. As in Study 2, I excluded 158 subjects (original sample N=575) based on the same pre-defined exclusion criteria. Seventeen subjects were excluded from questionnaire analysis due to answering at least one of the two catch questions incorrectly. An additional 90 subjects were excluded due to performance in the perceptual decision task being above 85% or below 60%, indicating that the staircase procedure was insufficient to produce threshold performance. 11 subjects were excluded because they chose a single confidence rating more than 90% of the time, and an additional 19 subjects were excluded for median confidence reaction times below 850 ms. Finally, 21 subjects were excluded due to a significant proportion of missed trials (>5%). Exclusion criteria followed similar procedures used in our lab (Rouault et al., 2018) and elsewhere (Oppenheimer et al., 2009). The overall exclusion rate (25%) was similar to other studies from our lab and was consistent with a recent meta-analysis which found that between 3% and 37% of the sample is typically excluded in web-based experiments (Chandler et al., 2014).

I applied these exclusion criteria to ensure high-quality data and prevent my results being influenced by people performing the task without care. However, I also
established that results were qualitatively similar in the absence of exclusions, with the composite measure of radicalism remaining associated with impaired metacognitive sensitivity (Study 2: $\beta=-.13$, $p=.008$; Study 3: $\beta=-.12$, $p=.01$) and reduced disconfirmatory evidence integration (Study 2: $\beta=-.22$, $p=.0002$; Study 3: $\beta=-.14$, $p=.009$).

4.2.5 Behavioral Analysis

4.2.5.1 Measurement of metacognitive ability

For assessment of metacognitive ability I calculated $meta - d'$ (Maniscalco & Lau, 2012), a signal detection theoretic measure of metacognitive sensitivity that is uncorrupted by the tendency to report high or low confidence (overconfidence bias (Fleming & Lau, 2014)). To estimate $meta - d'$ I employed a Bayesian estimation scheme (Fleming, 2017) (HMeta-d; https://github.com/smfleming/HMeta-d), using the non-hierarchical version of the model.

4.2.5.2 Measurement of post-decision evidence integration

I measured confirmatory and disconfirmatory evidence integration as changes in confidence induced by post-decision evidence. I constructed trial-by-trial linear models for every participant, separately for correct and incorrect trials across data pooled across Tasks 1 and 2, using post-decision evidence strength as a predictor (confidence task=0, low post-decision evidence=1, high post-decision evidence=2) and confidence ratings as the dependent variable. Individual beta weights for correct trials, indicating increases of confidence due to post-decision evidence, were estimated as measures of confirmatory evidence integration. Disconfirmatory evidence integration was estimated as the beta weight on incorrect trials (I reversed the sign of this beta weight in the figures such that higher values indicate greater disconfirmatory evidence integration).

I additionally tested whether sensitivity to post-decision evidence could be predicted from metacognitive sensitivity measured in Task 1. For this purpose, I calculated sensitivity to post-decision evidence based solely on trials from Task 2 to ensure independence from estimation of metacognitive ability in Task 1. For each subject I constructed a trial-by-trial linear model with confidence as dependent variable, entering the following predictors: accuracy (correct=1 and incorrect=-1), post-decision evidence strength (low post-decision evidence=1 and high post-decision evidence=2) and the critical accuracy $\times$ post-decision evidence strength interaction. This interaction term quantifies the extent to which confidence increases on correct trials and decreases on error trials at higher levels of post-decision evidence strength, thus forming a summary measure of sensitivity to additional evidence.
4.2.6 Factor analysis

There is extant debate about the underlying structure of political ideology and its relation to radical beliefs (Jost et al., 2009). Therefore, instead of relying on direct self-report measures of political orientation and radicalism, I combined multiple questionnaires related to political orientation, intolerance, dogmatism and authoritarianism, and conducted a factor analysis to identify the most parsimonious factor structure. Moreover, the factor analysis ensured that dogmatic, rigid and intolerant beliefs represented a separate factor that is distinguishable from political orientation. Regarding political orientation, I included questions that reflect putatively separate facets of social and economic conservatism (Jost et al., 2009). Participants filled out questions about the following issues: political orientation on a “liberal-conservative” dimension (general conservatism and separately for social and economic issues), voting behaviour and identification with the U.S. Democratic or Republican party, a social and economic conservatism scale (Everett, 2013), and attitudes towards specific political issues (Toner et al., 2013).

To measure dogmatism I employed a widely used scale that assays this construct (Altemeyer, 2002). Additionally, I administered previously used questions (Toner et al., 2013) about belief superiority and intolerance of opposing political opinions as these are known to show considerable conceptual overlap with dogmatism and have previously been reported as manifesting a quadratic relationship with political orientation (Toner et al., 2013). Authoritarianism is widely conceptualized as prevalent on the right side of the political spectrum together with more controversial proposals of a similar trait in left-wing individuals (Greenberg & Jonas, 2003; Van Hiel et al., 2006). Left-wing authoritarianism may be rarely reported due to problems with measurement (right-wing authoritarianism scales are not content-free but target conservative tendencies) or sample characteristics. To counteract such concerns here I included both left-wing (Van Hiel et al., 2006) and right-wing (Funke, 2005) authoritarianism scales.

An exploratory factor analysis was conducted on all 78 single questionnaire items using maximum likelihood estimation. Factor analysis was conducted using the fa() function from the Psych package in R, with an oblique rotation (oblimin). The number of factors was extracted based on the Cattell-Nelson-Gorsuch (Gorsuch, 1983) test using the nFactors package in R. The Cattell-Nelson-Gorsuch test revealed a three-factor solution as the best and most parsimonious solution for the covariance structure of the single items (see Figure 4.1A for the factor loadings of individual items). The pattern of factor loadings was qualitatively similar for both Study 1 and the combined data from all three Studies. To obtain precise estimates of factor loadings and thus more reliable factor scores I conducted the factor analysis on the pooled sample from all three studies when extracting factor scores for use in analysis of behavioural data in Studies 2 and 3.
Comrey & Lee (2013) have suggested a general rule of thumb that 300 participants are a large enough sample size for conducting a factor analysis. In comparison, Mundfrom et al. (2005) used simulation approaches to establish that a sample of around 100 participants is enough for deriving 3 factors of interest. To ensure that my sample size was large enough for conducting a factor analysis, I included more than 300 participants and combined samples across studies where possible to increase the accuracy of the derived factor structures.

4.2.7 Computational modelling

4.2.7.1 Modelling post-decision evidence integration

All computational models were adapted from those developed by Fleming et al. (2018) in a study of post-decision evidence integration during RDK decisions. I examined the potential of these models to explain individual differences in metacognitive sensitivity and confidence updating based on post-decision evidence observed in relation to radicalism. All models were grounded in signal detection theory, and simultaneously modelled choices and confidence ratings of Tasks 1 and 2. In the confidence task, subjects receive one internal sample, $X_{\text{pre}}$ generated from pre-decision evidence, whereas for the post-decision evidence integration task subjects receive two internal samples, $X_{\text{pre}}$ from pre-decision evidence and $X_{\text{post}}$ from post-decision evidence. These samples in turn were generated from a Gaussian whose sign depended on the location of objectively higher dot density (left=-1, right =1) and mean on internal evidence strength $\theta_{\text{pre}}$ or $\theta_{\text{post}}$:

\[
X_{\text{pre}} \sim N (d\theta_{\text{pre}}, 1) \tag{4.1}
\]

\[
X_{\text{post}} \sim N (d\theta_{\text{post}}, 1) \tag{4.2}
\]

The internal evidence strength depended on the dot difference and was always the same for $\theta_{\text{pre}} (\mu_{\text{low}})$, whereas the evidence strength could be either low or high for $\theta_{\text{post}} \in [\mu_{\text{low}}, \mu_{\text{high}}]$, where $\mu_{\text{low}}$ and $\mu_{\text{high}}$ are free parameters. The likelihood of $X_{\text{pre}}$ or $X_{\text{post}}$ was approximated by a Gaussian with mean $\mu$ and variance $\sigma^2$. For the confidence task (Task 1) in which only $X_{\text{pre}}$ was presented, I set $\mu = \theta_{\text{pre}}$ and $\sigma^2 = 1$. For the post-decision evidence task (Task 2), I approximated the likelihood of both $X_{\text{pre}}$ and $X_{\text{post}}$ as a single Gaussian with mean $\mu$ and variance $\sigma^2$ determined by a mixture of Gaussians across the two possible evidence strengths. Starting with $X_{\text{post}}$:

\[
P (X_{\text{post}}|d = 1) = \sum_{\theta_{\text{post}}} p (\theta_{\text{post}}) N (\theta_{\text{post}}, 1) \tag{4.3}
\]

As each of the two evidence strengths is equally likely by design one can define the mean as:

\[
\mu = \frac{\sum \theta_{\text{post}}}{2} \tag{4.4}
\]
Chapter 4. Metacognitive failure as a feature of those holding radical beliefs

The aggregate variance \( \sigma^2 \) can be decomposed into both between- and within condition variance. From the law of total variance:

\[
\sigma^2 = \sum_{\theta_{\text{post}}} p(\theta_{\text{post}}) [E[X_{\text{post}}|\theta_{\text{post}}] - \mu]^2 + \sum_{\theta_{\text{post}}} p(\theta_{\text{post}}) \text{Var}(X_{\text{post}}|\theta_{\text{post}}) \tag{4.5}
\]

\[
\sigma^2 = \sum_{\theta_{\text{post}}} p(\theta_{\text{post}}) [E[X_{\text{post}}|\theta_{\text{post}}] - \mu]^2 + 1 \tag{4.6}
\]

I assume that subjects are agnostic about the set of evidence strengths presented before and after the decision, such that \( \mu \) and \( \sigma^2 \) are the same for both \( X_{\text{pre}} \) and \( X_{\text{post}} \).

In both tasks, actions \( a \) are made by comparing \( X_{\text{pre}} \) to a criterion parameter \( m \) that accommodates any stimulus-independent biases towards the leftward or rightward response, \( a = X_{\text{pre}} > m \). Each piece of evidence, \( X_{\text{pre}} \) and \( X_{\text{post}} \), updates the log posterior odds of the rightward location containing a higher dot density, \( LO_{\text{dir}} \), which under flat priors is equal to the log likelihood:

\[
LO_{\text{dir}}^{\text{pre}} = \log \frac{P(d = 1|X_{\text{pre}})}{P(d = -1|X_{\text{pre}})} = \log \frac{P(X_{\text{pre}}|d = 1)}{P(X_{\text{pre}}|d = -1)} \tag{4.7}
\]

\[
LO_{\text{dir}}^{\text{post}} = \log \frac{P(d = 1|X_{\text{post}})}{P(d = -1|X_{\text{post}})} = \log \frac{P(X_{\text{post}}|d = 1)}{P(X_{\text{post}}|d = -1)} \tag{4.8}
\]

where, due to the Gaussian generative model for \( X \), \( LO_{\text{dir}} \) is equal to:

\[
LO_{\text{dir}} = \log \frac{e^{(\mu + X)^2 / 2\sigma^2}}{e^{(\mu - X)^2 / 2\sigma^2}} \tag{4.9}
\]

Positive values indicate greater belief in the higher dot density being on the right; negative values indicate greater belief in higher density on the left. To update confidence in one’s choice, the belief in dot density (\( LO_{\text{dir}} \)) is transformed into a belief about decision accuracy (\( LO_{\text{correct}} \)) conditional on the chosen action:

If \( a=1 \):

\[
LO_{\text{correct}} = LO_{\text{dir}} \tag{4.10}
\]

Otherwise:

\[
LO_{\text{correct}} = -LO_{\text{dir}} \tag{4.11}
\]

As for \( LO_{\text{dir}} \), \( LO_{\text{correct}} \) in the post-decision evidence task can be decomposed into pre- and post-decisional parts:

\[
LO_{\text{correct}}^{\text{total}} = LO_{\text{correct}}^{\text{pre}} + LO_{\text{correct}}^{\text{post}} \tag{4.12}
\]

For trials of the confidence task, \( LO^\text{post} \) was set to zero for all models as in those trials no post-decision evidence was presented. The final log odds correct is then transformed to a probability to generate a confidence rating on a 0-1 scale:

\[
\text{Confidence} = \frac{1}{1 + \exp \left(-LO_{\text{correct}}^{\text{total}}\right)} \tag{4.13}
\]
Model extensions accounting for differences in post-decision evidence integration

I considered different mechanisms that could account for reduced metacognitive sensitivity and changes of mind in radicals, adapted from Fleming et al. (2018) and Bronfman et al. (2015).

4.2.7.2.1 Temporal weighting

I considered participants may apply differential weighting to pre- and post-decision evidence when computing confidence. The “temporal weighting” model captured such differences via two free parameters ($w_{pre}$ and $w_{post}$) as follows:

$$LO_{total\ correct} = w_{pre} \times LO_{pre\ correct} + w_{post} \times LO_{post\ correct}$$ (4.14)

4.2.7.2.2 Choice weighting

An alternative model applies differential weighting to post-decision evidence depending on whether this evidence is in support of the chosen option (confirmatory) or unchosen option (disconfirmatory), as found in Talluri et al. (2018). To capture this effect I introduced two separate weighting parameters based on the correspondence between the decision and post-decision evidence:

If $\text{sign}(X_{post}) = \text{sign}(a)$:

$$LO_{total\ correct} = LO_{pre\ correct} + w_{confirmatory} \times LO_{post\ correct}$$ (4.15)

Otherwise:

$$LO_{total\ correct} = LO_{pre\ correct} + w_{disconfirmatory} \times LO_{post\ correct}$$ (4.16)

4.2.7.2.3 Choice bias

Finally I considered a model in which subjects become more confident in the option they chose, irrespective of the strength of post-decision evidence. This was implemented by adding a fixed amount of subjective probability to the chosen option which was controlled by a free parameter $w_{bias}$:

If $a = 1$:

$$LO_{bias} = \log \left( \frac{w_{bias}}{1 - w_{bias}} \right)$$ (4.17)

Otherwise:

$$LO_{bias} = \log \left( \frac{1 - w_{bias}}{w_{bias}} \right)$$ (4.18)

$$LO_{total\ dir} = LO_{pre\ dir} + LO_{post\ dir} + LO_{bias}$$ (4.19)

Note that the choice bias term (unlike the weighting parameters in alternative model extensions) also affects the predictions for the confidence task as it is applied independently of the level of post-decision evidence.
4.2.7.3 Model fitting

I used variational Bayesian inference implemented in STAN [Kucukelbir et al., 2015] to approximate draws from the posterior distribution of parameters given the world state $d$, subjects’ choices $a$ and their confidence ratings $r$. Since there were relatively few trials per subject, I used a hierarchical fitting procedure. I set the maximum number of iterations to 150,000 and a convergence tolerance on the relative norm of the objective to .0001 (this is a relative conservative approach regarding convergence; default options in STAN are 10,000 iterations and a convergence tolerance of .01). From the approximate posterior, 1000 samples were drawn for each of the parameters.

4.2.7.4 Model comparison

To compare between alternative models I assessed their ability to capture individual differences in radicalism. To this end, I constructed separate multiple regressions to predict each participants radicalism scores from the mean of the posterior draws of each model’s fitted parameters for each individual $j$. For each model I inputted $\mu_{\text{low},j}$ and $\mu_{\text{high},j}$ as predictors together with model-specific parameters (choice bias: $w_{\text{bias},j}$), weighting parameters for confirmatory and disconfirmatory evidence ($w_{\text{confirmatory},j}$ and $w_{\text{disconfirmatory},j}$) or weighting parameters for pre and post-decision evidence ($w_{\text{pre},j}$ and $w_{\text{post},j}$). I computed BIC scores for each multiple regression to identify model fits that best explained individual difference in radicalism. Note that this model comparison approach differs from standard approaches in that it is concerned with best capturing individual differences rather than an aggregate fit to the group.

Since the BIC score includes a penalty for model complexity (i.e. number of parameters), I wished to ensure that the choice bias model was not favoured due to its lower complexity alone. I therefore also considered multiple regressions that included only one of the fitted parameters from the more complex temporal weighting and choice weighting models ($w_{\text{confirmatory},j}$ or $w_{\text{disconfirmatory},j}$; $w_{\text{pre},j}$ or $w_{\text{post},j}$) as predictors and included these variants in the model comparison. The parameter combinations with the lowest BIC scores are presented in Figure 4.4A.

4.2.7.5 Model simulations

To visualize qualitative features of computational model fits and determine their ability to account for the patterns of confidence ratings in moderates and radicals (a posterior predictive check, see Figure 4.4B), I drew 100 samples from the posterior distributions of fitted parameters for each participant, and for each draw simulated 4000 trials per subject per condition (confidence task, low and high post-decision evidence) with these parameter settings.
4.2.8 Statistical analysis

In all regression analyses I employed robust fits by using the default robust option of the Matlab function `fitlm` which applies a “bisquare” weighting. \( R^2 \)-values for each predictor were calculated by comparing the explained variance of the full model including this predictor to a model excluding the predictor of interest. All effects for Studies 1 and 2 were tested two-tailed. Since I had strong \textit{a priori} hypotheses in the replication sample, effects in Study 3 were tested one-tailed based on the directional hypothesis derived from Study 2.

The following regression analyses were conducted:

1. To investigate the relation between political orientation, dogmatic intolerance and authoritarianism, I specified separate models with dogmatic intolerance and authoritarianism as dependent variables and political orientation as a predictor in a second-order polynomial regression, with both linear and quadratic terms for the predictor. The relationship of political orientation with dogmatic intolerance and authoritarianism was labelled as linear or quadratic via comparison of BIC scores of models with either or both terms.

2. To quantify the link between metacognitive sensitivity (measured in the confidence task) and sensitivity to post-decision evidence I conducted a multiple regression analysis with post-decision evidence sensitivity as the dependent variable and separate predictors for \textit{meta} – \( d' \) in Task 1, perceptual task performance (\( d' \)) averaged across Tasks 1 and 2, confidence bias in Task 1, objective evidence strength (logarithm of dot difference) and performance at higher post-decision evidence strength (as measured in the calibration phase).

3. To investigate the relation between metacognitive function and radicalism I implemented separate multiple regression models with the factor scores (dogmatic intolerance, authoritarianism and political orientation) as dependent variables and separate predictors for \textit{meta} – \( d' \) in Task 1, confidence bias in Task 1, confirmatory evidence integration in Task 2, disconfirmatory evidence integration in Task 2, perceptual task performance (\( d' \)) averaged across Tasks 1 and 2, objective stimulus strength (logarithm of dot difference), performance at the higher post-decision evidence strength (as measured in the calibration phase), age, gender and education.

4. Finally, to investigate whether radicalism was associated with reduced earnings in the task, I constructed a multiple regression model with earnings as dependent variable and radicalism as predictor, controlling for perceptual task performance (\( d' \)) averaged across Tasks 1 and 2 and objective stimulus strength (logarithm of dot difference).
4.3 Results

4.3.1 Factor analytic description of political beliefs

A factor analysis of individual items identified three latent factors (Figure 4.1A) which I labelled “political orientation”, “dogmatic intolerance” and “authoritarianism”. Together these three factors explained 40% of the variance in questionnaire responses.

The first factor tracked political orientation (liberal to conservative) as indicated by the highest loading item “Please rate your overall political attitude on the dimension from liberal to conservative” (factor loading = .90). The second factor loaded predominantly on the dogmatism and intolerance questionnaires, with the highest loading items concerning rigid and dogmatic world views, e.g. “My opinions are right and will stand the test of time” (factor loading = .73) and intolerance of opposing political beliefs, e.g. “My beliefs about the government’s role in helping people in need are totally correct (mine is the only correct view)” (factor loading = .58). The third factor was related to authoritarianism and showed the highest loadings on questions related to obedience to in-group authorities, e.g. “A revolutionary movement is justified in demanding obedience and conformity of its members” (factor loading = .43), group conventions, e.g. “The withdrawal from tradition will turn out to be a fatal fault one day” (factor loading = .37) and support of aggression to reach one’s political goals, e.g. “What my country really needs is a strong, determined Chancellor which will crush the evil and set us on my right way again” (factor loading = .4).

