Communicated Beliefs:

The interplay of evidence and truth values in erroneous belief acquisition and maintenance

Toby David Pilditch

Department of Experimental Psychology

University College London

Thesis submitted for the degree of

Doctor in Philosophy (PhD)

December, 2016

-This page intentionally left blank-

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

Signature:_____________________ London, December 15th, 2016

(Toby David Pilditch, BSc. MSc.)

-This page intentionally left blank-

ABSTRACT

This thesis explores the interplay between evidence (in terms of clarity, quantity and order) and source factors in the uptake and maintenance of second-hand, erroneous beliefs.

Through a series of online paradigms, we demonstrate the uptake of a communicated belief (e.g., “Option A is better than option B”) is conditional upon early experiences, given an unknown source. Further, we show that such a consolidation of belief then leads to a confirmation bias, wherein beliefs are maintained despite long-run contradictory evidence. Importantly, we demonstrate that such a bias occurs *despite* participants being motivated towards accuracy (as opposed to belief-maintenance), and the presence of counterfactual information. We accordingly forward an integrative confirmation bias account of consolidated belief maintenance.

The focus then turns to explore the gatekeeping role of early experiences. Using short-term fluctuations in evidence, we not only demonstrate the impact of the first few pieces of evidence in consolidating beliefs, but that such effects are interruptible. Taking insights yielded from the Bayesian source credibility model, the perceived expertise and trustworthiness of the source are then manipulated in conjunction with initial evidence. In line with predictions, beliefs from credible sources show consolidation prior to initial evidence, subsuming its role. Conversely, beliefs from dubious sources once again demonstrate the critical impact of initial evidence. Findings are related to the role of source cues and early experiences in increasing the confidence in a belief’s validity, placed within the wider theoretical context, and novel implications for reliability updating are demonstrated.

These empirical findings are then extrapolated to belief propagation in online networks using Agent-Based Modelling. This work demonstrates that the structure and incentives present in online networks exacerbate societal level erroneous belief uptake.

Implications are drawn to literature including persuasion, placebo effects, and opinion dynamics, along with phenomena including superstition and pseudoscientific beliefs.

This began with a question.

It ends with some answers, but even more questions.

It is a trade I make gladly.

ACKNOWLEDGEMENTS

There are many to whom I owe a debt of thanks for even being able to write these words. Although I shall try to be thorough in my gratitude, if, Dear Reader, you do not find your name where you feel it ought to be, then along with my thanks, you have scored yourself an apology to boot.

First and foremost, to Dr. Ruud Custers: You took a chance on this topic, and on me. That is no small thing, and for it, I will always be grateful. Your open-mindedness, enthusiasm, and guidance have all been infallible. I look forward to our continued collaboration in growing this research into something greater.

To the myriad faculty members who have informed this work, including (but not limited to); Professors Bradley Love and David Lagnado, Dr. Miguel Vadillo, Dr. Stephanie Baines, the Lagnado Lab, and MoBeLab: Your comments, criticisms, and counsel I have been grateful for in equal measure; they have remade this work stronger each time.

To the cast of room 201, past and present; Neil Bramley, Sarah Jenkins, Dr. Keith O'Brien, Sarah Smith, Dr. Laura De Moliere, Dr. Christos Bechlivanidis, and more: The conversations, advice, and general good humour I could not have done without – in fact, I would have likely been bloody miserable. Christos, I am particularly grateful for our discussions of causality, programming, experiments, and everything in between – time well spent.

To John Draper, Pia Horbacki, and Dr. Joseph Devlin: Your support, encouragement, and understanding have never failed to keep this ship on an even keel. Put simply, your attitude, support, and the opportunities you have afforded me have made life easier – you have my thanks.

To the Economic and Social Research Council, Dr. Leun Otten, and the ESRC support staff: Without your generous financial support, neither this research, nor a life in London, would have been possible for me. I hope the present work, and any publications derived from it, help assuage my debt to you. To use the official expression once more: The research in this thesis was supported by the Economic and Social Research Council [grant number ES/J500185/1].