In what follows I focus on the dogmatic intolerance and authoritarianism factor scores as summary indices of radicalism \cite{Rokeach1960, VanHiel2012, Eysenck1968}.

Notably, a clear quadratic relationship was evident between political orientation and dogmatic intolerance ($\beta = .37$, $p<10^{-11}$, Figure 4.1B), indicating that both the far left and far right of the political spectrum hold similarly intolerant and rigid beliefs, replicating previous findings \cite{vanProoijenKrouwel2017}. On the other hand, a linear relationship of authoritarianism with political orientation was found ($\beta = .38$, $p<10^{-11}$, Figure 4.1C), showing that those on the right of the political spectrum displayed higher levels of authoritarianism, also as reported previously \cite{AltemeyerAltemeyer1996}. Finally, dogmatic intolerance and authoritarianism were positively correlated ($\beta = .21$, $p<.0001$, Figure 4.1D).
Using factor analysis I investigated the underlying factor structure of multiple questionnaires about political issues. Three latent factors were identified and labelled “political orientation”, “dogmatic intolerance” and “authoritarianism”, according to the pattern of individual item loadings. Item loadings for each question (questionnaires indicated by different colours) are presented. B-D To investigate the relation between these constructs scores on the three factors were extracted for each individual. B I observed a quadratic relationship between political orientation and dogmatic intolerance, revealing that people on the extremes of the political spectrum are more rigid and dogmatic in their world views. C A linear relationship between political orientation and authoritarianism was observed, with people from the far right of the political spectrum showing more obedience to authorities and conventions. D Dogmatic intolerance and authoritarianism were positively correlated, indicating commonality between these two sub-components of radicalism.
4.3.2 Radical beliefs are related to reduced metacognitive ability

I next investigated whether metacognitive aspects of decision-making predict facets of radicalism. Subjects were asked to carry out a series of perceptual discrimination tasks assaying decision-making and metacognition (Figure 4.2) before filling out the same questionnaires administered in Study 1. A first experiment was conducted on a new sample of 381 US participants (Study 2) and all key findings were replicated in an independent sample of 417 US participants (Study 3). Importantly I also replicated both the 3-factor structure of questionnaire responses observed in Study 1 and the pattern of interrelations between factors (quadratic relationship between dogmatic intolerance and political orientation, Study 2: \( \beta =.42, p<10^{-16}; \) Study 3: \( \beta =.4, p<10^{-15} \); linear relationship between authoritarianism and political orientation, Study 2: \( \beta =.32, p<10^{-8}; \) Study 3: \( \beta =.38, p<10^{-12} \); positive association between authoritarianism and dogmatic intolerance, Study 2: \( \beta =.22, p<10^{-4}; \) Study 3: \( \beta =.29, p<10^{-7} \)). In the confidence task (Task 1), participants first completed a series of perceptual discrimination judgments as to which of two flickering patches contained a greater density of dots, followed by confidence ratings in their choices. Participants were rewarded according to the extent to which confidence ratings tracked their objective performance over 60 trials, and were thus incentivized to report their confidence as accurately as possible.

In line with my hypothesis, higher values of dogmatic intolerance were associated with reduced metacognitive sensitivity (Study 2: \( \beta =-.12, p=.032, R^2 =.01 \), see Figure 4.3A), in the absence of any effect on perceptual discrimination performance (Study 2: \( \beta =.02, p=.77 \)) and controlling for key demographic variables (i.e. age, gender, education). Importantly, there was also no relation between dogmatism and overconfidence (Study 2: \( \beta =.07, p=.26 \)), suggesting a specific reduction in the sensitivity with which confidence tracks performance, rather than a bias in confidence. I replicated this reduction of metacognitive sensitivity in dogmatic individuals in Study 3 (\( \beta =-.13, \) one-tailed \( p=.008, R^2 =.014 \)), again in the absence of any observed link with perceptual performance (\( \beta =.04, p=.60 \)) or confidence bias (\( \beta =0.07, p=.24 \)). These results show that more dogmatic people manifest a lowered capacity to discriminate between their correct and incorrect decisions, after controlling for differences in both primary task performance and confidence bias. I obtained a qualitatively similar pattern for authoritarianism (see Figure 4.3B), with trends of reduced metacognitive sensitivity (Study 2: \( \beta =-.11, p=.051; \) Study 3: \( \beta =-.08, \) one-tailed \( p=.08 \)), but no relation with perceptual performance or confidence bias (all \( p \)-values >.17). Across both factor scores, this failure in metacognition was driven by radicals holding unreasonably high confidence in incorrect decisions compared to moderates (Figure 4.4B).

In light of long-standing debates about whether the cognitive profile of radicals is
Figure 4.2: Behavioral tasks measuring metacognitive sensitivity and post-decision evidence integration. A Confidence task (Task 1): Participants were asked to judge which of two patches contained a greater number of flickering dots before rating their confidence in each decision. Task difficulty was determined by a fixed difference in dot number between the patches and was individually adjusted in an initial calibration phase to target approximately 71% correct performance. B Post-decision evidence integration task (Task 2): Participants performed the same perceptual decision as in part (A), but after each decision they were presented again with a new sample of flickering dots, before rating their confidence. In half of trials participants received the same evidence strength post-decision as pre-decision, while in the other half of trials they received stronger post-decision evidence (pre-adjusted to a strength that led to 80% performance).
more similar to those on the left or right sides of the political spectrum (Greenberg & Jonas, 2003; Jost et al., 2003), I also tested the relation between political orientation (rightward vs. leftward) and metacognition. Here, the pattern of results was qualitatively different, with no reduction of metacognitive sensitivity (Study 2: $\beta = -.08, p = .18$; Study 3: Study 3: $\beta = -.02, p = .73$) in more conservative participants. In contrast, more conservative participants showed an increased bias towards over-confidence (Study 2: $\beta = .15, p = .035, R^2 = .01$; Study 3: $\beta = .12$, one-tailed $p = .033, R^2 = .008$; see Figure 4.3C), as found previously (Moore & Healy, 2008).

4.3.3 Metacognitive ability predicts post-decision evidence integration

Metacognitive sensitivity is thought to be strongly linked to an integration of evidence following a decision, allowing latitude for the recognition and reversal of incorrect choices (Van Den Berg et al., 2016; Fleming et al., 2018; Yeung & Summerfield, 2012). Having demonstrated a specific decrease in metacognitive sensitivity in more radical participants, I next considered the same participants’ sensitivity to new evidence. To specifically probe such post-decisional processing, in a second phase of the experiment I inserted an additional sample of evidence (a new series of flickering dots) after subjects had committed to a choice, but prior to providing a confidence rating (Task 2). Following correct choices, additional evidence should normatively increase participants’ confidence (due to integration of confirmatory evidence; green markers in Figure 4.4B) whereas for incorrect choices, additional evidence should lead to a decrease in confidence (due to integration of disconfirmatory evidence; red markers in Figure 4.4B).

In line with the proposal of a post-decisional process supporting metacognition (Van Den Berg et al., 2016; Murphy et al., 2015), metacognitive sensitivity measured in Task 1 correlated with participants’ sensitivity to post-decision evidence in Task 2 (Study 2: $\beta = .1, p = .034$, Study 3: $\beta = .18$, one-tailed $p < .0001$).

4.3.4 Radical beliefs are characterized by reduced integration of disconfirming post-decision evidence

Furthermore, and consistent with a tripartite relationship between radicalism, metacognitive sensitivity and post-decision evidence integration, dogmatic intolerance was associated with a specific reduction in disconfirmatory evidence integration (Study 2: $\beta = -.15, p = .016, R^2 = .015$; Study 3: $\beta = -.1$, one-tailed $p = .034, R^2 = .008$; see Figure 4.3A), representing a smaller decrease in confidence on incorrect trials. Conversely there was no association between confirmatory evidence integration on correct trials and dogmatism (Study 2: $\beta = .06, p = .37$; Study 3: $\beta = -.09, p = .13$), i.e. more dogmatic people showed similar increases of confidence on correct trials as that
Figure 4.3: Impaired metacognitive sensitivity and reduced disconfirmatory evidence integration predict facets of radicalism. A-C Multiple regression analyses predicting factor scores (dogmatic intolerance, authoritarianism and political orientation) from metacognitive sensitivity and post-decision evidence integration, controlling for multiple demographic variables (gender, education, age) and other task-related variables (e.g. performance in the perceptual decision task). Perceptual performance was averaged across Tasks 1 and 2. I present standardized beta coefficients ± standard error of predictors for Study 2 (left markers, N=381) and Study 3 (right markers, N=417). A Dogmatic intolerance was associated with impaired metacognitive sensitivity and reduced disconfirmatory evidence integration, in the absence of differences in overconfidence or performance. B Authoritarianism showed qualitatively similar patterns of association as dogmatism. C Political orientation (higher values represent more conservative views) was consistently associated with a bias towards overconfidence but not changes in metacognitive sensitivity or post-decision evidence integration. Effects in Study 3 were tested one-tailed based on the directional hypothesis derived from Study 2. †p < .1; *p < .05; **p < .01, ***p < .001. Task 1 = Confidence task; Task 2 = Post-decision evidence integration task.
seen in moderates. I again found a similar pattern of results in relation to authoritarianism (see Figure 4.3B) with decreased disconfirmatory evidence integration in Study 2 ($\beta=-.19$, $p=.005$, $R^2=.019$) and the same trend in Study 3 ($\beta=-.09$, one-tailed $p=.05$, $R^2=.01$), despite no effect on confirmatory evidence integration (Study 2: $\beta=-.04$, $p=.53$; Study 3: $\beta=-.09$, $p=.16$). In contrast, while higher conservatism was related to reduced disconfirmatory evidence integration in Study 2 ($\beta=-.17$, $p=.012$), this effect was not replicated in Study 3 ($\beta=-.02$, one-tailed $p=.33$).

4.3.5 Computational mechanism underlying radical beliefs

In light of associations between dogmatic intolerance, authoritarianism and multiple behavioural measures of metacognitive sensitivity I next asked whether I could identify a core computational driver of radicalism. I first combined the factor scores of dogmatic intolerance and authoritarianism to construct a composite measure of radicalism. As expected, this combined measure showed similar relationships with metacognition as the individual components, with impaired metacognitive sensitivity (Study 2: $\beta=-.13$, $p=.01$, $R^2=.018$; Study 3: $\beta=-.13$, one-tailed $p=.006$, $R^2=.015$) and reduced disconfirmatory evidence integration (Study 2: $\beta=-.21$, $p=.001$, $R^2=.027$; Study 3: $\beta=-.12$, one-tailed $p=.015$, $R^2=.011$). I next used this score to identify putative mechanisms underpinning reduced metacognitive sensitivity and disconfirmatory evidence integration in more radical participants.

To this end, I compared alternative computational models of how post-decision evidence affects confidence (Fleming et al. 2018; Bronfman et al. 2015; Talluri et al. 2018). All models were grounded in signal detection theory, with two free parameters ($\mu_{\text{low}}$ and $\mu_{\text{high}}$) representing internal evidence strength for the weak and strong evidence conditions respectively. The models differed in how they updated their confidence in light of new evidence. A “temporal weighting” model allows an asymmetry in the overall weighting of pre- and post-decision evidence; a “choice bias” model adds evidence for the chosen response, without altering post-decision evidence integration; and a “choice weighting” model incorporates asymmetric weighting of confirmatory and disconfirmatory evidence. I fit the model simultaneously to data from the confidence task (Task 1, no post-decision evidence) and the post-decision evidence task (Task 2), and compared models based on how well variability in fitted parameters captured individual differences in radicalism in a linear regression.

The “choice bias” model best explained variations in radicalism (difference in BIC relative to next best model: Study 2 = 3.5 and Study 3 = 3.3; see Figure 4.4A) via a positive association with choice-dependent biases in confidence (Study 2: $\beta=.14$, $p=.012$; Study 3: $\beta=.18$, one-tailed $p=.0005$). This model accounts for a reduction in post-decisional processing in more radical participants by boosting confidence in chosen options, thereby making changes of mind less likely (Figure 4.4B).
Chapter 4. Metacognitive failure as a feature of those holding radical beliefs

Figure 4.4: Individual differences in radicalism are captured by a choice bias model.
A A choice bias model fitted to the confidence data across both tasks best accounted for variations in a composite measure of radicalism (summed factor scores of dogmatic intolerance and authoritarianism). I compared between three computational models within multiple regressions that predicted radicalism from fitted model parameters. I present the BIC of each regression against the lowest BIC in the model set (the best model has a difference in BIC of zero). B Radicals reduce their confidence less when new evidence indicates they are wrong (reduced disconfirmatory evidence integration). To visualize this effect, I combined data from Study 2 and Study 3 and compared the 10% most radical participants (based on the composite measure) against the rest of the sample. Aggregate confidence ratings are separated according to whether the decision was correct (green) or incorrect (red). Markers (circles and squares) show raw data (group averages ± 95% confidence interval) for each condition. Lines (solid line=moderates; dashed line=radicals) show posterior predictives from the choice bias model. Predictions were simulated from best-fitting parameters and represent group averages ± 95% confidence interval. Task 1 = Confidence task; Task 2 = Post-decision evidence integration task.
4.4 Discussion

Taken together, my data show that key facets of radicalism are associated with specific alterations in metacognitive abilities. The finding that decision performance per se was not associated with radicalism reveals that a specific change in information processing is manifest at a metacognitive, rather than cognitive, level. Importantly, my results show radicalism is associated with reductions in metacognitive sensitivity, i.e. the reliability with which subjects distinguish between their correct and incorrect beliefs. Thus my findings complement and extend previous studies documenting alterations in confidence in political radicals (Ortoleva & Snowberg, 2015; Toner et al., 2013; Brandt et al., 2015), but suggest that these alterations may stem from changes in metacognitive sensitivity. In contrast, for more right-wing subjects (as indexed by political orientation), a change in confidence bias was observed. Without the application of psychophysical measures of metacognition it has not, up until now, been possible to disentangle these two factors.

What is striking is my demonstration that these impairments are evident during performance of a low-level perceptual discrimination task, where participants are unlikely to have strong a priori vested interest in the outcome of their decisions, ruling out multiple possible confounds (e.g. prior knowledge and motivational factors). This contrasts with previous studies that have investigated changes of mind about political attitudes themselves, a context where there exists a strong motivation for people to maintain their current beliefs in order to sustain a positive (and consistent) self-image (Nyhan & Reifler, 2010; Kaplan et al., 2016; Van Bavel & Pereira, 2018). Thus, my results suggest a potential explanation for why it is notoriously difficult to change extreme beliefs by what would appear to be the simple expedience of confronting people with evidence that contradicts these beliefs. Before such information can update attitudes, the manner in which a recipient processes this information may need to be altered. I stress however that the results are entirely compatible with a complementary role of motivational factors as contributing to the maintenance of radical beliefs, and it is possible that motivational factors may themselves interact with metacognitive abilities.

The computational modelling results suggest that a reduction in changes of mind in radicals is driven by a boosting effect of choice, leading participants to assign undue probability to the option they chose without affecting the integration of post-decision evidence. At first, this modelling result might appear surprising as the behavioural trends seem to indicate a specific reduction in the integration of negative evidence with no boost in positive evidence. However, since the choice bias model adds evidence to each decision in a logarithmic space (the choice bias was added to the log-odds) this evidence is non-linearly transformed to derive confidence ratings. If a participant has already strong evidence in favour for a decision, e.g. when having received confirming post-decision evidence, such a bias will only have
small effects on the final confidence rating. In comparison, at intermediate levels of evidence, e.g. when a participant received disconfirming post-decision evidence after an initial decision, such a choice bias will strongly influence the resulting confidence. Therefore, the choice bias model implicitly leads to an asymmetric influence of confirming or disconfirming post-decision evidence.

This computational mechanism shares notable similarities with classical findings in psychology in which the act of making a choice itself affects subsequent preferences (Brehm, 1956; Sharot et al., 2009). In contrast, recent laboratory studies of post-decision evidence integration have found that subjects’ behaviour was best described either by a near-optimal Bayesian model (Fleming et al., 2018), by diminished sensitivity to post-decision evidence (Bronfman et al., 2015) or a selective gain for the incorporation of choice consistent information (Talluri et al., 2018), as also shown in chapter 3. However, the modelling approach aimed to find a model that best accounts for individual differences in radicalism (while also fitting the overall behavioural pattern). Thus, it might well be possible that “non-radical” participants are best described by models suggested in previous literature, while a choice-bias parameter was the computational alteration that best described the changes in radical participants (compared to non-radical participants). Moreover, since initial confidence and post-decision evidence integration were here measured in two separate tasks, it was not possible to fit the models from chapter 3 (i.e. a model in which initial confidence modulates post-decision evidence processing), as no trial-by-trial ratings of initial confidence was given in the post-decision evidence task. Therefore, these studies describe slightly different experimental set-ups.

In the current study I investigated independent judgments wherein participants integrate two consecutive samples of information. This is distinct from more elaborate beliefs formed over longer time scales which require integration of multiple samples of information. A useful future extension of this work will be to extrapolate these findings to situations where learning is required over extended periods of time (Zmigrod et al., 2018; Meyniel et al., 2015a). The computational model fits indicate that more radical participants assign undue probability to chosen options when updating their confidence, which over repeated exposure to multiple samples of evidence may summate, such that even small asymmetries in information processing could lead to a highly skewed representation of reality. In the current task, such resistance to updating is detrimental, leading to a loss of earnings. However, in other scenarios, such as if there were reason to distrust the fidelity of the new information, a reduction in belief flexibility may prove adaptive. Such considerations remain to be explored in future studies and point to the intriguing notion that metacognitive flexibility may itself be amenable to strategic or environmental influences.

Here I used perceptual decision-making as a model system which permitted precise control over performance so as to reveal relationships between radicalism and
metacognition. A question remains as to whether the metacognitive alterations shown here would extend to other types of decision (e.g. value-based, memory-based). Recent evidence points towards a core domain-general circuit supporting metacognitive abilities (Morales et al. 2018; Faivre et al. 2018), suggesting that metacognition as measured in the current task may represent an indicator of a more general metacognitive ability. Despite relatively small effect sizes, the findings linking radicalism to changes in metacognition are robust and replicable across two independent samples. However, I note that other, domain-specific facets of metacognition (e.g. insight into the validity of higher-level reasoning or certainty about value-based choices (De Martino et al. 2013)) are arguably closer to the drivers of radicalization of political and religious beliefs, suggesting the current results represent a lower bound for the strength of a relationship between metacognitive abilities and radicalism. Similarly, while these measures of radicalism were derived from questionnaires tapping into political attitudes, it is possible that impairments in metacognition may constitute a general feature of radicalism about political, religious and scientific issues.
Chapter 5

Dogmatism manifests in lowered information search under uncertainty

In the previous chapter, I showed that dogmatic participants are characterized by a general difficulty in recognizing and revising their mistakes when presented with new information. However, in everyday life we are rarely passively presented with new information but have to actively seek it out. Therefore, I aimed to extend the previous findings to this scenario and test whether more dogmatic participants would also show alterations regarding the active search for information.

5.1 Introduction

A never-ending flow of informational choices is a defining feature of modernity (Hills, 2019). On the one hand, we can decide whether we want to find out a stranger’s opinion of a restaurant. On the other, we are in charge of gathering information critical to our health (Garrett, 2020), the survival of democracies (Lazer et al., 2018) or the conservation of the planet (Lemos et al., 2012). These decisions are in turn a crucial determinant of our beliefs.

Unsurprisingly, cognitive science has studied information search extensively, providing us with a rich empirical and theoretical perspective on these choices (Sharot & Sunstein, 2020; Schulz & Gershman, 2019; Gottlieb & Oudeyer, 2018; Hills et al., 2015). This research indicates that people prefer to seek information that confirms their beliefs and has positive valence, exemplified in reading an additional news story about the victory of a favored political party. This type of motivated search is evident both in laboratory experiments (Charpentier et al., 2018; Gesiarz et al., 2019; Jonas et al., 2001) and in real-world data (Bakshy et al., 2015; Schmidt et al., 2017; Hart et al., 2009).

In contrast, normative frameworks propose that uncertainty, rather than valence,
should determine where and when we should seek information (Schulz & Gershman, 2019, Gottlieb & Oudeyer, 2018, Schulz et al., 2019). In the absence of external feedback, humans can only rely on internal confidence signals to guide information search. This confidence signal should be based on the initial evidence strength, thus representing a Bayesian probability of making a correct judgement (Pouget et al., 2016). In simple scenarios, such probability of being correct can be combined with the cost of further information and the expected information value (the value of the increased likelihood to make a correct decision after reception of more evidence) to calculate action-values for and against further information search (Dayan & Daw, 2008). Optimally, such computations should lead participants to seek out more information when they have low confidence in a decision as in this case the likelihood of receiving corrective (and thus valuable) information is highest. Empirical data bears out such predictions (Schulz et al., 2019, Desender et al., 2018), showing that people are more likely to seek information when they express low confidence (i.e., higher uncertainty) in their decisions (Desender et al., 2018, Boldt et al., 2019).

Both motivational influences and failures in uncertainty-guided information search can lead to biased or inaccurate belief formation, albeit via distinct mechanisms. For example, a person who does not believe in climate change is likely to show a preference for media that refutes its occurrence (Newman et al., 2018), reinforcing pre-existing beliefs. Alternatively, people with doubts about the science of global warming (Fischer et al., 2019) might fail to act on this uncertainty, and as a consequence not seek out further, potentially corrective evidence.

An unwillingness to seek out such corrective information is one potential source of dogmatism, a worldview that involves a rigid maintenance of one’s beliefs (Altemeyer, 2002, Toner et al., 2013, Rohach, 1960) regardless of their accuracy (Rollwage et al., 2018). The scope of this worldview is wide-ranging and transcends specific issues and positions, affecting political (van Prooijen & Krouwel, 2017), scientific (Fast & Horvitz, 2016), and religious debate (Long & Ziller, 1965, Altemeyer, 2002). Prior questionnaire-based research suggests a link between such a dogmatic style of thinking and a willingness to seek further information (Long & Ziller, 1965, Altemeyer, 2002). However, how motivation and uncertainty precisely contribute to this phenomenon, remains unknown.

Here, I address this question using a precise assay of uncertainty-guided information search in the context of a low-level perceptual decision-making task. Leveraging the computational precision afforded by this approach, I test (in both exploratory and replication samples) whether individual differences in sensitivity to uncertainty explain a disposition to hold dogmatic beliefs. This approach builds on previous research on the influence of confidence on information search (Desender et al., 2018, 2019b) and allows ruling out possible motivational influences: participants are unlikely to approach such a low-level task with vested interests or prior knowledge,
and should not hold differing appraisals of the helpfulness of further information. Moreover, eliciting trial-by-trial estimates of confidence enabled precise inference on how participants use uncertainty to guide their search.

I studied a sample of 370 US adults (study 1) and replicated all key findings in an independent second sample of 364 participants (study 2). Both samples were recruited through Amazon Mechanical Turk and comprised a wide range of ages and educational backgrounds. Participants first completed an information-seeking task and then answered a number of questionnaires that allowed measurement of general belief rigidity and dogmatism, political beliefs, authoritarianism, and intolerance to opposing political attitudes. This methodology builds on the study presented in the previous chapter.

5.2 Methods

5.2.1 Participants

Both studies were conducted online and recruited US adults through the online labor marketplace Amazon Mechanical Turk. They were approved by the Research Ethics Committee of University College London (1260-003) and subjects gave informed consent. In Study 1, 370 subjects’ data was analyzed. Subjects were paid a basic payment of $1 and earned a bonus of up to $6 based on their adequate completion of the questionnaires and their performance on the information-seeking task (see below). Participants were 50% percent female (50% male) and the mean age was 36.62 years (SD = 11.61, range: 19 to 81 years). In Study 2 (replication), I analyzed data from 364 participants with the same payment scheme as in Study 1. An a priori power analysis based on the information-seeking effect size from Study 1 determined the sample size in study 2, giving us a power above 80% to detect the association between dogmatism and average information search. The sample consisted of 52% women (48% male; mean age = 36.55, SD = 11.09, 18 to 74 years).

For study 1, I collected data from 500 participants. Out of this sample, I excluded 130 subjects to obtain a final analyzable sample of 370 participants. Exclusion criteria were defined a priori and were based on similar reasoning as in the previous chapter. For study 1, 6 subjects were excluded for failing to correctly respond to at least one of two catch questions that were placed randomly within the questionnaires. Furthermore, I excluded 78 subjects due to their perceptual performance in the initial decision falling outside of an interval between 60% and 85% correct indicating non-convergence of the staircase procedure (see below). An additional 34 participants were excluded because they chose the same confidence rating on more than 95% of trials. I also excluded 12 participants whose data contained more than 5% of total trials with missing data. For study 2, I again collected data from 500 participants. Out of this sample, I excluded 136 subjects, leaving a final sample
Chapter 5. Dogmatism manifests in lowered information search

of 364 subjects using the same exclusion criteria as for study 1. Specifically, I excluded 9 subjects for failing to correctly respond to the catch questions. A further 85 subjects were excluded due to their perceptual performance falling outside of the predefined range. I additionally excluded 36 subjects because they had chosen the same confidence rating on more than 95% of trials. Finally, I excluded six participants whose data contained more than 5% of trials with missing data.

5.2.2 Factor Analysis

The same questionnaires and factor analytical methods as in the previous chapter were used. To maximize the precision of the factor loading estimates and the factor scores, I pooled the present sample with the data from the previous chapter. This resulted in a total sample of 2,135 participants for the factor analysis. I observed qualitatively similar pattern of factor loadings for both the pooled sample of 2,135 participants and the two individual samples from this chapter.

5.2.3 Experimental design

5.2.3.1 Stimuli

The stimuli were exactly the same stimuli as presented in the previous chapter.

5.2.3.2 Task and procedure

Both studies followed the same protocol and participants spent around 45 minutes on the experiment, which was divided into three parts. Participants first received information and reported their demographic information. Following this, they then first completed a 120-trial calibration phase to individually determine task difficulty, identical to procedures in the previous chapter. There, participants simply had to indicate which of the two boxes contained more flickering dots by pressing the “2” or “6” key (indicating left and right) and received feedback about their correctness through a colored frame around their chosen option. This was followed by the information-seeking task (see Figure 5.1), in which subjects received no feedback about their correctness. The information-seeking task consisted of a total of 100 trials. Participants then went on to fill out the mentioned questionnaires.

5.2.3.3 Information-seeking task

Across the 100 trials of the information-seeking task, participants were presented with the stimulus strength determined in the calibration phase (study 1: mean performance = 73.80%, SD = 6.57%; study 2: mean performance = 73.67%, SD = 6.50%). As in the calibration phase, participants had to decide whether more dots were in the left or in the right box (the initial decision). Simultaneously, they
Figure 5.1: Information seeking task. Participants first had to judge whether a left or right square contained more flickering dots. They then chose whether they wanted to see a stronger, more helpful, version of this stimulus again, costing them either 5 or 20 points. After seeing either this additional stimulus or blank boxes, they again made a judgment as to which box contained more dots. I compensated participants for the accuracy of this final decision alone (100 points for a correct judgement, 0 points for an incorrect judgement). Participants rated their confidence (on a six-point scale from “sure left” to “sure right”) at both the initial and the final decision. The difficulty of the initial decision was fixed through an individually pre-determined difference in dot number that resulted in approximately 71% accuracy. The post decision-evidence strength was yoked to this predetermined dot difference, to make the final decision easier. Participants’ bonus payment was linked to their performance in the task: they received a $2 bonus for completing the task and an extra 4 cents for every 100 points they had earned on the task (average bonus payment, study 1: mean = 3.11 $, SD = .34 $; study 2: mean = 3.11 $, SD = .35 $).
Chapter 5. Dogmatism manifests in lowered information search

5.2.4 Statistical Analysis

5.2.4.1 Task behavior

I conducted several analyses to ensure participants understood the task and were able to perform it adequately. Within-participant effects were investigated using trial-by-trial hierarchical mixed effects models, using the “afex” package. Specifically, I constructed logistic models with binary outcomes as respective dependent variables and the corresponding predictors as fixed effects. I included per-participant random intercepts and slopes and employed likelihood-ratio tests to obtain p-values ([Singmann & Kellen 2017]). To quantify relationships between subjects’ average information seeking and their final decision accuracy I set up a general linear model (GLM) using the {	exttt{lm}}() function in R. All analyses were performed separately for the two studies.

5.2.4.2 Factor scores and behavioral variables

I conducted the following regression analysis using the {	exttt{lm}}() function in R. All analyses were performed separately for the two studies and effects were tested two-tailed if not stated otherwise.

1. To investigate the relationship between the factor scores I constructed polynomial regression models. Specifically, I built these models for each possible factor combination and compared (1) a linear fit, (2) a quadratic fit and (3) a combined linear and quadratic fit based on their BIC and used the best fitting model for inference on the parameters.

2. To investigate the relationship between information seeking and the factors observed through the questionnaire, I set up one GLM per factor, explaining the respective variance in this factor score through participants’ average information seeking. I controlled for the following covariates: age, gender, education, subjects’ average performance and confidence level on the initial decision, objective stimulus strength (indicated by the logarithm of the dot difference) and performance on the stronger version of the stimulus (as recorded during the calibration phase). For significant variables of interest, I calculated $R^2$ values by comparing the variance explained by a full model including information seeking relative to a model excluding this predictor.

3. Finally, to check whether dogmatism was associated with a reduction in points earned in the task, I set up the same model used for the information-seeking analysis, but replaced the information-seeking predictor with the points earned on the task.
4. To investigate whether dogmatism was linked to a reduction in final decision accuracy and whether this arose from a lowered propensity to seek out information, I conducted a mediation analysis. This analysis was conducted using the “mediate” package in R ([Tingley et al. 2014]) which uses a quasi-Bayesian Monte-Carlo method based on normal approximation to estimate the significance of the mediation effect ([Imai et al. 2010]). I again entered the covariates used for the original information-seeking analysis as control variables into all paths of the mediation analysis.

5.2.4.3 Computational modeling

To probe the underlying mechanisms contributing to dogmatic individuals’ information search, I set up a computational model that investigated the factors impacting an individual’s decision to seek out more information. Specifically, I modeled the information-seeking choices as a function of the confidence level and the current information cost:

\[ P(\text{InformationSeeking}) = \frac{1}{1 + \exp\left(- (\beta_0 + \beta_1 \cdot \text{Confidence} + \beta_2 \cdot \text{Cost})\right)} \] (5.1)

The three \( \beta \)'s capture three independent behavioral phenomena (see Figure 5.5A): Differences in the model’s intercept, \( \beta_0 \), represent a general shift in willingness to seek out information; \( \beta_1 \) represents how strongly participants’ information seeking choices are influenced by confidence; and \( \beta_2 \) indicates the influence of information cost on subjects’ willingness to seek out more information.

I was primarily interested in whether any of these parameters were associated with individual differences in dogmatism, i.e., whether dogmatism was linked to a general tendency to seek out less information (\( \beta_0 \)), a differential influence of confidence on information seeking (\( \beta_1 \)) or an altered sensitivity to information costs (\( \beta_2 \)).

Because classical maximum-likelihood based methods can frequently provide noisy estimates with so few data points, I employed a hierarchical fitting procedure ([Carpenter et al. 2017]). In such a hierarchical model, individual parameters, \( \beta_i \), are drawn from a group level prior distribution. For example, for the first parameters, \( \beta_{0,i} \), I can write:

\[ \beta_{0,i} \sim \mathcal{N}(\mu_{B_0}, \sigma_{B_0}) \] (5.2)

Here, \( \mu_{B_0} \) represents the population mean that then informs the estimation of \( \beta_{0,i} \), the individual parameters of \( \beta_0 \) for participant i, from a population distribution. Conventionally, parameters obtained through such an approach can then be correlated with an external measure of differences between individuals. However, this procedure is suboptimal because it assumes no variability in the mean of the
population in the initial model fit, possibly distorting or minimizing potential relationships between the parameter and external factors [Moutoussis et al. 2018]. To maintain the advantages of hierarchical fitting while avoiding such pitfalls regarding individual differences, here I employ a procedure recently prescribed by Moutoussis et al. (2018). There (see also Figure 5.5C), the relationship between the parameters and individual differences is embedded into the estimation of the parameters themselves through the prior, so that:

$$\beta_{0,i} \sim \mathcal{N}(\mu_{B0} + \rho_0 \ast \text{Dogmatism}_i, \sigma_{B0})$$  (5.3)

To capture interindividual differences in the parameter, I allow the mean of the population distribution to vary as a function of dogmatism through the embedded parameter $\rho_0$. To enable accurate hierarchical estimation, I pooled the samples from both studies and only included subjects that sought out information on at least 5% and at most on 95% of trials. In doing so, I achieved a total sample of 568 subjects. I built the computational model using the programming language Stan (Carpenter et al., 2017) which uses a form of Markov chain Monte Carlo sampling, Hamilton Monte Carlo sampling, to estimate posteriors over parameters.

5.3 Results

5.3.1 Measuring Dogmatism

I derived a measure of dogmatism from a factor analysis applied to the same questionnaire battery as in the previous chapter (see Figure 5.2A). The breadth of the battery allows to quantitatively distinguish dogmatism from other, possibly related, constructs and study their interplay. The most parsimonious factor structure contained three factors, capturing 40% of questionnaire variance. A first factor represented individuals’ position on a left-right political spectrum, and a second factor described their domain-general dogmatism (Altemeyer, 2002). A third factor captured variance in beliefs as to the superiority of participants’ policy preferences (a factor related to but also theoretically independent of dogmatism) (Toner et al., 2013).

I note that the factor structure obtained using data from the current study (N = 734), as well as when pooling data with the data set from the previous chapter (Rollwage et al. 2018, N = 2,135), is qualitatively different from the structure obtained in the previous chapter, despite using the same set of questionnaires and factor analytic approach. Two key changes surfaced in these newer analyses: (1) Political orientation and authoritarianism are unified in one factor, and (2) dogmatism and political belief superiority are separated into two factors. Potential reasons for these differences are discussed below.
First, the political orientation factor includes items from the authoritarianism scales, unlike in the previous chapter where the left- and right-wing authoritarianism questionnaires loaded most heavily on a separate factor that was labelled “authoritarianism”. However, it is noteworthy that the political orientation and authoritarianism factors were correlated, hinting at an underlying relationship, even in the data of the previous chapter. One possible explanation behind this change in factor structure is a change in the structure of US political attitudes over time (the data set from the previous chapter was acquired in early 2017, whereas this data set was acquired in early 2019). Policy positions and partisan identifications might have unified increasingly into one continuum that now encompasses left- and right-wing authoritarianism. This is congruent with a trend in U.S. politics that has seen an increasing division into two clearly defined camps, each endorsing more extreme beliefs over time (Center, 2017; Kohut et al., 2012).

Second, a previous factor labeled “dogmatic intolerance” encompassed items from two factors that were separately identified in the present study, a domain-general “dogmatism” factor and a policy-focused “political belief superiority” factor. While these two constructs might initially seem highly related, a separation along these lines is consistent with theoretical perspectives on dogmatism. Specifically, whereas dogmatism describes a general certainty in one’s attitudes (Altemeyer, 2002; Duckitt, 2009), political belief superiority involves a judgement of other people’s specific policy opinions (Toner et al., 2013; Petty et al., 2007). While this last factor is specific to political policy, dogmatism itself is a broader construct that describes the general way beliefs are held and acted upon (Rokeach, 1960; Snow & Cross, 2011; Wintrobe, 2006). Dogmatism thereby transcends specific political preferences evident, for example, in a reported link between dogmatism and religious fundamentalism (Altemeyer, 2002). With respect to the current hypothesis, a domain general measure of dogmatism appears to be more likely be linked to computational alterations in information processing.

5.3.2 Interrelation between factor scores

I explored inter-relations between individuals’ political orientation, dogmatism, and political belief superiority. I found both a positive linear (study 1: $\beta_{\text{linear}} = .16, p = .001$; study 2: $\beta_{\text{linear}} = .24, p < 10^{-6}$) and quadratic (study 1: $\beta_{\text{quadratic}} = .35, p < 10^{-13}$; study 2: $\beta_{\text{quadratic}} = .37, p < 10^{-13}$) relationship between political orientation and dogmatism across both samples. These findings indicate that individuals on both the far left and far right of the political spectrum show enhanced dogmatism, but interestingly this increase in dogmatism is more marked for those on the far right (see Figure 5.2B). Conversely, a negative linear (study 1: $\beta_{\text{linear}} = -.33, p < 10^{-10}$; study 2: $\beta_{\text{linear}} = -.32, p < 10^{-9}$) and positive quadratic relationship (study 1: $\beta_{\text{quadratic}} = .43, p < 10^{-20}$; study 2: $\beta_{\text{quadratic}} = .34, p < 10^{-10}$) between political
orientation and political belief superiority reveals that individuals on both the far left and far right show heightened beliefs in the superiority of their respective policy positions, but more so on the far left (see Figure 5.2C). Finally, I found a positive linear relationship between political belief superiority and dogmatism, indicating more dogmatic subjects also tended to be more confident in the superiority of their specific political convictions (study 1: $\beta_{\text{linear}} = .26$, $p < 10^{-6}$; study 2: $\beta_{\text{linear}} = .14$, $p = .006$, see Figure 5.2D).

5.3.3 Measuring Information Search

I next tested my primary hypothesis of a link between dogmatism and uncertainty-guided information search. To probe this, I presented participants with a perceptual information-seeking task (see Figure 5.1) where they received monetary reward for correctly judging which of two flickering boxes contained the greater number of dots.

First, I validated that participants used the additional information adaptively. Participants chose to see additional information more often after initial mistakes (study 1: $\beta = -.76$, $p < 10^{-69}$; study 2: $\beta = -.77$, $p < 10^{-52}$, Figure 5.3A) and were more likely to make an accurate final decision after having decided to see additional information (study 1: $\beta = 1.23$, $p < 10^{-80}$; study 2: $\beta = 1.12$, $p < 10^{-74}$, Figure 5.3B). Similarly, examining individual differences in information search revealed that participants who sought additional information more often also performed better in their final decisions (study 1: $\beta = .68$, $p < 10^{-50}$; study 2: $\beta = .71$, $p < 10^{-56}$, Figure 5.3C), and received a higher pay-off (study 1: $\beta = .43$, $p < 10^{-17}$, study 2: $\beta = .49$, $p < 10^{-22}$). Importantly participants were also sensitive to cost, seeking information less often when it was more expensive (study 1: $\beta = -1.22$, $p < 10^{-58}$; study 2: $\beta = -1.34$, $p < 10^{-66}$, Figure 5.3D). Finally, participants sought out additional evidence less often when they were more confident in their initial decision (study 1: $\beta = -2.02$, $p < 10^{-126}$; study 2: $\beta = -1.93$, $p < 10^{-130}$, Figure 5.3E), demonstrating that internal signals of uncertainty were used to guide information search.

5.3.4 Information Search and Dogmatism

I next asked whether more dogmatic participants differed in their propensity to seek out information. To that end, I sought to explain variance in dogmatism factor scores using behavioral measures derived from the information-seeking task. In line with my hypothesis, higher levels of dogmatism were associated with a reduced willingness to seek out information (study 1: $\beta = -.15$, $p = .005$, $R^2 = .02$, Figure 5.4A). No significant relationships with initial decision accuracy (study 1: $\beta = .02$, $p = .72$) or overall confidence level (study 1: $\beta = -.03$, $p = .61$) were found and there was no significant association with metacognitive ability (study 1: $\beta = -.04$, $p = .52$), which is in contrast to the findings from the previous chapter. This discrepancy
Chapter 5. Dogmatism manifests in lowered information search

Figure 5.2: Liberal and conservative extremes of the political spectrum predict levels of dogmatism and belief superiority. Data presented for both studies combined (N=734). A A factor analysis revealed a three-factor structure underlying responses to multiple questionnaires assessing political convictions, authoritarianism, and belief rigidity. The three factors identified (1) “political orientation” (liberal to conservative), (2) “dogmatism” representing a domain-general belief certainty and (3) “political belief superiority” characterizing participants’ confidence in specific political convictions. Item loadings for each question are presented with individual questionnaires indicated by colors. B-D I examined the relationships between these constructs by investigating individual scores for each factor (combined data for study 1 and 2 are plotted). B-C I observed a combined linear-quadratic model provided the best fit to the relationship between both political orientation and dogmatism as well as between political orientation and political belief superiority. C A linear relationship between political orientation and authoritarianism was observed, with people from the far right of the political spectrum showing more obedience to authorities and conventions. D The relationship between dogmatism and political belief superiority was best characterized by a linear relationship.
Figure 5.3: Participants adaptively seek additional information. 