To Dr. Jens Madsen: There are many things I could (and should) say, but I will have to settle for just one: Your presence has helped me more than you can know.

To my family; Mum, Dad, Alex, Rosemary, and Ted: You have all contributed in equal measure. Ted especially, actually[1].

To Riikka: I am not a big fan of sentiment, but I can summarise your impact accordingly: If, over the months and years of this thesis, I have become a better man – that is down to you.

[1] For those who do not know, Ted is, at time of writing, an 8-year-old black Labrador. He is a good dog.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

"*When doubt ceases, mental action on the subject comes to an end; and, if it did go on, it would be without purpose.*" – Charles Sanders Peirce, 1877

Between 2001 and 2007, a meta-analysis in the Journal of the Royal Society of Medicine found that not only did spinal manipulations (or chiropractic treatment) show no conclusive medical benefits beyond placebo, but in fact lead to routine adverse effects (Ernst, 2007). These included (but were not limited to) dissection of vertebral and carotid arteries, followed by stroke, and various forms of nerve damage. However, the number of registered chiropractors has risen steadily with demand since 2007, to over 3000 practitioners as of 2015[2].

Similarly, the National Health Service of the United Kingdom spends £4million of taxpayer money annually on homeopathy (excluding running and maintenance costs)[3] despite evidence demonstrating such treatment initiatives are no better than placebo (Ernst, 2002).

Despite the availability of evidence, such erroneous beliefs are still prevalent in modern society. Critically, the most prevalent beliefs, and consequently most harmful for society, are spread via second-hand information. Psychology is well equipped to address the natural question that arises from this: How are communicated beliefs adopted and maintained without supporting evidence?

Although such a question has provoked research in the past, the present

[2] Taken from the General Chiropractic Council report 2015: http://www.gcc-uk.org/UserFiles/Docs/Registrations/Report%20on%20the%202015%20registration%20year%20160616.pdf

[3] Taken from House of Lords Enquiry into alternative treatment availability on the National Health Service: http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/45/4504.htm

approach provides novel insights into the process, with particular regard to the interaction of communicated beliefs and the early experiences of recipients. A holistic approach is adopted, in which cognitive processes, motivational goals, social cues, and the structure of both evidence and beliefs are considered. To help establish this approach, Chapter 2 starts by providing a broad overview of the literature on learning, gradually turning to the study of second-hand information based learning. From this broad platform, the review then focuses on when second-hand information (mis)matches the evidence at hand. Mechanisms of belief transference and maintenance are then discussed, with particular reference to cognitive bias and motivated reasoning accounts. These elements are then synthesized into a theoretical framework that incorporates order effects, the role of ambiguity (both of belief and evidence), and the overweighting of confirmatory (belief-matching) instances as evidence is integrated.

With the theoretical backdrop established, Chapter 3 provides an empirical grounding for the investigation of the biasing consequents of beliefs, in line with previous research in the related fields of study. The experiments detailed within build the case for an integrative account of the belief-biasing effect, ruling out alternative explanations and identifying the role of initial evidence. Chapter 4 then concentrates on this factor, improving the methodology and answering some concerns raised in Chapter 3. This empirical work demonstrates the potency of both second-hand beliefs and initial evidence independently on choices and judgements in the short-term, and the impact of their conjunction. Chapter 4 also considers and tests factors that impede or interrupt such effects, with some success.

Up until this point, the empirical work presented has demonstrated the role of initial evidence as a gatekeeper to subsequent belief-preserving biases, when the sources of such beliefs are purposefully stripped of credibility cues. Consequently, Chapter 5 then builds off this work through the incorporation of source credibility elements. By manipulating the perceived trustworthiness and expertise of the source, the role of initial evidence as a belief validator is further exposed. Specifically, if cues indicate the source of a belief is reliable, the role of initial evidence is diminished. Conversely, when such cues instead cast doubt on the reliability of a source, initial evidence comes to the fore. This chapter also provides an important demonstration of the power of high trust sources in evidence integration processes (and an illustration of how these sources benefit from this deception), along with self-fulfilling prophetic effects of low trust sources and increased skepticism.