A Participants were less likely to seek out information after a correct initial decision. B On a within-subject level, a participant was more likely to be correct in the final decision after having decided to see the additional information and C on a between-subjects level participants who used the additional information more often tended to be more accurate in the final decision. D Higher cost of information lowered participant’s willingness to seek out additional information. E Participants reduced their information seeking as a function of their pre-decision confidence. A-E Circles indicate individual data point, while the box plots indicate group medians ± the middle quartiles whereby the whiskers indicate the upper and lower quartile respectively. *** p < .001.
in results might be driven by the different incentive structures between the tasks (in the task used for this chapter, rewards were not tied to reporting initial confidence but instead for maximizing performance in the final decision).

these analyses controlled for key demographic variables including age, gender, and education (see Figure 5.4A). I replicated this lowered tendency for dogmatic subjects to seek out information in a second, independent sample in study 2 ($\beta = -0.10, p = 0.039, \text{one-tailed, } R^2 = 0.01$), again in the absence of differences in initial decision accuracy ($\beta = -0.09, p = 0.13$) and confidence ($\beta = -0.03, p = 0.54$). These findings highlight that, even in the absence of motivational factors, dogmatic participants seek out less information before committing to a decision – even when this information would be helpful.

A key question arising from this finding is whether dogmatic individuals’ final accuracy and payoff suffered because of their lowered information search, or whether they simply sought information more efficiently. Here, a mediation analysis (Figure 5.4B) showed that more dogmatic participants were in fact less likely to be accurate in their final decision (total effect: study 1, $\beta = -0.11, p = 0.001$; study 2, $\beta = -0.09, p = 0.01$, one-tailed), and that this effect was fully mediated by a lowered willingness to seek information (mediation effect: study 1: $\beta = -0.08, p = 0.005$; study 2: $\beta = -0.05, \text{one-tailed } p = 0.038$; corrected direct effect, study 1: $\beta = -0.03, p = 0.097$; study 2: $\beta = -0.03, p = 0.12$). To better assess the overall strength of the relationship between dogmatism, information seeking and final performance, I pooled data from both studies to establish that the effect was stable across conditions (total effect: $\beta = -0.098, p = 0.0001$, mediation effect: $-0.062, p = 0.002$, corrected direct effect: $-0.035, p = 0.016$). This analysis revealed that the relation between dogmatism and final accuracy was partially mediated through information seeking, indicating that more dogmatic participants processed post-decision evidence less efficiently even after seeking it out. This result replicates findings from the previous chapter. Dogmatic participants also earned less money overall indicating that their lowered information seeking did not entail any strategic benefits (study 1: $\beta = -0.24, p = 0.008, R^2 = 0.02$; study 2: $\beta = -0.21, p = 0.009$, one-tailed, $R^2 = 0.01$).

### 5.3.5 Computational Modeling of Information Search

I next sought to develop a more detailed account of how dogmatic individuals’ trial-by-trial information seeking choices were informed by confidence judgments and the cost of information through a simple computational model. This model can be expressed as a logistic regression predicting the choice to seek information:

$$P(\text{InformationSeeking}) = \frac{1}{1 + \exp(- (\beta_0 + \beta_1 \times \text{Confidence} + \beta_2 \times \text{Cost}))}$$

The three $\beta$’s capture three independent behavioral phenomena (see Figure 5.5A): Differences in the model’s intercept, $\beta_0$, represent a general shift in willing-
Dogmatism manifests in lowered information search

Figure 5.4: Dogmatism is characterized by a reduction in information search, leading to less veridical judgements. A Dogmatism was predicted by a reduced willingness to seek out more information before committing to a decision, controlling for several demographic and task variables. Standardized beta coefficients ± standard error of predictors are presented for study 1 (left markers, N = 370) and study 2 (right markers, N = 364). Effects in study 2 were tested one-tailed based on the directional hypothesis derived from study 1. *p<.05, **p<.01, ***p<.001. B A reduction in information search mediated the less accurate overall final judgements in more dogmatic participants (mediation results for study 1 are presented in the figure, see main text for results from study 2).
ness to seek out information; $\beta_1$ represents how strongly participants’ information seeking choices are influenced by confidence; and $\beta_2$ indicates the influence of information cost on subjects’ willingness to seek out more information. I was primarily interested in whether any of these parameters were associated with individual differences in dogmatism, i.e., whether dogmatism was linked to a general tendency to seek out less information ($\beta_0$), a differential influence of confidence on information seeking ($\beta_1$) or an altered sensitivity to information costs ($\beta_2$). Due to a limited number of trials per participant I fitted this model hierarchically to ensure more reliable estimates. The relation between each parameter, $\beta$, and dogmatism was thereby estimated directly within this hierarchical framework, such that each individual’s parameter (indexed by i) was a function of a group mean ($\mu_\beta$) and their dogmatism score (Dogmatism$_i$). For instance, the relation between dogmatism and a general tendency to seek out less information can be formalized as follows:

$$\beta_{0,i} = \mu_{B_0} + \rho_0 \ast \text{Dogmatism}_i + \sigma_{B_0}$$

(5.5)

Here, $\rho_0$ describes the relation between dogmatism and $\beta_0$, while $\sigma_{B_0}$ represents individual variation in this parameter that is not explained by dogmatism. If the confidence interval of $\rho_0$ does not include zero, this indicates a significant association between dogmatism and $\beta_0$ [Moutoussis et al., 2018].

I found more dogmatic subjects had lower values of $\beta_0$ (Figure 5.5D; CI $\rho_0 = -.40, -.07$), in accordance with my model-agnostic findings that dogmatic participants show lower information-seeking behavior. While I found no association between dogmatism and cost sensitivity (CI $\rho_0 = -.06, .07$), dogmatism was associated with higher values of the confidence parameter $\beta_1$ (CI $\rho_1 = .02, .27$). Because $\beta_1$-values were generally negative (see Figure 5.5B, for distribution of individual parameter values), this positive shift suggests that more dogmatic individuals’ information search was less coupled to fluctuations in subjective confidence compared to less dogmatic individuals. In other words, more dogmatic participants were less likely to use feelings of confidence or uncertainty to guide their search for more information. Together, this dual shift in both $\beta_0$ and $\beta_1$ parameters combine to produce marked differences in information search under low confidence (high uncertainty). On these trials, less dogmatic individuals were more likely to (adaptively) seek out new information, whereas highly dogmatic individuals were not. In contrast, both high and low dogmatism participants showed similar profiles of information-seeking behavior when they were highly confident in their decision (Figure 5.5E-F).

### 5.3.6 Information Search and Other Factor Scores

In light of a long-standing debate over diverging cognitive profiles of liberals and conservatives [Jost, 2017; Greenberg & Jonas, 2003], I also investigated the relationship between information search and political orientation. Here, I found that
Chapter 5. Dogmatism manifests in lowered information search

Figure 5.5: Individual differences in confidence-driven information search as captured by a computational model. A The decision to seek additional information was captured using a model with three parameters: an intercept $\beta_0$, a confidence parameter $\beta_1$, and a cost-parameter $\beta_2$ (not depicted here). B Distribution of individual level parameters displaying the generally negative influence of higher confidence ($\beta_1$) and higher cost ($\beta_2$) on information search. C I captured dogmatism-related differences in these model parameters through a hierarchical fitting procedure, whereby each parameter’s empirical prior varies as a function of another set of parameters $\rho$ that encode the influence of subjects’ dogmatism scores in a hierarchical estimation scheme. D Posterior distribution of embedding parameters $\rho$ encoding dogmatism-driven shifts in parameter means. I found a dogmatism-related decrease in the parameter capturing baseline information search ($\rho_0$) and an increase in the parameter capturing the tuning of information search to participants’ confidence ($\rho_1$). No effect of dogmatism on the cost parameter was observed ($\rho_2$). The dotted vertical line represents a null effect. *p<.05, **p<.01. E-F Dogmatic individuals seek out less information than moderates when they are uncertain. To visualize this effect, I compared the 10% most dogmatic participants to the remainder of the sample. I plot (E) model predictions and (F) actual data (medians with upper/lower quartiles), averaged over both levels of information cost.
position on the political spectrum (right vs. left) was not predicted by a willingness to seek information (study 1: $\beta = -.07$, $p = .19$; study 2: $\beta = -.07$, $p = .23$).

Similarly, policy-specific political belief superiority was not related to changes in information seeking (study 1: $\beta = .03$, $p = .62$; study 2: $\beta = -.07$, $p = .24$).

5.4 Discussion

In this chapter I show that dogmatic individuals are less likely to seek out additional information before committing to a decision. By foregoing this opportunity, they in turn tend to form less accurate overall judgements. Modeling analyses revealed that two factors drove dogmatic individuals’ altered information seeking: (a) a shift in the general willingness to seek information and (b) a decreased influence of confidence on information seeking behavior. Together, these effects gave rise to a distinct pattern: whereas dogmatism had little effect on information seeking after high confidence decisions, more dogmatic subjects were less likely (relative to moderates) to seek out additional information when they were uncertain about their decision.

The mediation analysis between dogmatism, final accuracy and information seeking revealed the reduction of final accuracy in more dogmatic participants was partially mediated by their reduced tendency to seek out additional information. However, even when controlling for the tendency for information seeking, more dogmatic participants performed worse in the final decision. This points towards a reduction in the sensitivity to post-decision evidence in dogmatic participants, even when they choose to seek out additional evidence. This finding replicates results from the previous chapter, indicating a reduced sensitivity to post-decision evidence. Thus, the reluctance to incorporate new information in dogmatic participants can be described as a combination of a lowered tendency to seek out information combined with a reduced sensitivity to process the information if it is obtained.

A key aspect of these results is that these patterns of disadvantageous information-seeking were observed in a low-level perceptual decision-making task. This stands in contrast to previous studies on information seeking in the political domain which tend to solely rely on questionnaires or experimental tasks with overt political content (Long & Ziller 1965; Tappin et al. 2020). By capitalizing on the neutral valence and personal irrelevance of simple dot stimuli I was able to isolate uncertainty-driven information seeking behavior from possible confounding effects of motivated reasoning. Observing such an effect in this neutral setting is consistent with a proposal that domain-general cognitive factors contribute to real-world attitudes (Zmigrod et al. 2018; Zmigrod 2020; Rollwage et al. 2019). Nevertheless, in most real-world decision-making scenarios it is likely that both motivational and cognitive (uncertainty-driven) effects contribute to biases in information seeking (Sharot & Sunstein 2020), and it is interesting to consider that the latter may even
become magnified in the presence of affective influences.

The computational model fits revealed that while participants generally use internal signals of uncertainty (as assayed by confidence ratings) to guide information search, dogmatic individuals did so to a lesser extent. This points to a general alteration in the way that confidence guides actions, a process usually described as metacognitive control (Nelson & Narens, 1990). Metacognitive control is hypothesized to not only regulate information search, but also other phenomena in which effort must be weighed against accuracy, such as cognitive offloading (Risko & Gilbert, 2016) or speed-accuracy trade-offs (Desender et al., 2019a). From a theoretical perspective, metacognitive control complements metacognitive monitoring (Nelson & Narens, 1990) which describes a process that gives rise to, and updates, representations of confidence. However, while metacognitive monitoring has received considerable attention from a neural (Fleming et al., 2010; Vaccaro & Fleming, 2018) and individual differences perspective (Rollwage et al., 2018; Rouault et al., 2018), metacognitive control processes have received less attention. Such research might therefore provide fruitful for understanding the drivers of altered information search.

More dogmatic individuals displayed lowered information search in situations of uncertainty compared to their peers. At a single-trial level, this is consistent with basing an overall judgement on less evidence, leading to less accurate judgments overall. Because uncertain decisions are also less likely to be correct, this meant more dogmatic individuals were less likely to seek out contradictory evidence when they were wrong – a form of confirmation bias. Over a longer time horizon, and in the absence of external feedback (Rouault et al., 2019), such a self-reinforcing feedback loop might in turn entrain dogmatic individuals with a viewpoint that their initial judgments are already sufficiently optimal, such that investing in acquiring more information becomes unnecessary. A useful extension of the current work will be to investigate how more dogmatic individuals manage information search in situations that span more than one trial and require iterative learning. In such scenarios, adequately managing the exploration/exploitation trade-off is central to effective learning (Schulz et al., 2019; Hills et al., 2015; Boldt et al., 2019) such that small differences in a tendency towards or against uncertainty-driven information search may summate and lead to skewed representations of reality.

While there was no effect of dogmatism on $\beta_2$ (i.e. the sensitivity to information cost), it is possible that a categorical reluctance to pay for information affects the general tendency to seek out information (as captured by $\beta_0$). While I varied the cost of information between two levels, there was no condition in which information was completely free. Such a condition could help to further disentangle a general tendency to seek out less information from a categorical reluctance to pay for information. In future research, the inclusion of this condition could be useful to further understand the underlying mechanisms that drive reduced information seeking in
Chapter 5. Dogmatism manifests in lowered information search

dogmatic participants.

While a psychophysical approach provides the precise control required to characterize dogmatic individuals’ information search, the current task is necessarily contrived relative to real-world decision problems. It remains unknown whether the types of search behavior observed here are representative of real-world search behavior, for instance on the internet (Hills, 2019). However, I am cautiously optimistic about the generalizability of the current results, given the domain-general nature of the task and recent observations that real-life behavior adheres to cognitive models of uncertainty-based exploration (Schulz & Gershman, 2019). One difference between the current paradigm and real-world decisions is the guaranteed helpfulness of future information. The calculus changes when a first source is trustworthy, but future information might be unreliable. In that case, it might be adaptive to solely rely on one’s initial judgment, and refrain from seeking new information even when uncertain.

The causality of this relationship between information seeking and dogmatism is difficult to establish with correlational studies. It is possible that differences in information seeking predispose people to develop dogmatic views, as they will be less likely to encounter corrective and disagreeing information. However, it is also plausible that holding dogmatic views (and interacting in associated social circles) might reduce the willingness to seek out corrective information, as in certain environments it might not be seen as beneficial to change one’s mind in light of new evidence. Independently of the direction of causality (as so often in psychology it might be an interaction of the two), it is easy to imagine that such a reluctance to use new information will increase entrenchment of opinions and sustain dogmatic beliefs.

In sum, I highlight a generic resistance to seek out additional information in more dogmatic individuals, a difference that is most marked when initial decisions are uncertain. This is disconcerting in the current cultural landscape. While the internet has heralded access to a plethora of well-vetted information, fake news remains rife (Hills, 2019; Lemos et al., 2012). In such cases, the mere availability of correcting information might not be enough to prevent the formation of unsupportable beliefs in highly dogmatic individuals, because even feelings of uncertainty would not sufficiently trigger corrective information-seeking behavior. On a systemic level, such results indicate that the veracity of first contact with a news story is therefore critical (Pennycook & Rand, 2019). On an individual level, instilling successful uncertainty-based search may be enabled by the extension of training of metacognitive monitoring (Carpenter et al., 2019) to metacognitive control. Finally, my research shows that psychophysical paradigms in conjunction with computational modeling provide important tools for identifying mechanisms behind dogmatism, polarization and their consequences (Rollwage et al., 2019).
Chapter 6

Confirmation bias is adaptive when coupled with self-awareness

6.1 Overview

In the previous chapters, I have shown that high confidence induces a neural confirmation bias and that this altered processing of new information is associated with rigid, dogmatic and radical beliefs. These studies mainly focussed on the detrimental effects of selective information processing. In the last two chapters, I will shift the focus and investigate ways to alleviate these negative effects. Specifically, I will center my attention on the role that metacognitive abilities might play for the manifestation of negative effects caused by selective evidence integration. In this chapter, I use simulation-based modeling to theoretically show the impact that metacognition has on post-decision evidence integration. In chapter 7, I will put these theoretical considerations to an empirical test and investigate whether a metacognitive training procedure can improve openness for new information.

6.2 Introduction

As shown in the previous chapters, humans often resign to one-sided consideration of evidence. On societal level, such skewed information intake might lead to entrenched beliefs and societal polarization (Lilienfeld et al., 2009; Rollwage et al., 2019).

In chapter 4 I have shown that dogmatic participants are characterized by two cognitive alterations in the context of a perceptual decision-making task. First, dogmatic participants showed a reduction in metacognitive ability, manifesting as a selective overconfidence after making errors. Second, metacognitive ability (i.e. the accuracy of confidence judgments) was predictive of post-decision evidence integration, where people with poorer metacognition showed less sensitivity for corrective information.

These results indicate that confidence acts as an internal control signal that
guides future information processing (Atiya et al., 2019; Desender et al., 2018, 2019b; Meyniel et al., 2015a; Meyniel & Dehaene, 2017).

In chapter 3, I further showed that confidence strikingly modulated the extent of neural post-decision processing. Evidence accumulation was largely unbiased after low confidence decisions but displayed a strong confirmation bias after high confidence decisions. In other words, people appear especially resistant to corrective information when they hold high confidence in a wrong decision. However, when confidence is well aligned with performance – when metacognitive ability is high – such selective post-decision integration is likely to be less problematic, as people will tend to be open to new information when they need it (when they are less confident after making errors). This line of reasoning indicates that people’s metacognitive ability might be a crucial driver of the degree to which selective information processing leads to negative behavioural outcomes.

Here I test a hypothesis that selective information integration might be adaptive when coupled with high metacognitive ability. This proposal is in line with the broader hypothesis that the ubiquitous nature of selective evidence integration makes it unlikely that this cognitive characteristic is always maladaptive (Gigerenzer, 2008; Klayman & Ha, 1987). For instance, others have interpreted confirmation bias as a heuristic that reduces computational complexity (Evans, 1989) or allows for robustness against noise (Lefebvre et al., 2020; Tsetsos et al., 2016). Here I offer an alternative perspective: that confirmation bias is adaptive to the extent it is tempered by the metacognitive ability to effectively monitor and recognise when we might be wrong (Fleming et al., 2012). I use simulation-based modelling to compare different evidence integration strategies and test their respective performance. Specifically, I compared unbiased evidence integration with a simple confirmation bias as well as with a confidence-weighted confirmation bias (as observed in chapter 3). Since the performance of a confidence-weighted confirmation bias might depend on the reliability with which confidence judgments indicate choice accuracy, I also investigated the influence of metacognitive ability on the adaptiveness of confirmation bias.

6.3 Modelling behavioural impact of selective evidence integration strategies

Here I model a simple situation in which agents make a binary decision between two choice options based on noisy information drawn from a world state that is unknown to the agent. This situation is easily modelled with existing frameworks for characterising belief updating (Bronfman et al., 2015; Fleming et al., 2018; Talluri et al., 2018), and closely resembles common perceptual decision-making paradigms. This setting also acts as a minimal framework within which more complex decision
problems can be modelled, as exemplified by debates about the existence of climate change. For instance, to reach an opinion about whether human activity causes global warming (the ground truth), we have to form beliefs based on multiple noisy evidence samples (e.g. scientific publications and newspaper articles). Importantly, this process requires the updating of pre-existing beliefs whenever more evidence becomes available. In such a situation, Bayesian belief updating is often used as benchmark model (Mathys et al., 2011; O’Reilly et al., 2013). Here this model is used as “unbiased agent” against which the other evidence integration strategies will be compared. A Bayesian model keeps track of the graded belief for the choice by calculating the probability for the chosen option, given the evidence, compared to the alternative:

\[
P(\text{choice}|\text{evidence}) / P(\text{alternative}|\text{evidence})
\] (6.1)

For simplicity, I simulate a situation in which participants only receive two samples of information: \( X_{\text{pre}} \) represents the initial information and \( X_{\text{post}} \) represents the additional (or post-decision) evidence. Both \( X_{\text{pre}} \) and \( X_{\text{post}} \) are sampled from normal distributions:

\[
X_{\text{pre}} \sim N(\mu, \sigma_{\text{pre}}^2)
\] (6.2)

\[
X_{\text{post}} \sim N(\mu, \sigma_{\text{post}}^2)
\] (6.3)

\[
\mu = [-1, 1]
\] (6.4)

where the mean (\( \mu \)) of these distributions corresponds to the actual world state that needs to be inferred. While pre- and post-decision evidence distributions have the same mean (i.e. indicate the same underlying world state), they might differ in their variance. The variance represents the reliability of the information, with higher variances indicating less reliable information. I simulate different reliabilities of pre and post-decision evidence (see Figure 6.1A) as the consequences of a confirmation bias might depend on this balance.

After receiving initial information, agents make an initial decision which depends solely on \( X_{\text{pre}} \). The decision equals 1, if \( X_{\text{pre}} \) has a positive value (\( X_{\text{pre}}>0 \)), whereas the decision equals -1 if \( X_{\text{pre}} \) has a negative value (\( X_{\text{pre}}<0 \)). A sense of confidence in this initial decision is derived by calculating the log-odds in favor for the chosen world state:

\[
\text{LogDir}_{\text{pre}} = 2 * \mu * X_{\text{pre}} / \sigma_{\text{pre}}^2
\] (6.5)

These log-odds can be transformed into a confidence rating between 0 and 1 as follows:

\[
\text{confidence}_{\text{initial}} = 1 / (1 + e^{-\text{LogDir}_{\text{pre}}})
\] (6.6)
Chapter 6. Confirmation bias is adaptive when coupled with self-awareness

After the initial decision the agent receives additional information $X_{\text{post}}$. In order to reach a final decision, both evidence samples can be integrated in an unbiased Bayesian fashion by simply summing the log-odds:

$$\log \text{Dir}_{\text{final}} = \log \text{Dir}_{\text{pre}} + \log \text{Dir}_{\text{post}} \quad (6.7)$$

Note that by adding the log-odds of pre- and post-decision evidence, the certainty/reliability of these two evidence samples is implicitly considered, i.e. the evidence is combined in line with Bayesian principles. The final decision depends on the sign of the posterior log-odds ($\log \text{Dir}_{\text{final}}$). If the final decision corresponds to the actual state of the world, the agent can be said to have formed an accurate belief and performs the task correctly. The average decision accuracy over many simulated trials forms a measure of the agent’s performance under different decision strategies (I simulated 20,000 trials for each agent and each level of reliability). Depending on the different reliabilities of $X_{\text{pre}}$ and $X_{\text{post}}$, an unbiased Bayesian observer will achieve different final performances (see Figure 6.1A), with more reliable information yielding better performance.

A confirmation bias can be modelled as an altered incorporation of the post-decision evidence dependent on whether this new information confirms or disconfirms the initial decision:

If $\text{sign}(X_{\text{post}}) = \text{sign}(\text{decision}_{\text{initial}})$:

$$\log \text{Dir}_{\text{final}} = \log \text{Dir}_{\text{pre}} + w_{\text{confirm}} \times \log \text{Dir}_{\text{post}} \quad (6.8)$$

Else if $\text{sign}(X_{\text{post}}) \neq \text{sign}(\text{decision}_{\text{initial}})$:

$$\log \text{Dir}_{\text{final}} = \log \text{Dir}_{\text{pre}} + w_{\text{disconfirm}} \times \log \text{Dir}_{\text{post}} \quad (6.9)$$

As observed empirically (Rollwage et al., 2020a; Talluri et al., 2018), confirmation bias manifests as an amplification of confirmatory evidence ($w_{\text{confirm}} > 1$) and a reduction in the processing of disconfirmatory evidence ($w_{\text{disconfirm}} < 1$). This form of confirmation bias can range from .5 (no bias) to 1 (where a maximal confirmation bias is obtained when processing of disconfirmatory information is abolished).

$$w_{\text{confirm}} = \text{confirmation bias} \times 2 \quad (6.10)$$

$$w_{\text{disconfirm}} = (1 - \text{confirmation bias}) \times 2 \quad (6.11)$$

the $ws$ vary between 0 and 2, whereby a value of 0 represents an absence of processing and a value 2 represents maximal processing of post-decision evidence. Values of 1 represent unbiased information processing.
Chapter 6. Confirmation bias is adaptive when coupled with self-awareness

Figure 6.1: Comparison of agents’ performance with different biases in information processing. A Performance of an agent that integrates initial and additional information in an unbiased (i.e. Bayesian) manner. Depending on the evidence strength, this agent shows different levels of accuracy, with better performance when both evidence samples are strong/reliable. B Difference in performance between an unbiased agent and a confirmation bias agent as function of the reliability of the initial and additional evidence. A confirmation bias has especially detrimental effects when initial evidence is relatively weak. C Comparison of agents with a simple confirmation bias and a metacognitive agent against an unbiased agent (detrimental performance of zero would indicate the same performance as an unbiased agent), as a function of confirmation bias strength. Here, performance is averaged over all combinations of initial and additional evidence strengths. The vertical line indicates the strength of confirmation bias used in panels B and D. D Difference in performance between an unbiased agent and a metacognitive agent as a function of the reliability of initial and additional evidence. Overall, the metacognitive agent shows only a relatively small disadvantage in comparison to an unbiased agent. In comparison to a simple confirmation bias, the metacognitive agent suffers less detriment in situations with weak initial evidence. B & D Dark colours indicate more detrimental performance of the confirmation bias strategies when compared to unbiased evidence integration.
6.4 Results

I modelled agents with gradually increasing levels of confirmation bias and investigated their performance relative to an unbiased agent (see Figure 6.1C). As hypothesized, a selective accumulation of information results in detrimental performance, with higher levels of confirmation bias leading to a more pronounced detriment. This effect was most starkly present when the agent received relatively weak initial information but reliable post-decision evidence (see Figure 6.1B). In other words, a confirmation bias has its most damaging effects when initial beliefs are already based on weak information, as the bias prevents the incorporation of new and potentially more reliable corrective information.

As shown in chapter 3, a modulation of confirmation bias by confidence can be empirically observed such that participants were relatively unbiased in their use of new evidence when less confident, but showed an enhanced confirmation bias after high confidence decisions. Here I also simulate an agent that shows a modulation of confirmation bias by initial confidence which I term a “metacognitive” agent. Such an agent shows a canonical confirmation bias when it is confident in an initial choice (confidence=1), but is unbiased when unsure (confidence=.5)

\[ w_{\text{Conf}}^{\text{confirm}} = \text{baseline processing} + (\text{confirmation bias} \times \text{confidence}) \]

\[ w_{\text{Conf}}^{\text{disconfirm}} = \text{baseline processing} - (\text{confirmation bias} \times \text{confidence}) \]

A metacognitive agent differs from a Bayesian agent in that the initial confidence directly modulates the extent to which post-decision evidence is incorporated (instead of just being used for combining prior and new information). For a Bayesian agent there is also a sense in which confidence “weights” the subsequent incorporation of evidence, in that a highly confident decision will require more disconfirming evidence to be overturned. But such updates are in keeping with the linear accumulation of the log-odds of one or the other hypothesis. In contrast, the metacognitive agent shuts down the processing of disconfirmatory evidence to the extent it is confident, representing a non-linear effect of confidence on the incorporation of post-decisional log-odds. Thus, this metacognitive agent shows similarities with a circular inference model developed by Bouttier et al. (2019), in which prior beliefs directly corrupt new incoming evidence.

Interestingly, such a confidence-weighted confirmation bias outperforms a simple confirmation bias in all settings and only shows slight impairments in relation to an

\[ w_{\text{Conf}}^{\text{confirm}} = 1 + (w_{\text{confirm}} - 1) \times ((\text{confidence} - .5) \times 2) \]

\[ w_{\text{Conf}}^{\text{disconfirm}} = 1 + (w_{\text{disconfirm}} - 1) \times ((\text{confidence} - .5) \times 2) \]

\(^1\)the presented equations are not absolutely correct and serve mainly illustration purposes. The actual equations are as follows:
unbiased agent (see Figure 6.1C). When comparing a confidence-weighted confirmation bias to a simple confirmation bias, we find that a metacognitive agent clearly outperforms a simple confirmation bias in situations when the initial evidence is weak (see Figure 6.1D). While a simple confirmation bias has the strongest decrement in performance in these situations, a metacognitive agent “realizes” when having weak initial evidence (by assigning low confidence to these decisions) and thus shows a more equal sensitivity to confirming and disconfirming information.

Up until now I have assumed that agents calculate confidence in an initial choice by directly evaluating the reliability of evidence that informed the decision. In this case, the agent uses exactly the same information for both their decision and simulated confidence rating. However, empirical studies have shown that confidence judgments and decisions may rely on (partially) separate information (Cortese et al., 2016; Odegaard et al., 2018; Peters et al., 2017; Zylberberg et al., 2012; Miyoshi & Lau, 2020). The type (and quality) of evidence that informs confidence judgments in turn determines the degree to which confidence is a good indicator of belief accuracy (Maniscalco & Lau, 2012). This correspondence between confidence and performance can be formally quantified as the ratio of meta-$d'/d'$ (known as metacognitive efficiency) within a signal detection theoretic framework, where meta-$d'$ reflects metacognitive sensitivity, and $d'$ reflects primary task performance. Several reasons for a dissociation between meta-$d'$ and $d'$ have been suggested. For instance, confidence may reflect a noisy read-out of the decision evidence or a decline of decision evidence in working memory before a confidence judgment is elicited (Maniscalco & Lau, 2012), leading to a meta-$d'/d'$ ratio of less than 1. On the other hand, confidence might be informed by evidence that was not available at the time of decision (Moran et al., 2015; Murphy et al., 2015; Pleskac & Busemeyer, 2010), or on correlated evidence accumulated in parallel (Fleming & Daw, 2017; Miyoshi & Lau, 2020) both of which may lead meta-$d'/d'$ ratios to surpass 1.

Here, I assessed the degree to which metacognitive efficiency influences the performance of the metacognitive agent. To do so, I relaxed the assumption that confidence is directly derived from the evidence informing the initial decision. Instead, the evidence informing decisions ($X_{\text{pre}}$) and confidence estimates ($X_{\text{conf}}$) were distinct but correlated, and I allowed the reliability of the confidence ($\sigma^2_{\text{conf}}$) and decision ($\sigma^2_{\text{pre}}$) samples to differ. Hence, $X_{\text{pre}}$ and $X_{\text{conf}}$ were sampled from a bivariate normal distribution with the same mean, and the following covariance:

$$
\begin{bmatrix}
    X_{\text{pre}} \\
    X_{\text{conf}}
\end{bmatrix} \sim N(\mu, \Sigma) \tag{6.16}
$$

$$
\Sigma = 
\begin{bmatrix}
    \sigma^2_{\text{pre}} & \rho \star \sigma_{\text{pre}} \star \sigma_{\text{conf}} \\
    \rho \star \sigma_{\text{pre}} \star \sigma_{\text{conf}} & \sigma^2_{\text{conf}}
\end{bmatrix} \tag{6.17}
$$

$\Sigma$ describes the covariance between $X_{\text{pre}}$ and $X_{\text{conf}}$, whereby $\rho$ represents the correlation between these variables. As suggested by Fleming & Daw (2017), in
Chapter 6. Confirmation bias is adaptive when coupled with self-awareness

this situation confidence can be inferred based on a combination of \( X_{\text{conf}} \), the initial
decision and the covariance of \( X_{\text{pre}} \) and \( X_{\text{conf}} \) (see the Appendix of Fleming & Daw
(2017) for further details on this calculation):

\[
\text{Confidence} = P(\text{decision}_{\text{initial}} = \text{worldstate} \mid X_{\text{conf}}, \text{decision}_{\text{initial}}, \Sigma) \quad (6.18)
\]

Importantly, modelling separate samples of \( X_{\text{conf}} \) and \( X_{\text{pre}} \) allows for dissociations
between meta \( - d' \) and \( d' \), thus making it possible to simulate varying degrees of
metacognitive efficiency.

![Figure 6.2:](image)

A decrease in the reliability of the evidence informing the confidence rating
\( (\sigma^2_{\text{conf}}) \) naturally reduces metacognitive efficiency and leads confidence judgments to
be less reliable predictors of choice accuracy Fleming & Daw (2017), when keeping \( \rho \)
constant (for this condition I fixed \( \rho=.8 \), however, these findings are not dependent
on the specific value of \( \rho \) ). The simulation shows that with respect to post-decision
evidence integration, reduced metacognitive efficiency leads to impaired performance
(see left side of Figure 6.2A), as the weighting of new information by confidence is
less effective. I also modelled a situation in which confidence and decision information have the same reliability \( (\sigma^2_{\text{pre}}=\sigma^2_{\text{conf}}) \), but the evidence samples have a variable
correlation \( (\rho=|.1-.8|) \). Such a setting can result in increased metacognitive efficiency, as the “confidence rater” has additional information on which to base an
evaluation of the decider \cite{Fleming & Daw 2017}. Surprisingly, when metacognitive efficiency is high ($meta - d'/d' > 1$) a metacognitive agent can even outperform an unbiased agent (see Figure 6.2A right side). This result is striking as it suggests that a selective integration of information is not necessarily a “bias” but represent an advantageous strategy for achieving optimal performance in the context of a realistic cognitive architecture (i.e. one in which metacognition is particularly efficient).

6.5 Discussion

Here I investigated the effects of confirmation bias on the accuracy of belief formation. The central proposal is that when confirmation bias is a feature of a self-aware (metacognitive) agent, it ceases to be detrimental, and may even become adaptive. I used simulation-based modelling to compare the performance of agents with different forms of confirmation bias against an unbiased agent. In general, a simple (non-metacognitive) confirmation bias showed detrimental effects compared to an unbiased agent in all settings. In comparison, a metacognitive agent which modulates the extent of a confirmation bias by its confidence (as shown empirically in human observers in chapter 3) outperformed a simple confirmation bias agent, and was in many cases not substantially worse than an unbiased agent. In turn, by simulating varying degrees of self-awareness (metacognitive efficiency), I found that the performance of this metacognitive agent was highly sensitive to the level of self-awareness. Strikingly, a metacognitive agent with high self-awareness could in some cases even outperform an unbiased agent, indicating that selective information processing might be particularly adaptive when coupled with good metacognitive abilities.

Why should a metacognitive agent have an advantage in this case? The core mechanism appears to be the capacity of a confidence estimate to provide a “second look” on a decision, similarly to how an external adviser might give us a separate view on a topic. The adaptivity of this information processing mechanism crucially depends on the agent’s metacognitive ability – as agents with good metacognition provide the most effective “internal” advisory signals. Interestingly, however, a confidence-weighted confirmation bias outperformed a simple confirmation bias in all settings, even when metacognitive ability was relatively low (see Figure 6.2B). Conversely, when metacognition is high, this evidence-weighting strategy can even outperform an unbiased agent. These results are in accordance with a view that biases evolved for an evolutionary reason and may turn out to be adaptive when considered in the context of the agent’s environment, including its broader mental toolkit \cite{Gigerenzer 2008}.

By incorporating confirmation bias as part of a broader cognitive architecture in which different mental processes can interact with each other (e.g. decisional
and metacognitive processes), selective information processing may become adaptive when compared to the same “bias” considered in isolation. In the same spirit, it has been argued that heuristics that often appear as biases in simple and constrained laboratory tasks become beneficial in more complex environments (Pleskac & Hertwig, 2014). More broadly, this study indicates that considering cognitive biases in isolation from other mental processes might lead to the wrong conclusions about the impact of a particular cognitive feature on behaviour.

However, a behavioural benefit for confirmation bias was only present when simulating agents with “hyper” metacognitive efficiency (i.e. \( \text{meta} - d'/d > 1 \)), such that metacognition becomes more accurate than first-order task performance. This might seem odd at first glance, as it implies that the system is not using all the information available to it at the time of making an initial choice, and only afterwards becomes more sensitive to whether it was right or wrong. However, this kind of pattern is commonly observed in empirical data (Moreira et al., 2018; Fleming & Daw, 2017), and is thought to be driven either by additional post-decisional processing or (as simulated here) a parallel processing of information that allows the system to rapidly detect and correct errors (Moreira et al., 2018). The capacity for rapid error detection is well established in human studies (Murphy et al., 2015; Rabbitt, 1966; Yeung & Summerfield, 2012) and thus it is reasonable to assume that hyper metacognitive efficiency may be common in the healthy population.

Importantly, the confidence system does not need to be more reliable than the decision system to achieve high metacognitive ability: it need only to rely on partially separate information. In this respect, these results also contribute to a long-running debate over why it might be useful for the brain to encode a separate confidence signal that is distinguishable from decision evidence (Meyniel et al., 2015b; Maniscalco et al., 2016). A metacognitive agent that realizes its own mistake (and assigns low confidence to this decision) will then tend to become more open to new information, due to the confidence weighting applied to selective information processing (and in fact, if confidence falls below 0.5 in a two-choice scenario the simulations would predict that metacognitive agents would show a “disconfirmation” bias). These results suggest that this benefit is only seen when metacognition is partly independent of first-order cognition.

I only simulated a narrow range of situations which consider only one outcome variable (belief accuracy). In this context only “hyper” metacognitive abilities led to benefits of a confirmation bias. However, confirmation bias might have other benefits for scenarios or outcome variables that are not considered here (e.g. saving time and processing resources, or avoiding the effort of changing one’s actions). Therefore, it is possible that confirmation bias might be beneficial when considering broader contexts even in the absence of high metacognitive ability.

Selective information processing has been assumed to lead to skewed, entrenched
and potentially inaccurate beliefs about a range of societal and political issues (Rollwage et al., 2019; Zmigrod, 2020). As shown here, the detrimental effects of selective information processing are highly dependent on people’s self-awareness. Therefore, metacognitive deficits might represent a core driver of polarised or radical beliefs, due to their consequence for maladaptive confirmation bias. Interestingly, this is in line with the empirical observations from chapter 4, showing that more dogmatic participants show reduced metacognitive sensitivity which in turn is predictive of reduced post-decision evidence processing. Recognising metacognition as a central driver of belief polarization may make it possible to develop new strategies for debiasing decision-making (Rollwage et al., 2019). Excitingly, there are existing interventions that have been shown to boost people’s metacognitive ability (Baird et al., 2014; Carpenter et al., 2019). These results indicate that cognitive training which improves domain-general self-awareness and metacognitive efficiency may help to alleviate the negative behavioural outcomes of selective information processing. Metacognitive training might be a feasible and efficient interventions that could foster resilience against misinformation and counteract belief polarization.
Chapter 7

Boosting metacognition to reduce confirmation bias

7.1 Introduction

As shown in the previous chapter, selective information processing does not necessarily have detrimental effects. More specifically I have shown that the consequences of a confirmation bias crucially depend on a person’s metacognitive ability. The capacity to effectively assign confidence to beliefs and detect own mistakes appears to be a form of self-awareness that is especially important for promoting openness to new information (Lilienfeld et al., 2009).

As shown in chapter 3, these ideas are also borne out by empirical results, showing that confirmation bias is modulated by participants’ confidence. Participants process new information in an unbiased manner when they have low confidence in an initial decision (see Figure 7.1A upper panel). However, holding high confidence in a decision induces a selective gating for choice-consistent information and abolishes processing of disconfirming information (see Figure 7.1A upper panel). These results were supported by MEG recordings showing that confidence directly altered the neural sensitivity for new information. This key role of confidence in the manifestation of confirmation bias might be a key to unlock the puzzle of how to counteract this pervasive cognitive bias. In other words, I expect selective information processing to have the most detrimental effects when people are highly confident in a wrong belief and intervention strategies should be most effective if they can help to avoid this situation.