The above empirical chapters are designed to be broadly self-contained, and as such may be read in isolation. Any resultant overlaps in areas of the introductory sections are an unavoidable consequence of this structuring, but can hopefully serve as useful reminders to the dedicated reader. Chapter 3 is based on the article "Communicated Beliefs about Action-Outcomes: The Role of Initial Confirmation in the Adoption and Maintenance of Unsupported Beliefs" by Pilditch and Custers, submitted at time of writing. Chapter 4 is based on the article "Communicated Falsehoods in Medical Decision Making: The Impact of Initial Evidence and Counterfactuals" by Pilditch and Custers, which at time of writing is in preparation. Finally, Chapter 5 is based on the article "False Prophets and Cassandra's Curse: A little trust goes a long way." by Pilditch and Custers, in collaboration with Dr Jens Madsen (Oxford University), which at time of writing is also in preparation.

The empirical findings of these three chapters are then extrapolated to a societal level using an Agent-Based Modelling technique in Chapter 6. This form of modelling ascribes general rules and values to individuals (or "agents"; e.g., buying behavior rules, and values to represent the agent's current wealth), then creating a multitude of agents within a simulation space. These agents can incorporate stochastic processes and draw individual values from distributions, creating heterogeneity in the population. They are then free to interact within the constraints of the system over time, creating conditions under which dynamic processes may evolve. Of particular interest are forms of emergent behavior that are not predictable from base (individual-level) conditions. All processes and behaviors occurring with the system are measurable, allowing for outcomes to be assessed in relation to different system conditions.

The work in Chapter 6 simulates agents within a social-network setting, demonstrating various belief propagation outcomes across networks, including penetrative capacity, speed of spread, and degree of clustering. These outcomes are found to be critically impacted by the actions of those who refute communicated beliefs, and the degree of interconnectivity within a system, among other factors. These exploratory insights not only highlight the exacerbating impact of interactive network structures on the individual-based findings and proposed mechanisms of the empirical work, but serve as a demonstration of the unique insights Agent-Based Modelling can provide to accounts of cognitive processes.

The implications of these findings, along with the conclusions from the preceding empirical chapters, are then synthesized in the General Discussion (Chapter 7). This highlights the key findings, placing them within the context of the overarching narrative, and draws fresh insights when considering the work as a coherent whole. It is at this point that the limitations of the work are brought forward, along with considerations of further research. Finally, implications are drawn to literatures including impression formation, persuasion, and placebo research, along with more applied domains including personal health, politics, pseudoscience, and internet-based communications.

CHAPTER 2: LITERATURE REVIEW

The purpose of this chapter is to serve as a general review of psychological literature concerned with the transfer of knowledge from secondary sources. It initially focuses on the obvious learning advantage communication provides when encountering new environments, including an overview of propositional learning, before spelling out the basic cognitive and motivational explanations behind such a capacity. The review then turns to areas in which communicated knowledge and objective evidence are not in accordance. This discrepancy is framed in terms of the benefits of adherence to a communicated belief versus the costs of acting out-of-step with the objective evidence. Through the lens of this cost-benefit analysis of belief, attention is brought to areas in which the benefit of belief outweighs potential costs (such as impression formation, placebo effects and certain superstitious beliefs) to the individual actor. From here the focus will shift towards the opposing side of the equation – when the potential costs of adhering to a demonstrably false belief outweigh the benefits.

This final circumstance provides the context and motive for the line of research in this thesis. It is at this point that the review turns to the logical question that follows from the premise of maladaptive belief adherence: what mechanisms are behind how such false beliefs are adopted and maintained? In answering this question, both cognitive and motivational accounts will be reviewed and synthesised into a broad theoretical framework that is explored throughout the remainder of the thesis.

2.1 The Study of Knowledge Transference

The capacity to transfer knowledge from one agent to another is possessed by many species within the animal kingdom, including bees (Frisch, 1950), chimpanzees (Bradbury & Vehrencamp, 1998), and of course, humans. The methods through which knowledge is communicated are diverse, from visual displays such as body language and facial expressions in humans, to the bee's "waggle dance" to indicate food locations (Frisch, 1950), through chemical signalling in insects (Francke & Dettner, 2005; Hartmann, D'Ettorre, Jones, & Heinze, 2005), and back to vocal communication in humans, primates, and many other species.

One can think of the act of communication in terms of a cost-benefit analysis. On the one hand communicating requires an expenditure of time and energy (and in some cases an increased risk of predation), but on the other, an evolutionary advantage is gained – the recipient(s) receive information without the cost of direct-experience, whether this is food locations, quality of a potential mate, avoidance of danger or understanding of a social hierarchy (Pearce, 2013). In other words, the act of transferring knowledge from one agent to another (or multiple others) provides the advantage of experience to those *without it*, thus providing rules on how to act within a new environment without the inefficiency of having to discover the rule oneself.

Among all the life on earth, it is humans who have taken this capacity to unprecedented new heights. From passing on stories around the campfire and making crude markings on cave walls, we have developed the capacity to convey knowledge on a grander scale – across generations. The resulting accumulation of knowledge from consecutive lifetimes is generally credited as the source of all human culture (Donald, 1993). Over the course of history, humans have developed technologies to facilitate this communication of knowledge; from the development of language (both spoken and written) and the progression of mediums through which to convey it – in short order: chisel and stone; parchment and ink; the printing press (important for expanding the *availability* of knowledge beyond the privileged); telegram; telephone; radio; television; electronic documents; texting; and most recently the internet. As our world becomes more interconnected, and both physical and temporal distance becomes less important, the capacity to communicate information is more relevant than ever.

From a psychological standpoint, questions regarding the transference of knowledge (and learning more generally) have existed in various forms since the field’s conception (James, 1890). In reviewing the literature on how (and why) knowledge (broadly defined here as beliefs about one’s environment, such as action-outcome relations) is transferred, learnt from, and maintained, it is best to organise knowledge transfer by its outcomes. These outcomes can be ordered in terms of a simple cost-benefit analysis to the recipient, for the uptake and maintenance of the belief, which is based upon the accuracy of the belief relative to the realities of the environment (and the agent’s own outcomes within it). In other words, is the agent that has a belief better off when acting in the environment than an agent without such a belief?

When beliefs match evidence, the result is a net benefit to belief uptake and maintenance (the aforementioned advantage of knowing how to navigate an environment without having had to experience it first hand), and has been investigated in social learning theory (Bandura, 1977) and propositional learning (De Houwer, 2009; Mitchell, De Houwer, & Lovibond, 2009), among other domains. Following this are instances in which the communicated belief does not necessarily match the true state of the environment, but in maintaining the belief, one can still sustain a net benefit (Abbott & Sherratt, 2011). Examples include placebo research (see Wager & Atlas, 2015) both in medical and task performance contexts (Damisch, Stoberock, & Mussweiler, 2010), as well as self-efficacy and learned helplessness (Seligman, 1972). This benefit-despite-inaccuracy section introduces the psychological concepts and mechanisms that allow unsupported beliefs to be maintained (and even flourish), including various misperceptions of evidence and forms of bias. This is expanded further when turning to costly adherence outcomes, where the discrepancy between the content of the communicated belief and the objective evidence *should* disfavour adherence to the former. This final outcome category is the primary focus of the thesis, and the rationale behind the empirical chapters herein.

2.1.1 Communication of Knowledge for Effective Learning

To understand how communicated beliefs (or “second-hand” evidence) are used by an individual, it is necessary to take a step back to provide a context into which this form of learning might fit. As a field of interest in psychology, learning is among the oldest and broadest areas of research. Stretching back over a century to seminal work on conditioning (Pavlov, 1927) and the rise of behaviourism (Skinner, 1948), focusing on models of how an agent learns to associate one stimuli/event/action with another (Thorndike, 1931). Accordingly, before incorporating second-hand information into learning, one must first understand its place among these prior works and theories in first-hand learning, as the study of the latter has shaped the ground on which the former stands.