In line with the theoretical work presented in the previous chapter, empirical findings identified metacognitive ability as a crucial factor that determines the amount of confirmation bias a person displays (Rollwage et al., 2018). People with low metacognitive ability will have poor confidence alignment, i.e. they often do not realize when they are making a mistake and assign high confidence to these decisions leading to reduced processing of corrective information on error trials, exactly when
it is most needed (see Figure 7.1A lower panel). In comparison, high metacognitive ability describes good confidence alignment, which should boost the processing of post-decision evidence (see Figure 7.1A lower panel). These effects play out over many choices, as high metacognition leads to more decisions in which confidence is well aligned and thus participants are receptive to post-decision evidence, making them on average less susceptible to the negative effects of confirmation bias (see Figure 7.1A).

Importantly, metacognitive ability is not a fixed trait but a skill that can be nourished by cognitive training (Baird et al., 2014; Carpenter et al., 2019). Direct, repeated feedback about task performance and confidence alignment (a form of feedback that is rare in everyday life) has been shown to improve participants’ metacognitive abilities by enabling them to generate more veridical confidence judgments. Importantly, these training effects generalize to untrained tasks, indicating that this procedure could be an avenue to increase metacognition in a domain-general manner and through this potentially reduce the negative consequences of confirmation bias in a variety of scenarios. This seems especially relevant as previous attempts to design robust and generalizable interventions to counteract confirmation bias have led to mixed results (Lilienfeld et al., 2009; Willingham, 2008; Lord et al., 1984; Kahneman et al., 1982; Milkman et al., 2009).

Here I test a hypothesis that cognitive training can increase participants metacognitive ability and through this reduce the negative effects of confirmation bias.

Participants underwent an initial assessment of their post-decision evidence processing before conducting the training procedure. In this session, participants performed a simple perceptual task (the same task as used in chapter 3) in which they had to judge the motion direction of a cloud of dots, while we monitored their task performance and EEG to derive behavioural and neural metrics of post-decision evidence integration. Participants were presented with a sample of moving dots (pre-decision evidence) before indicating their initial decision (motion to the left or right) and their confidence in this choice (see Figure 7.1B). They were then presented with a helpful second sample of moving dots (post-decision evidence), before making a final choice and providing a confidence estimate. The degree to which the second sample of moving dots affects the final choice is used as an indicator of post-decision evidence processing.

After the initial assessment, participants completed 7 training sessions, equally spread over a period of two weeks (see Figure 7.1B). During each training session, participants conducted a dot motion discrimination task and rated their confidence (similar to the task during the initial assessment but without post-decision evidence). Every 10 trials, participants received feedback about their task performance. I tested two different feedback procedures: in group 1 participants only received feedback about their task performance, whereas in group 2 they received feedback about their
Chapter 7. Boosting metacognition to reduce confirmation bias

Task performance and confidence alignment (Figure 7.1B, see Methods for more detailed description of the feedback). The more detailed feedback provided to group 2 has been shown to increase metacognitive ability (Carpenter et al., 2019). After the training, participants were invited back to the lab to conduct a final assessment, again measuring post-decision evidence processing on both a behavioural and neural levels.

Figure 7.1: Training hypothesis and task design. A The upper panel shows the influence of confidence on processing of confirmatory and disconfirmatory evidence, reproducing the main results from chapter 3. This upper panel shows the processing of information on a single trial level. Importantly, confirming post-decision evidence was received after correct initial decisions while disconfirming information was received after initial mistakes. The dotted lines represent trials with metacognitively incorrect confidence alignment (e.g. high confidence when being wrong) that are predominantly experienced by people with low metacognitive ability, whereas the solid lines represent trials with correct confidence alignment that are predominantly experienced by participants with high metacognitive ability. The lower panel shows the predicted sensitivity for post-decision evidence from people with high and low metacognitive ability as a result of their differential confidence alignment. In this lower panel, I present the sensitivity expected for a participant averaged over many trials. High metacognitive ability will lead people to be more often in a situation where confidence is well aligned and thus post-decision evidence is beneficial. B The middle panel shows the overall study structure, consisting of two assessment sessions before and after the training respectively, and seven training sessions that were conducted online. The upper panel shows the perceptual task that was performed in combination with EEG recordings before and after the training procedure in order to obtain behavioural and neural measures of post-decision evidence processing. The lower panel shows the perceptual task that was performed during the training sessions.

7.2 Methods

7.2.1 Participants

I analysed data from 40 participants (M_age = 24.40; SD_age = 5.93; 26 female) that completed all training and assessment sessions. Data of two participants could not
be analysed because they performed at chance level at the initial assessment session (and were thus not included in the training procedure and the final assessment), and data of one participant could not be analysed due to problems with trigger recordings for the EEG analysis. For all sessions combined, participants received an average payment of £118.77 (SD=£6.35, range £100-£130). For the initial assessment participants received a basic payment of £15 and a performance-based bonus of up to £10 (see below details of the performance based incentivization), whereas for the final assessment they received a basic payment of £20 a performance-based bonus of up to £15 (to ensure that participants would come back after the training procedure). Payment for the training sessions was entirely performance-based (to ensure motivated participation in the training) with an average bonus payment of £8.69 per session. The study was approved by the Research Ethics Committee of University College London (8231-001) and all subjects gave written informed consent.

7.2.2 Stimuli

7.2.2.1 EEG sessions

The psychophysical task used in the initial and final assessment session (EEG sessions) was an adaptation of the task used in chapter 3, and programmed in MATLAB 2012a (Mathworks Inc., USA) using Psychtoolbox-3.0.14. Stimuli were random dot motion kinetograms (RDKs), viewed at a distance of approximately 60 cm. The stimuli were presented using a 240Hz refresh rate monitor (Alienware Full HD Gaming Monitor). The RDKs were clouds of white dots (0.15° diameter) within a white circular aperture with a radius of 10° on a grey background. The RDK lasted for 500 ms. The direction of motion was rightward or leftward along the horizontal meridian. The speed of movement was 5° per second and the density of dots in the whole experiment was set to 60 dots per degree. Dots moved coherently in both the target direction and the opposite direction. The remaining dots moved randomly (percentages described below). The proportion of dots moving in the incorrect direction was set to 10% and the proportion moving in the correct direction was a higher percentage, staircased to ensure the targeted performance level (71% correct). The post-decision evidence had two coherence levels. Weak post-decision had the same coherence level as the pre-decision evidence, whereas strong post-decision evidence stimuli were derived by multiplying this coherence level (in the correct direction) by a factor of 1.3.

I intended to use rapid frequency tagging (Zhigalov et al., 2019; Drijvers et al., 2020) to potentially improve EEG decoding performance. Therefore, I presented the dots moving in one direction with a sinusoidal flicker frequency of 30Hz and the dots moving in the opposite direction with a sinusoidal flicker frequency of
40Hz. Importantly, both flickers were above the perceptual threshold and thus not consciously perceivable, so that participants had the impression of stable dot presentation. The assignment of each frequency was completely counterbalanced between left and right motion and correct versus incorrect motion direction, with pseudo-randomized assignment on each trial, to ensure no confounding effects on behaviour. However, since this manipulation did not induce the expected steady-state evoke potentials (Vialatte et al., 2010), I did not use frequency information for the decoding analysis.

7.2.2.2 Training sessions

The training sessions were conducted online and participants accessed the task via their web-browser. The task was programmed in JavaScript using JsPsych (version 5.0.3) and was hosted on the online research platform Gorilla (https://gorilla.sc/). Participants were required to enter full-screen mode to complete the task. The presented stimuli were also RDKs, using the JsPsych toolbox developed by Rajananda et al. (2017). I designed the RDK for the training as similar as possible to the RDKs used in the assessment sessions. The RDKs were clouds of 150 dark dots on a white background, lasting for 500 ms. The direction of motion was rightward or leftward along the horizontal meridian. Dots moved coherently in both the target direction and the opposite direction. The remaining dots moved randomly (percentages described below). The proportion of dots moving in the incorrect direction was set to 10% and the proportion moving in the correct direction was a higher percentage, staircased to ensure the targeted performance level (71% correct). No post-decision evidence was presented in the training task.

7.2.3 Procedure

This experiment involved three phases: an assessment session before the training (including EEG recordings), 7 training sessions, and an assessment session after the training (including EEG recordings), giving a total of 9 sessions. The initial and final assessment were conducted in the laboratory and were separated by two weeks. In both the initial and final assessment, participants completed a post-decision evidence integration task while EEG was recorded. This assessment was used to derive behavioural and neural measures of post-decision evidence integration.

7.2.3.1 EEG Sessions

Participants first performed 90 trials of a calibration phase before performing the main task which consisted of 352 trials.

In the calibration phase, subjects judged whether the dots were moving to the left or to the right side of the screen, without rating their confidence or seeing
additional post-decision evidence. The response had to be given within 1.5 sec after stimulus offset. The coherence of the target direction was adapted using a 2-down 1-up staircasing procedure designed to obtain 71% performance (García-Pérez, 1998). After the calibration phase the the staircasing was continued to ensure stable performance.

In the main task, participants were first presented with a moving dot stimulus before they made a left versus right decision and gave a binary high/low confidence rating. After this initial decision, participants received a second sample of moving dots (i.e. post-decision evidence) which was always in the same (correct) direction as the pre-decision evidence presentation, but of variable strength. Subjects were instructed that this evidence was bonus information that could be used to inform their final decision and confidence. After the post-decision evidence, participants were again asked to judge the motion direction and indicate their confidence.

To disentangle participants’ decisions (left/right and high/low confidence) from the motor responses they had to perform (pressing the up or down key on a key pad), the mapping between decision options and key presses was randomized. Specifically, on any given trial leftward motion could be indicated by pressing the up key and on another trial by pressing the down key. Similarly, high confidence could be indicated in one trial by pressing the up key and in a different trial by pressing the down key. The mapping between decisions and motor responses was revealed once responding was possible, by presenting the letters L or R (and H or L for confidence ratings) above/below the horizontal plane. This approach ensured that decoding of motion direction was not trivially confounded by motor preparation signals. Additionally, I introduced delays of 350ms after the presentation of each stimulus but before participants were informed about the response mappings to allow decoding analysis to be applied in a time window when subjects could form an abstract decision about motion direction but were not yet able to prepare a response.

7.2.3.2 Training sessions

The experiment involved seven training sessions, spread over a period of 2 weeks, each taking approximately 20 minutes. Subjects were pseudo-randomly allocated into two groups, with groups differing in their feedback type. Each session consisted of 270 trials presented in 27 blocks. Participants performed a random dot motion task and rated their confidence on each trial. All subjects received feedback regarding their performance every 10 trials. Group 1 received feedback on their perceptual performance while group 2 received feedback on perceptual performance and accuracy of their confidence ratings (see below for further explanation of the feedback conditions). Participants received a bonus payment based on the metric that they received feedback on (group 1: task performance, group 2: task performance combined with confidence accuracy). This bonus payment was accumulated
over all training sessions and paid to the participants once they came back for the final EEG assessment. On completion of the first and last training session, subjects also completed several questionnaires (discussed below).

7.2.4 Scoring and bonus payment: EEG sessions

During EEG sessions, participants were asked to rate their confidence as a subjective probability of being correct. They were awarded points for the accuracy of their final decision, using a Quadratic Scoring Rule (Brier, 1950) which incentivised them to accurately align their confidence and accuracy.

The points obtained on each trial were aggregated and subjects were given a £1 bonus payment for every 3000 points earned during the initial assessment session and for every 2250 points in the sessions after the training. Subjects were informed of their current points total after each block. Participants received no other performance feedback during the EEG sessions.

7.2.5 Feedback and bonus payment: Training sessions

During the training sessions, the feedback and incentivization differed between the two groups. In group 1 participants received a bonus payment based on their perceptual performance. Since the performance was staircased throughout each session, improvements in task performance led to increased stimuli difficulty. Therefore, participants were rewarded by a multiplication of stimulus difficulty and choice accuracy (e.g. 70% of maximal difficulty and 10 correct decisions: \(0.7 \times 10 = 7\) points), with a correct decision leading to more points when the task was difficult. Participants were fully informed about this incentivization scheme and received feedback based on this reward structure. They were presented with the average difficulty over the last block (e.g. “the difficulty was 70% of the maximal difficulty level”) and the amount of points they earned in this block (e.g. “you earned 7 points in this block”) which directly allowed them to infer the number of correct decisions (e.g. in this case 10 correct decisions).

In group 2, participants received rewards based on a combination of their task accuracy and confidence alignment. They received 1 point when correct and highly confident, but lost 1 point when highly confident and wrong. When they rated low confidence they neither won or lost points. Therefore, to maximize points participants should try to be most often correct and highly confident, but avoid misaligned high confidence to not lose points. Every 10 trials, they were presented with a 2 × 2 table showing the number of correct and incorrect decisions split according to high and low confidence.
7.2.6 Behavioural measures

7.2.6.1 Measurement of metacognitive ability

I assessed participants’ metacognitive ability based on the behavioural task during the initial and final assessment. I calculated \( meta - d' \) based on participants’ confidence ratings in their initial decisions (before post-decision evidence was presented), which measures their ability to discriminate between incorrect and correct trials. To directly control for perceptual performance, I calculated the ratio \( meta - d'/d' \) by employing a Bayesian estimation scheme (Fleming, 2017), using the non-hierarchical version of the model.

7.2.6.2 Measurement of behavioural post-decision evidence integration

To assess post-decision evidence integration, I calculated the change in accuracy (\( d' \)) between the initial and final decision. An improvement in the accuracy from initial to final decision is a behavioural indicator for the integration of post-decision evidence.

7.2.7 Belief updating: climate change

In addition to analysing behavioural measures from the perceptual decision-making task, I sought to investigate a potential transfer of training effects to attitudes towards real-world beliefs. Specifically, I was interested in participants attitude and belief updating regarding climate change. First, I assessed participants’ attitudes regarding the severity of climate change (Konisky et al., 2016), before answering 8 climate change knowledge question and giving a confidence rating after each question (Fischer et al., 2019). After answering all questions, participants were presented with the actual correct answers and finally indicated their attitude about climate change severity again. If participants incorporated this new information, they should change their attitude. I measured this shift in attitudes regarding the severity of climate change. Since the data was highly right-skewed (i.e. most participants showed small changes in attitude whereas large changes were rare), I used the logarithm of this belief updating measure to derive a more normally distributed variable. Since participants should only update their attitudes if they received new information (i.e. they answered some of the knowledge questions wrong) I controlled for climate change knowledge when investigating a link between metacognitive training effects and climate change related belief updating.

I also collected self-report measures of dogmatism and a confirmation bias inventory (Rassin, 2008; Altemeyer, 2002), but found no relation to the effects of cognitive training, indicating a potential dissociation between self-report and behavioural measures of confirmation bias.
7.2.8 EEG pre-processing

EEG was recorded continuously at 2000 samples/second using a 64-channel using a Brain products active-electrode system (Brain Products GmbH). Data was segmented into 8200 ms segments from 200 ms before to 8000 ms after the trial onset, where each segment encompassed one trial. Each epoch was aligned to the onset of the trial or, for analysis of the post-decisional phase, was realigned to the onset of post-decision evidence (to minimize any presentation delays that may have occurred during the trial). The data were resampled to 200 Hz to conserve processing time and improve signal to noise ratio, resulting in data samples spaced every 5 ms. All data were then high-pass filtered at 0.5 Hz to remove slow drift. All analyses were performed directly on the filtered, cleaned EEG signal.

7.2.9 Neural measure of post-decision evidence processing

The analysis method to determine neural post-decision evidence processing followed the same procedure as described in chapter 3, please see section 3.2.7 for more details.

Briefly, I built a machine-learning classification algorithm to predict participants’ decisions on each trial (leftward vs. rightward motion) at each timepoint during the post-decision phase and used the probabilistic prediction of this classifier as a measure of the (decodable) neural evidence in favour for left versus right motion. After applying the classifier to every trial and time point during the post-decision phase, I obtained a timeseries of neural evidence accumulation within each trial. The accumulation process can be summarized by fitting a linear regression to the time series on each trial, where the slope is analogous to the drift rate in a drift-diffusion model, and the intercept analogous to the starting point. The slopes extracted from this analysis are signed, such that positive values indicate stronger prediction of a rightward choice and negative values stronger evidence for a leftward choice. In order to obtain an unsigned metric of evidence accumulation strength, I flipped the sign of the slopes extracted from trials in which leftward motion was presented. This measure quantifies a propensity to correctly integrate the presented information, by tracking an internal DV, where higher values indicate stronger sensitivity to the presented stimulus.

7.2.10 Belief updating: General knowledge - Sinclair et al. 2019

In order to test the robustness of the results and to confirm a link between metacognitive ability and belief updating in an independent, larger sample, I re-analysed a publicly available data set from [Sinclair et al. (2019)](Sinclair et al. (2019)) (for a more detailed description of the task please see their publication). This study was conducted online, recruiting
participants from Amazon Mechanical Turk. The data set includes a total of 278 participants from two studies. I excluded another 39 participants from the current analysis as they had fewer than 70 trials recorded (out of 120 trials) indicating a potential lack of focus and making it difficult to reliably estimate their metacognitive ability \( (\text{meta} - d') \). Thus, I analysed a final sample of 239 participants.

The study was split into three parts. First, participants conducted a training session in which they answered 120 general knowledge questions and rated their confidence in these answers. After each answer (and confidence rating) they directly received feedback about the correctness of their choice, a procedure showing strong similarity with the current metacognitive training approach. After this, participants completed an immediate test session. Here, they were tasked to answer half of the questions (60 questions) again and to rate their confidence. The other half of questions (60 questions) were asked 2 weeks later in a delayed test session. I derived a measure of belief updating as suggested by Sinclair et al. (2019), by calculating the percentage of initially wrong answers (in the training session) that were subsequently revised (in the test session).

Since each answer was accompanied by a confidence rating, it was possible to calculate \( \text{meta} - d' \) and \( d' \) for each participant. Since the performance in this task was not staircased, \( d' \) could vary widely, reaching values close to zero (or even going negative). Therefore, I used \( \text{meta} - d' \) instead of the ratio \( \text{meta} - d'/d' \) as an indicator for metacognitive ability, since using the ratio resulted in extreme values for some participants. To ensure first-order performance was still controlled for I entered \( d' \) (derived from the training session) as a covariate in the linear regression models.

I calculated \( \text{meta} - d' \) for the training session which represents a mix of baseline metacognitive ability and improvements in metacognition due to the feedback. To disentangle baseline metacognition from training effects, I also calculated \( \text{meta} - d' \) for the test session as purer measure of a training effect. However, incorporating feedback for a specific question should influence one’s subsequent answer and confidence rating. Therefore, calculating \( \text{meta} - d' \) and belief updating based on the same questions and then relating both measures to each other might lead to circular inference. Therefore, I calculated \( \text{meta} - d' \) based on the questions in the immediate test session, whereas belief updating was calculated based on the questions in the delayed follow-up session (2 weeks later) to ensure two independent measures.

### 7.2.11 Statistical analysis

All statistical analyses were conducted using Matlab R2017b, applying two-tailed tests. In all regression analyses I employed robust fits to reduce influences of outliers, by using the default robust option of the MATLAB function fitlm which applies a “bisquare” weighting.
7.2.11.1 Behavioral analysis

I compared behavioural variables from the assessment before the training with behaviour from the session after the training to evaluate the training. I used repeated measure ANOVAs, whereby the training session (before training = 1, after training = 2) was entered as a within-subject factor and the group (performance feedback = 1, performance and confidence feedback = 2) as a between-subject factor. To ensure that training effects were not driven by nonspecific changes in performance or stimulus characteristics I entered these potentially confounding variables as covariates.

As metacognitive ability is tightly linked to first-order performance (Maniscalco & Lau 2012), I used the ratio $\text{meta-d'}/d'$ to directly control for potential differences in performance (Fleming & Lau 2014). Stimulus characteristics such as the stimulus strength and variability have also been reported to be related to metacognitive ability (Rahnev & Fleming 2019) and therefore I also controlled for these covariates when investigating training effects on metacognition. When investigating training effects on post-decision evidence integration I also controlled for these same covariates (performance in initial decision, stimulus strength, stimulus variability).