This early work was the precursor to the associative framework for describing the mechanisms of learning (Shanks, 1995). In broad terms, associative learning is defined as the bottom-up linkage of events or action, through repeated co-occurrence (Thorndike, 1931), otherwise known as conditioning. According to proponents of this theory, most forms of learning can be reduced to the consequents of directly experienced event / stimuli associations. Among the most famous theorists in associative learning, Robert Rescorla and Allan Wagner developed associative theories based on their learning paradigms using prediction error (Rescorla & Wagner, 1972). The Rescorla-Wagner model (R-W; see equations 1 & 2 below) symbolised this line of research, and has been used in various subdomains of learning since its inception.

$$V_X^{n+1} = V_X^n + \Delta V_X^{n+1} \tag{1}$$

$$\Delta V_{\mathrm{X}}^{n+1} = \alpha_X \beta(\lambda - V_{tot}) \tag{2}$$

In the model, the current strength of the association of X is represented by V_X^n in equation 1. The strength of this association on the next trial (V_X^{n+1}) is the sum of the current strength and change in association strength (ΔV_X^{n+1}). This change is calculated in equation 2 as the product of the salience of the CS, or conditioned stimulus (α), the association value for the US, or unconditioned stimulus (β), and the difference between the maximum conditioning possible for the US (λ) and the total associative strength of all CSs (V_{tot}). One of the principal benefits of such a model is its capacity to process event representations via both intensity and unexpectedness, and has as a model generated clear (and successful) predictions within the associative learning literature (for a review, see Siegel & Allan, 1996). Within the context of this section, the R-W model provides a parsimonious (there are relatively few free parameters within the model) mathematical formalisation of the associative learning process; most critically a description of how automatic updating of associative expectancies can occur. Accordingly, such a model is a useful point of mechanistic reference when discussing subsequent associative learning literature.

A prominent paradigm of research in which associative learning theories have been developed is Evaluative Conditioning (EC). It should be noted that EC is but one of many lines of learning research, and is given prominence in the present review only to serve as a singular, coherent approach through which relevant effects and theory development may be illustrated. This research stems from the premise that likes and dislikes about objects or events are predominantly *learned* rather than innate (Rozin & Millman, 1987). EC procedures look at the change in liking (or valence) of a stimulus (the CS) that results from being paired with either a positive or negative stimulus (the US). Thus, the negative / positive valence of the US, through co-occurrence, becomes associated with the CS (for a thorough review on EC research, see De Houwer, Thomas,

& Baeyens, 2001). In this sense, EC is a form of Pavlovian conditioning, wherein associations are learnt in a bottom-up manner.

EC paradigms have used stimuli from a diverse range of domains, including visual (Levey & Martin, 1975); haptic (Hammerl & Grabitz, 2000); and even cross-modal (Reekum, Marije, van den Berg, & Frijda, 1999). A typical EC paradigm, such as that originally used by Levey and Martin (1975), uses a forward conditioning procedure, whereby presentation of the CS (target for valence change; in this case pictures of paintings that participants originally judged to be neutral – neither positive or negative) precedes the presentation of the US (stimuli providing the valence; in this case pictures of paintings initially judged by participants as positive or negative). These pairings are repeatedly presented in counterbalanced conditions with a "control" pairing – in the case of Levey and Martin, a neutral-positive pairing, a neutral-negative pairing and a neutral-neutral baseline "control". After an acquisition phase in which these pairings were each presented 20 times, participants completed a series of scales assessing their degree of liking (-100 for maximum disliking, to +100 for maximum liking) for each of the pictures once again, based on their global and spontaneous impressions for each stimulus.

Through this paradigm, Levey and Martin (1975) successfully demonstrated that the rating of a CS shifts towards the valence of the paired US. In other words, a previously rated neutral picture is subsequently rated more positively if it was paired with a positive US, with an equivalent shift in the negative direction if the US was of negative valence. The implication for associative theory was the demonstration of associative transfer (in this case of valence), which was subsequently demonstrated to be possible without awareness of CS-US contingencies, simply as a consequence of events or stimuli co-occurring. From this foundation, research on EC, and by proxy associative theory, expanded to look at the aforementioned alternative mediums (such as cross-modal), and in doing so searched for possible boundary conditions (for a review see De Houwer et al., 2001). One such area of expansion, which bares particular relevance to both of the theories of learning under discussion, as it does not rely exclusively on *direct* experience, and as it shifts closer to the methodological ancestors of the paradigms developed in this thesis, is Observational EC.

Work on EC has shown that effects can also occur through observation of another individual (Baeyens, Vansteenwegen, De Houwer, & Crombez, 1996). By looking at Observational EC, we start to move away from learning purely as a consequence of direct experience. In this way, the contingency between CS (in this case the consumption of coloured drinks of various flavours) and US was vicariously experienced. The children who participated consumed the drink (CS) and then witnessed an actor on videotape drinking the same drink, and then making either a neutral or disgusted (negative) facial expression (US). Accordingly, the negative valence associated with the actor's expression (having imbibed), led to participants rating their own similar drinks as more negative – a transfer of valence. In summary, this work demonstrated, through the communication of affective states in association with a stimulus, that associative links can be learnt without *direct* experience.

EC is not the only area to broach the topic of observation in learning. Social Learning theory (Bandura, 1977) in the social psychology literature demonstrated that behaviours could be learnt vicariously through witnessing the acts and experience of another individual, most famously with children replicating the aggressive behaviours towards a doll having witnessed an adult behaving aggressively. It is important to note, however, that vicarious experience (such as through observing another actor) does not equate fully to second-hand evidence in the communicated belief sense, as evidence is still being *witnessed* first-hand. However, returning to the observational work in EC starts to show the efficacy of communicated information, in this case instruction in learning.

The aforementioned observational EC research (Baeyens et al., 1996) was replicated and extended to include manipulations of the observer's beliefs (Baeyens, Eelen, Crombez, & De Houwer, 2001) regarding the relationship between the drinks of the actor and the participant (same / different / no information). Interestingly, conditioning effects were only found when there was either no information or information indicating the drinks were the same. Given that participants in these conditions were, however, unaware of the source for the dislike (i.e. no CS-US contingency awareness), the causal involvement of belief in a conscious manner that influences cognitive inference processes cannot necessarily be inferred. In fact, the lack of awareness lends support to the argument that the observed stimulus of a facial expression acts simply as an automatically associated US (as in traditional EC paradigms). Nevertheless, the effect of "communicated" information in this context lends some early support to the notion of beliefs being able to affect learning, in this case by modulating the degree of attention to the model's facial expressions.

2.1.1.1 Propositional Learning

Before formally delving into the role of propositions (De Houwer, 2009, 2014; Mitchell et al., 2009), the notion of "instruction effects", which one can broadly classify the above Baeyans and colleagues (2001) study as illustrating, is worth casting further light upon. Instruction effects relate to a fundamental pillar of experimental psychology – how to describe a task to a participant in order to control for different interpretations of how each participant may naturally interpret how to act within the task space. This "constraining" that instructions provide can be seen as something of a double-edged sword. On the one hand, you are increasing your experimental control and in theory providing homogeneity of context across participants. On the other, the participant is being directed towards the desired effects, and given this artificiality, undermining the validity of the effects themselves. Issues regarding the role of instructions (and their interpretation) have previously been noted (Hertwig & Ortmann, 2001). Research in social cognition has started to manipulate instruction (known as “instruction effects”) to investigate the impact that these formative propositional statements (which one could reasonably define a typical experimental instruction as being, for example, “Press “x” when you see [stimulus]”; “Choose between option A and B, depending on which you think is best.”) have on various psychological effects (Doll, Jacobs, Sanfey, & Frank, 2009; Mertens & De Houwer, 2016; Roswarski & Proctor, 2003; Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel, Gawronski, Smith, & De Houwer, 2016). One example of this work, in investigating the power of such statements, in terms of a *psychological* effect, is Bernard Hommel’s (1993) work on the Simon effect.