To investigate a specific relationship between the training effects on metacognition and the training effects on post-decision evidence, I calculated an improvement score (metacognitive efficiency after the training - metacognitive efficiency before the training; post-decision evidence integration after the training - post-decision evidence integration before the training) and correlated these two training effects. I again controlled for similar changes in performance in the initial decision, stimulus strength and stimulus variability.

To formally test whether improvements in post-decision evidence integration were driven by training effects on metacognition, I conducted a mediation analysis, using the Multilevel Mediation and Moderation (M3) Toolbox (Wager et al. 2008). Mediation analysis assesses whether covariance between two variables (predictor and dependent variable) is explained by a third mediator variable. Significant mediation is obtained when inclusion of the mediator significantly alters the slope of the relationship of predictor and dependent variable (evaluated as the product of the predictor-mediator and mediator-dependent variable path coefficients). Training session (i.e. before training=0, after training=1) was entered as the predictor variable, changes in post-decision evidence processing was entered as dependent variable and changes of metacognitive efficiency as the mediator. I controlled for covariates that may could have a confounding influence on these linkages such as changes in initial decision performance, changes in stimulus strength and changes in stimulus variability. The following effects of interest were simultaneously tested: the impact of training on metacognitive efficiency (path a); the impact of metacognitive efficiency on post-decision evidence integration (path b); and the formal mediation of metacognitive efficiency on post-decision evidence integration (path a × b). The
direct effect of training on post-decision integration before and after controlling for metacognitive efficiency was also estimated (paths c and c’, respectively). Parameter estimates for each path (a, b, c, a × b, c’) were obtained by bootstrapping 10,000 times with replacement, producing two-tailed p-values and 95% confidence intervals.

### 7.2.11.2 Neural evidence integration

The EEG machine-learning analysis yields separate trial-by-trial measures of the starting point (intercept) as well as rate (slope) of post-decision evidence integration. Since my hypothesis was about the rate of evidence integration, I focused on the changes in the slope of neural evidence integration due to training. However, since the intercept and slope of each trial were highly anti-correlated (r=-.59; i.e. when a trial starts with a strong neural representation of the stimulus, there is only limited room for changes in this neural representation) I statistically controlled for the starting-point in all analysis to derive pure metrics of evidence accumulation.

Since I had a trial-by-trial measure of post-decision evidence integration for each participant and each session, I could use a mixed-effect linear model to analyse the data, with random effects for participants. Based on my previous findings (Rollwage et al., 2020) I expected that the slope of post-decision evidence integration would be influenced by the initial decision (confirmatory post-decision evidence =1; disconfirmatory post-decision evidence=-1), initial confidence (low confidence= -1; high confidence =1) and their interaction. Since I was interested in the effects of training and potential group differences on post-decision evidence integration, I also entered training (before training=-1, after training=1) and group (group 1= -1, group 2=1) as predictors. Therefore, training, group, initial decision, initial confidence and the interactions between these variables were entered as predictors for post-decision evidence integration. To ensure that any observed training effects in neural evidence processing were not driven by changes in decodability or stimulus characteristics, I entered the overall decodability of each session and the trial-by-trial stimulus coherence (this differed for every trial as difficulty was constantly staircased) as covariates. Significance was inferred based on group-level fixed effects of each predictor. My hypothesis focussed on training effects on general neural evidence integration, which should manifest as a main effect of training.

Additionally, I tested whether individual differences in the training effect on neural evidence processing were related to training effects on metacognition. To this end, I derived individual measures of the change in neural evidence integration due to training for each participant, by allowing the training effect to vary as random effect for each participant.
7.2.11.3 Belief updating: Sinclair et al., 2019

To test for a general link between metacognition and belief updating, I entered \( \text{meta} - \hat{d}' \) derived from the training session to predict belief updating measured in the test sessions. Additionally, I simultaneously entered metacognitive ability during the training session and the improvement of metacognitive ability (\( \text{meta} - \hat{d}' \) in immediate test session – \( \text{meta} - \hat{d}' \) in training session) as predictors for belief updating (derived from the follow up test session that was conducted with a delay of two weeks). To ensure isolation of specific effects of metacognition from general cognitive abilities, I controlled for general knowledge performance (\( \hat{d}' \) during the training session) in these regression models.

7.3 Results

7.3.1 Behavioral training effects

In total, 40 participants (N=20 in each group respectively) completed all training and assessment sessions. As expected, I found an increase in participants’ metacognitive efficiency (main effect of training: \( F(1,38)=5.45, p=.025, \text{Cohen’s } d=.38 \), see Figure 7.2A). However, there were no group differences (interaction effect of group \( \times \) training: \( F(1,38)=0.24, p=.63 \)), indicating that both types of performance feedback were sufficient to boost participants’ metacognition (since both groups showed a similar increase in metacognitive efficiency, I collapsed across both groups for subsequent analysis). Importantly, this improvement did not reflect nonspecific practice effects as there was no change in participants’ performance on the initial decision (main effect of training: \( F(1,38)=0.11, p=.74, \text{Cohen’s } d=.05 \), see Figure 7.2B) and the improvement in metacognitive efficiency was not explained by changes in stimulus strength or stimulus variability (Rahnev & Fleming, 2019) (p = .002 when controlling for these effects). Instead, incremental learning over training sessions was observed leading to general improvements in metacognitive ability (see grey markers in Figure 7.2A).

Having established that training increased participants’ metacognitive efficiency, I next tested whether this procedure also improved participants’ processing of post-decision evidence (i.e. reduced confirmation bias). Post-decision evidence integration was measured behaviourally as the improvement in performance from the initial to the final decision. There was a clear training effect, with an improvement in post-decision evidence processing (main effect of training: \( F(1,38)=11.58, p=.0016, \text{Cohen’s } d=.54 \) seen in both groups (interaction effect of group \( \times \) training: \( F(1,38)=0.2, p=.66 \)). The magnitude of this effect was striking, with a 53% increase of post-decisional integration after the training (see Figure 7.2C). This effect was not explained by nonspecific changes such as alterations in initial decision
Chapter 7. Boosting metacognition to reduce confirmation bias

Figure 7.2: Behavioural training effects (n=40 participants). A The training improved participants’ metacognitive efficiency (meta – d'/d') from before to after the training. The grey markers and lines represent the metacognitive efficiency for each training session, showing an incremental improvement over sessions. B The performance (d') of the initial decision was staircased in each session. As expected, initial decision performance was stable before and after the training. C The training improved post-decision evidence processing. D This effect was driven by an increased revision of initial mistakes, but no change in revisions of initially correct decisions. Percentage of revised initial decisions is presented as a function of whether the initial decision was correct (upper panel) or incorrect (lower panel). A-D Data are presented as mean values +/- SEM, combined for both training groups. E Individual differences in the improvement of metacognitive efficiency (meta–d'/d') are correlated with individual variation in the change in post-decision evidence integration. F The training effects on post-decision evidence integration were fully mediated by changes in metacognitive efficiency. *p<.05, **p<.01, ***p<.001
performance or stimulus strength (p=.0009 when controlling for these effects). Importantly, this behavioural improvement in post-decision evidence processing after training was driven by participants revising their initial mistakes more often (main effect of training: F(1,38)=7.15, p=.011, Cohen’s d=.43, see Figure 7.2D lower panel), while revisions of initially correct answers remained unaltered (main effect of training: F(1,38)=0.01, p=.97, Cohen’s d=.02, see Figure 7.2D upper panel). This suggests that the procedure indeed reduced participants’ confirmation bias, making them more likely to correct their originally incorrect decisions when presented with new information.

As predicted, the training boosted both metacognitive ability and reduced confirmation bias – a pattern of changes I interpret as being driven by a facilitatory effect of metacognition on the processing of post-decision evidence. To evaluate this proposed mechanism, I conducted an additional analysis of individual differences. In line with my hypothesis, participants who showed a stronger training effect on their metacognitive ability also showed a more pronounced boost in post-decision evidence processing ($\beta=.51$, t(38)=3.64, p=.0008, $R^2=.26$, see Figure 7.2E). This effect was not explained by other non-specific changes in performance or stimulus characteristics (p=.004, when controlling for these effects). I formally tested whether metacognitive improvements were the crucial factor for increasing post-decision processing, by conducting a mediation analysis. The training effect on post-decision evidence processing was fully mediated by changes in metacognitive ability (indirect effect $a \times b$: $\beta=.05$, p=.002, see Figure 7.2F), indicating that a person who shows limited improvement in metacognition is also unlikely to show an improvement in post-decision evidence integration (direct effect $c$: $\beta=.15$, p=.001; corrected direct effect $c'$: $\beta=.09$, p=.35).

### 7.3.2 Neural training effects

In addition to behavioural variables, I also recorded EEG before and after the training. This enabled me to evaluate the degree to which participants processed post-decision evidence on a neuronal level by applying machine-learning classification algorithms to decode participants’ decisions (the same analysis conducted in chapter 3). Changes in this prediction due to the presentation of new evidence can be seen as a neural indicator of post-decision evidence processing. A positively increasing classifier prediction indicates that participants correctly integrate the new post-decision evidence over time. Importantly, my previous experiments indicate that an improvement in metacognitive ability should increase the processing of both disconfirming and confirming evidence (see Figure 7.1A lower panel).

First, I tested whether there was a selective neural integration of post-decision evidence, as found in chapter 3. Replicating the findings (now using EEG instead of MEG), I found an enhanced processing of confirming compared to disconfirming
Chapter 7. Boosting metacognition to reduce confirmation bias

Figure 7.3: Training effects on neural post-decision evidence processing. **A&B** Neural metrics of post-decision accumulation **A)** before and **B)** after the training. **A-C** Neural evidence processing is separated by confirming (consistent with initial decision) and disconfirming (inconsistent with initial decision) post-decision evidence. More positive values on the y-axis indicate better (more veridical) representation of the actual motion direction, such that positive slopes indicate stronger integration of post-decision evidence. Weighted group averages (grand average) are presented and regression lines are fits to this averaged data. **C** Time point by time point difference in neural processing due to the training. **D Overall change in neural processing (averaged over confirming and disconfirming information) due to training. Data are presented as mean values +/- SEM, the grey dots represent individual participants’ data. **E Individual differences in the improvement of metacognitive efficiency ($meta - d'/d'$) are correlated with individual differences in increases of neural post-decision evidence integration. *p<.05
post-decision evidence ($\beta=.03$, $t(26196)=5.34$, $p=9.42\times10^{-8}$) an effect that was not altered by training ($\beta=.003$, $t(26196)=.47$, $p=.64$).

As discussed beforehand, I expected to see a training effect in the form of enhanced processing of disconfirming and confirming information. Therefore, the relative processing of these two types of information should remain unchanged. In line with this prediction, the overall processing of post-decision evidence was increased after training ($\beta=.012$, $t(26196)=2.17$, $p=.03$; see Figure 7.3A-D). Importantly, these effects were not driven by nonspecific differences between the two sessions, such as stimulus characteristics or general classification accuracy, as the reported analysis controlled for these potentially confounding effects. Taken together, these findings suggest that post-decision evidence can be more accurately decoded after training, complementing and supporting the behavioural indices of post-decision processing.

To further validate that these changes in neural processing were driven by improved metacognition, I again leveraged individual differences in the training effects. I found that participants who improved their metacognitive efficiency more due to the training also showed a stronger increase in post-decisional neural processing after the training ($\beta=.36$, $t(38)=2.21$, $p=.033$, $R^2=.11$, see Figure 7.3D). These results further support that a boost in metacognitive ability alters the way participants’ brains process information, making people more open to the integration of new information.

7.3.3 Generalization of training effects to real-world beliefs

While these findings indicate that training improved both metacognition and post-decision processing, these results are obtained in the context of simple perceptual decisions. In contrast, confirmation bias has received most attention for its potential influence on real-world beliefs about societal issues (Whitmarsh 2011; Nyhan & Reifler 2010; Kaplan et al. 2016; Sunstein et al. 2016). To evaluate whether a low-level training approach in the context of a perceptual decision-making task would also generalise to reduce confirmation bias in such settings, I also collected attitudes and belief updating regarding real-world issues after training was completed.

Specifically, I was interested in beliefs and attitudes about climate change, as metacognitive abilities regarding climate change knowledge have been reported to crucially influence attitudes on this issue (Fischer et al. 2019; Fischer & Said 2020). First, I assessed each individual’s attitudes regarding the severity of climate change. Then participants had to answer eight climate change related factual knowledge questions and were asked to rate their confidence in these answers (same procedure as in Fischer et al. 2019), before being presented with the correct answer, i.e. participants received new factual information. Finally, I assessed each individual’s opinion about the severity of climate change again. While most participants changed
Chapter 7. Boosting metacognition to reduce confirmation bias

their view about climate change based on these facts (23 out of 40 participants changed their view), there was individual variation in the degree of belief updating (see Figure 7.4A). Interestingly, variation in climate change belief updating was related to individual differences in metacognitive training effects (as measured in the perceptual task). Participants for whom the training improved metacognitive ability were also more likely to update their climate change beliefs more presented with new factual information ($\beta = .46$, $t(38) = 2.84$, $p = .007$, $R^2 = .18$, see Figure 7.4A). This effect was driven by participants with enhanced metacognitive efficiency (after training) updating their views more ($\beta = .36$, $t(35) = 2.09$, $p = .04$, $R^2 = .09$), and being more sensitive to the feedback information when changing their views about climate change ($\beta = .41$, $t(34) = 2.52$, $p = .017$, $R^2 = .03$). This could indicate that the effects of metacognitive training might generalise to affect the updating of higher-level attitudes and make people generally more receptive for new information.

I note, that due to the time-consuming nature of this study, the sample size was relatively small. To validate the robustness of my results, I aimed to confirm a similar link between metacognitive ability and belief updating in an independent, larger sample. To this end, I re-analysed a publicly available data set from Sinclair et al. (2019). This data set includes a total of 239 participants that were analysable for the present purposes. In this study, participants answered 120 factual knowledge questions and rated their confidence. After each answer, participants received feedback about the correctness of their answer (following a similar protocol as in my cognitive training procedure). After this training session, participants were asked the same knowledge questions again (half of them were asked immediately after the training, the other half were asked 2 weeks later) and their belief updating was measured as the proportion of questions on which participants revised their initially incorrect answers.

I was interested whether metacognitive ability during the training session predicted the degree to which participants would update their beliefs in the testing sessions. In line with this prediction, participants with better metacognitive ability ($\text{meta} - d'$) in the training session revised their beliefs more often based on the feedback information ($\beta = .44$, $t(237) = 7.81$, $p = 1.8 \times 10^{-13}$, $R^2 = .21$, see Figure 7.4B). However, since Sinclair et al., did not assess participants metacognitive ability before the training session it is difficult to disentangle baseline metacognitive ability from effects of training. Nevertheless, participants rated their confidence also during the test sessions. Hence, it is possible to evaluate their metacognitive ability both during and after training, and a change in metacognitive ability from training to test can be seen as pure indicator of a training effect. When entering both metacognitive ability during the training session as well as the change from training to testing (on the same day) to predict belief updating (two weeks later), both predictor explained unique variance in the degree of belief updating (baseline metacognition: $\beta = .36$, 
Chapter 7. Boosting metacognition to reduce confirmation bias

133

t(237)=7.76, p= 2.6*10^{-13}, R^2=.21; change in metacognition: \( \beta = .53, t(237)=11.65, p= 10^{-20}, R^2=.29; \) combined \( R^2=.5). \) Importantly, these effects of metacognition on belief updating were not explained by general cognitive abilities, as the observed effects remained significant when controlling for participants’ knowledge in the training session (p-values<.0001 when controlling for this effect).

**Figure 7.4:** Association of metacognitive ability and belief updating about real-world issues A&B The upper panels exemplify the procedure to measure belief updating. A Training effects on metacognitive efficiency (as measured in the perceptual task) predicted individual differences in belief updating regarding climate change. Participants gave their opinion regarding the severity of climate change and were presented with new factual information before their climate change attitude was assessed again. Belief updating is measured as the absolute change in attitude regarding the severity climate change due to the additional information. Since the belief updating distribution was highly right skewed (small changes were common, while large belief updates were rare), the logarithm of belief updates was used. N=40 participants. B I re-analysed a publicly available data set from Sinclair et al., 2019 to assess participants belief updating for general knowledge questions. Participants answered a series of questions and gave a confidence rating before receiving feedback about the correctness of their answer. Afterwards, they were asked to answer the same questions again, whereby belief updating is measured as the percentage of initially wrong answers that were revised. Since participants also gave confidence ratings for each knowledge question, it is possible to evaluate their metacognitive ability for these questions. Participants with higher metacognitive ability (\( meta – d’ \)) updated their beliefs more due to feedback. N= 239 participants. **p<.01, ***p<.001

7.4 Discussion

In summary, these results suggest that cognitive training procedures that deliver feedback on task accuracy and confidence alignment are effective ways for improving participants metacognitive ability. In turn, such boost in metacognitive ability
seems to enable participants to be more open to new information and revise their beliefs when needed. I further showed that an increased neural processing of new information may underpin such reductions in confirmation bias. Finally, these results seem to generalize to belief updating about real-world attitudes such as climate change and general knowledge, indicating the potential of metacognitive training procedures to promote openness to new information about a variety of issues.

Given previous literature (Carpenter et al., 2019), it was surprising to see improvements of metacognitive efficiency in the performance feedback group, as participants did not receive explicit feedback about their confidence alignment. Nonetheless, even in this group, participants received detailed performance feedback with relatively high frequency (every 10 trials) that allowed a direct inference on the number of correct and incorrect decisions. This information may have been sufficient for participants to indirectly infer their confidence alignment and thus use the feedback to improve their metacognitive efficiency. In comparison, in previous studies, participants received performance related feedback that did not allow a direct inference about the number of correct decisions and was less frequent (every 27 trials) (Carpenter et al., 2019). Further research investigating which feedback components are sufficient for metacognitive improvements will be valuable in distilling the active ingredient of metacognitive training and enable the identification of efficient procedures for boosting metacognition.

To my knowledge, the current study is one of the first interventions that may effectively counteract a confirmation bias. One reason for why it has proven difficult to develop interventions is the relatively limited mechanistic understanding of the factors underlying confirmation bias (Lilienfeld et al., 2009; Klayman, 1995). My previous findings revealed a key role of confidence in driving a confirmation bias, and proved crucial in generating a hypothesis-driven intervention to target a core mechanism underlying confirmation bias. Identifying and targeting an underlying mechanism in this way is rare in the literature of debiasing interventions (Lilienfeld et al., 2009; Klayman, 1995). This deep understanding was crucial in designing the most effective intervention. More broadly, by applying behavioural tasks and computational techniques to identify the drivers of other cognitive biases, similar targeted interventions may become feasible.

While the results seem to paint a coherent picture that improvements in metacognition drive increased post-decision evidence processing, the lack of a control group makes strong causal claims problematic. For instance, it could be that some participants were more engaged with the study and thus showed improvement on the different variables, explaining the correlation between the different outcome measures (e.g. metacognitive ability and climate change belief updating). Importantly, I assessed a multitude of additional tasks variables (e.g. performance in the initial decision and task difficulty) which showed no change due to the training and did not
Chapter 7. Boosting metacognition to reduce confirmation bias

explain the link between improvements in metacognition and post-decision evidence processing. Nevertheless, it is not possible to completely rule out some people being in general more likely to change their behaviour in response to training and thus driving the correlation of improvements between participants. To derive stronger claims regarding the causality of these effects, an additional control group would be need. Such a control group should show no changes in metacognitive abilities as a function of training. In order to achieve this, future studies could contrast a condition in which participants undergo the same training as here (including the feedback) compared to a group that undergoes training without any feedback, with the expectation that only the feedback group should show improvements in metacognition and post-decision processing.

An alternative approach to “debiasing” decision-making has been to educate people about the existence of cognitive biases with the implicit assumption that this knowledge is enough to successfully resist these mental shortcomings (Miller, 2016). However, since many cognitive processes are not available for introspection, such education strategies might not be sufficient (Oswald & Grosjean, 2004; Tavris & Aronson, 2008; Croskerry et al., 2013; Pronin et al., 2004). The current approach differed in this respect, as it relied on repeated feedback over hundreds of judgments. This repeated training approach might be important as the brain is known to learn incrementally over extended periods of time (see Figure 7.2A), for instance by strengthening and forming new synaptic connections (Malenka & Nicoll, 1999; Kandel, 2007; Liu et al., 2010; Petrov et al., 2005), and similar time investments may be required to shape higher-level parameters governing decision-making.

Alternative training approaches require deliberate contemplation and awareness of the potential influence of cognitive biases for each decision (Lord et al., 1984; Hirt & Markman, 1995; Sellier et al., 2019; van Brussel et al., 2020), a process that requires mental resources and time (Stanovich & West, 2000; Kahneman, 2003). Unfortunately, cognitive biases have been shown to be strongest when time and mental capacities are limited (Evans & Curtis-Holmes, 2005; Gilbert et al., 1993), potentially explaining why these strategies are prone to fail in the situations where they are most needed. In comparison, the training used here did not require participants to explicitly focus on counteracting cognitive biases. In fact, I never mentioned cognitive biases to the participants in order to avoid experimental demand characteristics. In other studies, participants were explicitly instructed about these biases, making robust and generalizable effect outside the experimental setting less likely (Sellier et al., 2019; van Brussel et al., 2020). Nonetheless, it would be interesting to test whether additional information about confirmation bias combined with a metacognitive training could yield even more powerful results.

The current study revealed training effects on the neural sensitivity for the processing of new information. Interestingly, the training did not alter the relative
processing of confirming and disconfirming post-decision evidence, but enhanced
the processing of both evidence types. While post-decision evidence is crucial for re-
vising initial mistakes, it is also important to not to become distracted by additional
information if a decision is correct. Therefore, the processing of confirming infor-
mation is also a vital part of adaptive information usage and the training improved
both forms of evidence integration.

Strikingly, the current training utilised simple perceptual decision-making tasks
while still generalizing to real-world beliefs such as climate change. This points
towards a domain-general improvement in the processing of new information. Con-
firmation bias has been assumed to contribute to societal polarization and misin-
formed beliefs in a wide range of issues (Lilienfeld et al., 2009; Rollwage et al., 2019).
In light of this, my results show promise in highlighting a simple cognitive training
procedure that might be effective in promoting the openness for new information
and ameliorate some of the negative societal effects of confirmation bias.
Chapter 8

General discussion

8.1 Overview

This thesis investigated the cognitive, computational and neural mechanisms underlying altered post-decision evidence integration. Specifically, I probed the role confidence plays for the processing of new information. I found that confidence induces a selective gain for choice-consistent information, i.e. causing a confirmation bias (chapter 3). This contradicts predictions from Bayesian belief updating which would have suggested a symmetric influence of confidence on confirming and disconfirming information (Meyniel et al. 2015a; Meyniel & Dehaene 2017).

More broadly, these findings shed light on the mechanisms that drive confirmation bias. These insights strongly suggest that metacognitive ability might be a crucial determinant for the negative consequences of selective post-decision processing. Theoretical modelling confirmed this intuition (chapter 6), demonstrating that low metacognitive ability leads to detrimental effects of a confidence-induced confirmation bias, while high metacognitive ability can render this strategy even superior compared to an unbiased observer. Guided by this hypothesis, I then tested whether a metacognitive training procedure can increase participants’ sensitivity for new information (chapter 7). Indeed, a boost in metacognitive ability was observed which in turn increased participants’ openness for post-decision evidence.

While these findings paint a coherent picture, all findings were obtained in simple perceptual tasks. However, biased information processing has received most interest with respect to real-world issues such as political polarization (Lilienfeld et al. 2009). Therefore, I also explicitly tested for a link between alterations in information processing and dogmatic (political) beliefs. I found that people who hold dogmatic and rigid attitudes were characterized by reduced metacognitive ability, abolished processing of disconfirming post-decision evidence (i.e. increased confirmation bias) and a lower willingness to seek out additional information in order to refine their decisions (chapter 4 and 5).

In the following chapter I will discuss these finding in more detail.
8.2 Confidence induces biased post-decision information processing instead of Bayes optimal weighting

I investigated the influence of confidence on the processing of new information. My findings indicate a selective boost in the processing of confirming post-decision evidence while disconfirming evidence processing is abolished. This is in contrast to Bayesian accounts in which confidence should symmetrically scale the weight of new incoming information (Atiya et al., 2019; Meyniel et al., 2015a; Meyniel & Dehaene, 2017; Meyniel, 2020). Using drift-diffusion modelling enabled me to disentangle an influence of the initial decision and confidence on the starting point and the drift-rate of a post-decision accumulation process. The change in post-decisional drift-rate was a clear indication of altered processing of new information, leading to a confirmation bias. When comparing my results to previous studies that showed a symmetric influence of confidence on belief-updating, one crucial difference is that I investigated the influence of participants’ subjective confidence (i.e. their actual confidence ratings), while Meyniel and colleagues investigated the influence of Bayes optimal confidence (i.e. the confidence a Bayesian observer should have based on the received information) on information processing (Meyniel et al., 2015a; Meyniel & Dehaene, 2017; Meyniel, 2020). Their use of this theoretical quantity, derived from Bayes theorem (rather than the actual confidence ratings of participants), might have pushed their results towards favoring a Bayesian account. This difference might partly explain the observed discrepancies.

My finding that the initial decision and confidence directly modulated the neural processing of new information shows a strong resemblance to the observation that prior expectations can influence the processing of new information (Braun et al., 2018), whereby the neural processing of expected stimuli is enhanced (Aitken et al., 2020; Kok et al., 2012, 2013). In my task, the initial decision represents a form of prior belief with respect to the post-decision evidence, whereby the initial confidence might indicate how strong this prior belief is. This suggests that the influence of prior knowledge on posterior beliefs might be implemented through a modulation of neural post-decision processing. This has mainly been interpreted as support for Bayesian belief updating (Meyniel & Dehaene, 2017; Meyniel, 2020), however a direct modulation of new information (i.e. the likelihood) by a prior belief might also indicate potentially maladaptive circular inference (Bouttier et al., 2019; Jardri et al., 2017). Therefore, more work is needed in order to disentangle whether this modulation of new evidence by prior information is Bayes optimal or causes a confirmation bias.

While my data points towards the latter possibility, it is interesting to consider that the accuracy of confidence judgments (i.e. metacognitive ability) is a crucial determinant for the positive or negative consequences of such information process-
ing strategies (see chapter 6 and 7). Excitingly, my simulations showed that a 
confidence induced confirmation bias can outperform a perfectly Bayesian strategy 
when metacognitive ability is high. These findings might reconcile this apparent dis-
crepancy, showing that such information processing can remain optimal even when 
deviating from Bayesian principles.

8.3 Mechanisms underlying confirmation bias

Besides revealing a role of confidence in post-decision evidence processing, my re-
results also expose a mechanisms that drives confirmation bias. While many different 
cognitive mechanisms might contribute to a reluctance to change one’s mind in light 
of new information, few studies were able to pin down the exact cognitive and neural 
mechanisms leading to this bias.

I show that the integration of post-decision evidence is altered after decision-
commitment. This fits well with findings from other cognitive neuroscience studies. 
These studies consistently identified a selective incorporation of post-decision evi-
dence as underlying mechanisms of confirmation bias (Talluri et al., 2018; Palminteri
et al., 2017; Cheadle et al., 2014; Talluri et al., 2020).

Moreover, the neuroimaging data shows that this selective incorporation of ev-
dence is apparent in brain activity after a few hundred milliseconds. In line with 
findings from Kappes et al. (2020), this further supports that contradictory evi-
dence may not even be processed by the brain in some situations. This goes against 
other accounts of confirmation bias which assume that disconfirming and confirm-
ing evidence are processed similarly, but disconfirming evidence is later discarded 
when forming a decision (Doll et al., 2011; Gilbert et al., 1993; Westen et al., 2006).

Importantly, this insight has direct implications for potential alleviation strategies. 
In particular, the enhancement of selective processing by confidence indicates that 
holding high confidence in a wrong belief will lead participants to not even to pro-
cess disconfirming information. Therefore, this situation needs to be avoided all 
together in order to bypass the negative consequences of such bias, rather than try-
ing to implement strategies that would tackle the weighting of information later on 
in the processing stream (Mellers et al., 2014). The use of metacognitive training in 
chapter 7 helped participants to better align their confidence in order to avoid high 
confidence mistakes and was in turn effective in promoting post-decision evidence 
processing.

In light of this, I hope my thesis exemplifies the power of mechanistic understand-
and the merits of a computational assessment of the processes that underly 
cognitive biases. Only a precise understanding of the contributing factors enabled a 
theory-driven intervention.

While I sought to eliminate motivational factors from my tasks, in everyday life
motivational and cognitive mechanisms are likely to interact to influence whether or not we change our mind. After having shown these motivational and cognitive contributions to post-decision information processing in isolation (Gesiarz et al., 2019; Leong et al., 2019; Talluri et al., 2018; Rollwage et al., 2020a), it would be an exciting avenue to test how these factors interact. For instance, the influence of confidence on evidence processing might be enhanced when accompanied by motivational and social factors. Investigating the impact of multiple factors in conjunction might bring us closer to a full understanding of information processing and cognitive biases in real world situations.

8.4 Altered information processing and dogmatic beliefs

Similar to the study of cognitive biases, political psychology has mainly used methods from social sciences (Greenberg & Jonas, 2003; Jost et al., 2003; Jost, 2017). One main aim in this field is to identify “cognitive styles” – content-free styles of thinking – that are linked to specific political ideologies (see Jost (2017) for a comprehensive review). An initial wave of findings has enabled researchers to sketch out a conceptual landscape that maps cognition onto politics (Jost, 2017), whereby recent efforts have mainly focussed on the cognitive styles associated with extreme and radical political beliefs (van Prooijen & Krouwel, 2019; Zmigrod, 2020). However, in a majority of studies the definition of cognitive styles remains qualitative in nature, operationalised by subjective self-reports from questionnaires, with considerable variability in their definition between studies (Jost, 2017). This renders it difficult to critically appraise and unify existing findings in order to identify cognitive processes supporting the development of specific beliefs.

In contrast, I utilized a new approach that involves the use of behavioural tasks in conjunction with formal computational models to uncover an algorithmic basis for cognitive styles associated with political radicalism. Especially well-validated behavioural tasks (informed by findings in cognitive neuroscience) can reveal differences in computational model parameters and enable discovery of candidate neural processes from which distinct cognitive styles may emerge.

The promise of such computational approach to political psychology is in identifying computational building blocks which lead to a mechanistic understanding of cognitive styles. Notably, however, such building blocks may themselves only explain limited variance in political attitudes. This is to be expected when using basic cognitive tasks with content that is unrelated to the political concepts of interest. Nevertheless, such mechanistic understanding is key for developing interventions to counteract political polarization, even when the association between computational alterations and real-world attitudes shows small effect sizes.
As an analogy, if a doctor attempts to predict whether a person will have a heart attack within 5 years, the best predictor might be the degree to which the coronary arteries contain plaque deposits, with large effect sizes. However, this knowledge does not tell us much about the mechanisms that create plaque deposits and will be unlikely to result in new treatments. Moreover, the contributors to plaque deposits are likely to be multifactorial (e.g. high levels of cholesterol, high blood pressure, etc.) and each of these factors may have only limited influence on the plaque deposit (such that effect sizes for links between individual mechanisms and deposits might be relatively small). Crucially, however, identifying small, reliable effect sizes associated with underlying mechanisms may bring us closer to the possibility of reducing heart diseases through targeted interventions such as a low-cholesterol diet and increased exercise.

Similarly, identifying computational alterations that underpin cognitive styles (even when these associations are small) which contribute to a radical mindset holds the promise of mechanistic understanding and thus the potential for targeted intervention.

In line with other recent findings in this field of computational political psychology (Sinclair et al., 2019; Zmigrod, 2020; Zmigrod et al., 2019a,b) the current findings suggest that altered usage of information might be related to rigid and dogmatic beliefs, at both ends of the political spectrum. This might explain why it is so difficult to convince people at the political extremes with new arguments or facts.

My results point towards a specific role of people’s confidence in this process. As shown in chapter 4, people holding dogmatic beliefs showed reduced metacognitive ability. Metacognition was in turn predictive of the incorporation of post-decision evidence, which was also impaired in more dogmatic participants. In chapter 5, I found that dogmatic participants were less likely to seek out additional information in order to refine their decisions. Again, this effect was (partly) driven by an altered use of internal confidence signals, whereby more dogmatic participants used their confidence estimates less adaptively to determine whether they should seek out more information or not. This suggests that metacognitive monitoring and metacognitive control are both reduced in participants who hold dogmatic views and in turn impairs their usage of additional information in order to correct initially wrong decisions.

The causality of this relationship is difficult to establish with correlational studies. It is possible that differences in metacognition predispose people to develop dogmatic views. However, it is also plausible that holding dogmatic views (and interacting in associated social circles) might reduce metacognitive abilities as in certain environments it might not be seen as beneficial to change one’s mind in light of new evidence (Van Bavel & Pereira, 2018). Independently of the direction of causality (as so often in psychology it might be an interaction between both), it is easy to imagine that such reluctance to use new information will increase en-
trenchment of opinions and sustain rigid beliefs. Since these entrenched beliefs are an increasing source of problems in our society, it might be desirable to counteract such rigidity. My findings suggest that metacognitive training could be potentially utilized in order to increase openness to new ideas and facilitate societal discourse.

8.5 Metacognitive training enhances the sensitivity for new information

Potentially the most important finding from this thesis is that a boost in metacognition increases participants’ sensitivity for post-decision processing (i.e. reduces confirmation bias). Throughout this thesis, I have argued that the merit of a mechanistic understanding is to ultimately use this knowledge in order to counteract negative consequence of the observed effects. Similar arguments have been used in the field of psychiatry, where “computational psychiatry” has been advocated for, based on the promise for mental health treatments (Huys et al. 2016, Montague et al. 2012, Wang & Krystal 2014). However, this translation has not always been straightforward (Berwian et al. 2020, Grosskurth et al. 2019). Therefore, although I had strong theoretical predictions regarding the effects and underlying mechanisms that metacognitive training should have on post-decision evidence processing, it is still notable that these predictions largely played out in empirical data. In this respect, my data supports the general notion that mechanistic understanding can have major benefits for the development of effective interventions.

Besides proving a theoretical point, my results have direct practical implications. Confirmation bias has been shown to influence a plethora of decision-making contexts, therefore it has been argued that an intervention to counteract this bias might be one of the important challenges of our times (Lilienfeld et al. 2009, Lorenz-Spreen et al. 2020). However, despite this desire for interventions, there have been few successful attempts (Lilienfeld et al. 2009). Besides a lack of mechanistic understanding, a main reasons for unsuccessful “debiasing” strategies might be the exclusive focus on educating people about the existence of cognitive biases (Miller 2016). While this might be a first step in the right direction, relying solely on these educational strategies might not be enough to overcome cognitive biases (Lilienfeld et al. 2009, Croskerry et al. 2013, Oswald & Grosjean 2004), especially since many of these biases are by definition unconscious (Pronin et al. 2004).

Even if people were aware of their biases, such educational strategies rely on people making a constant effort to detect and monitor own biases (Chaiken & Trope 1999, Kahemani 2011, 2003). Therefore, these strategies may only be helpful when people have the time and resources to actively counteract them. However, as shown empirically, cognitive biases have the strongest effect when people have no time or cognitive resources to perform this kind of analytical thinking (Evans & Curtis-
Holmes (2005) Gilbert et al. (1993). Therefore, educational strategies are most likely to fail in situations when they are most needed.

Although metacognitive abilities would also be classed as “analytic” thinking within a dual-process theory (Chaiken & Trope 1999; Kahneman 2011, 2003), internal feelings of confidence automatically accompany most (if not all) decisions. The current training approach aimed to improve the alignment of this internal feeling of confidence and through this to indirectly improve the processing of post-decision evidence. Participants did not have to explicitly remind themselves of the potential influences of confirmation bias. In fact, I never mentioned to the participants that I was interested in cognitive biases or the improvement of post-decision processing, in order to avoid demand characteristics. Therefore, it seems safe to assume that the training boosted post-decision processing without the need for participants to constantly be aware of potential biases.

I decided to not inform participants about the purpose of the study in order to make the results as interpretable as possible. If I had informed participants about the purpose of the study, they would have likely tried to behave according to the expectations within the experimental setting, making a robust and generalizable training effect less likely. Nevertheless, for practical applicability it will be important to maximize the effectiveness of the training and an educational component on cognitive biases could be helpful in this respect. It might be interesting to investigate whether additional information about confirmation bias in combination with a metacognitive training procedure might increase the observed effects on post-decision evidence processing.

One aspect that seems to be crucial for the effects of the metacognitive training is the continued performance feedback, over hundreds of trials. This component is shared with the few other interventions that have been shown to reduce confirmation bias (Sellier et al. 2019; van Brussel et al. 2020). These interventions also included some form of repeated feedback. This is in line with known principles of how the brain learns through incremental prediction errors (O’Doherty et al. 2004; Schultz et al. 1997), potentially driven by strengthening or formation of new synaptic connections (Kandel 2007; Malenka & Nicoll 1999; Liu et al. 2010). While this process might take extended periods of training, split over multiple sessions (as in my study), such changes in the neuronal processing might yield more robust and sustainable effects than other “debiasing” approaches.

With respect to the previous point, it will be very interesting for future research to investigate the sustainability of these training effects. I trained participants over a period of 2 weeks and tested them directly after this period. However, in order to have practical utility, such effects would need to last for extended periods of time. This makes it critical to test how long the observed training effects last.

Since training effects were tested in a perceptual decision-making task, the de-
gree to which these results generalize to more complex beliefs remains unknown. A
generalization to belief updating for climate change knowledge is a promising indi-
cator that these effects might transcend a wide range of societal issues. Importantly,
the approach of performance and confidence feedback can be easily implemented in
different tasks (e.g. one could directly train metacognition related to climate change
knowledge ([Fischer & Said 2020] [Fischer et al. 2019]). Stronger transfer would be
expected if the trained and tested domain are more similar to each other ([Taatgen
2013]. As long as a ground truth exists (i.e. there is an objectively right or wrong
answer), metacognitive training can be conducted for a range of tasks or knowledge
domains. It would be exciting to test this intervention in different domains in order
to see whether similar effects can be observed.

While there is more work to be done in order to optimize such cognitive training,
the results hold the promise of potentially broader societal implications.

\section{Conclusion}

In this thesis, I have investigated the cognitive, computational and neural mecha-
nisms that underly altered post-decision processing. I found that confidence in an
initial decision is a crucial factor inducing a neural confirmation bias. Based on
this knowledge, I designed and tested a metacognitive training procedure that im-
proved participants’ confidence judgment accuracy and through this increased their
post-decision processing. Further, I have shown that these cognitive alterations play
a crucial role in determining rigid and dogmatic beliefs, and thus have real-world
relevance.

Similar approaches might be helpful for studying motivational and societal influ-
ences on confirmation bias, as well as other cognitive biases. I hope that this thesis
convincingly advocates for the merits of computational approaches when studying
cognitive biases as this knowledge can directly inform targeted interventions and
improve decision-making.

Helping people to process information objectively and to make better decisions
might be one of the biggest challenges for human kind, in order to tackle issues
such as global pandemics, societal polarization and climate change. Ultimately, all
knowledge and scientific evidence is worthless if people do not take it into account
and act upon it.
Bibliography


relapse prediction after antidepressant discontinuation in a naturalistic setting. *bioRxiv.*


Center, P.R. (2017). The partisan divide on political values grows even wider. *Pew Research Center*.


Bibliography


Kaplan, J.T., Gimbel, S.I. & Harris, S. (2016). Neural correlates of maintaining one’s political beliefs in the face of counterevidence. Scientific reports, 6, 39589.


van Brussel, S., Timmermans, M., Verkoeijen, P. & Paas, F. (2020). ‘Consider the Opposite’–Effects of elaborative feedback and correct answer feedback on reducing
confirmation bias–A pre-registered study. *Contemporary Educational Psychology, 60*, 101844.