Literature on the Simon effect measures the associations between *actions* and *outcomes* – unlike the affective valence transfers of EC effects. Initially demonstrated by Simon (1969), the Simon effect is a decrease in choice reaction times when there is congruency on a particular dimension (e.g., spatial position) between the action and the outcome – even when that dimension is irrelevant to the task. For example, a typical Simon effect paradigm presents participants with red or green lights (Hedge & Marsh, 1975), for which they are instructed to press the left-hand button when the green light appears and the right-hand button when the red light appears. Both lights can appear on either the left or right side of the participant’s visual field. Even though the location of the light stimuli is irrelevant to the task, there is a facilitation effect of spatial correspondence between stimuli and intended action– when the green light happens to be on the left-hand side, the left-hand key press is faster in response than when the green light is presented on the right.

Simon effect research has typically kept the role of instructions to a minimum, whereby the participant is only instructed to respond with a particular key to a particular stimulus – *nothing* regarding spatial locations. Importantly, in reference to the discussion of propositional learning, work by Bernard Hommel (1993) demonstrated that the Simon effect can be reversed, depending on how the participant is instructed. Participants performed a task in which they responded to either a high or low tone, presented on their left or right hand side. Upon hearing the tone, the participant needed to press the correct key (e.g., left-hand key for a high tone), which would light up an LED either to their left or right. These three categories could then be arranged into different conditions and trial types, to test various questions. Of particular note is the comparison between trials where the key press would light up the LED on the opposite side, where participants were instructed that their goal in responding to the tone was either to press the particular key, or light up the particular LED. This manipulation of instruction led to a reversal of the Simon effect, wherein the instruction to focus on turning on the LEDs, rather than the pressing of keys, resulted in decreased reaction times when LED side matched the tone presentation side, even though the key pressed to turn on the LED was on the opposite side (incongruent). When the LED was on the opposite side from the tone (incongruent), but key press was on the same side (congruent), if the instructed goal was to light LEDs, then a delay in reaction times was found. As such, these instructions led to a reversal of the key-press – tone stimulus facilitation effect found in typical Simon effects. Such an effect was still found by Hommel when instructions focused on keys, rather than LEDs.

In summary, Hommel's work neatly demonstrates the power of instructional cues in how individuals approach (and respond to) a new environment. Such a demonstration leads back to the EC literature, where the role of instruction, and more generally the development of propositional learning theory, has in recent years gained traction (De Houwer, 2009; Gawronski & Strack, 2004; Mitchell et al., 2009), with instruction effects found on fear conditioning (Raes, De Houwer, De Schryver, Brass, & Kalisch, 2014), extinction and counterconditioning, (Gast & De Houwer, 2013) and approach-avoidance effects (Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016).

Propositional learning theory's principal theoretical distinction from associative theories is the form of mental representations within learning. Associative theories assume learning is represented solely as a function of repeated co-occurrences of CS and US, which in the case of EC then results in a transfer of valence (De Houwer et al., 2001). Propositional theories instead propose a non-automatic, top-down effect of propositional statements regarding associations, wherein learning through the experience of events can be influenced by factors including prior knowledge (Waldman & Holyoak, 1992; Lopez, Cobos, & Caño, 2005), instruction (Hommel, 1993, as we have touched upon above; De Houwer, 2002), deductive reasoning (De Houwer & Beckers, 2002) and interventions (Waldman & Hagmayer, 2005). In this way, associative effects occur through the formation and evaluation of propositions (De Houwer, 2009).

Within contingency learning, the phenomenon of "blocking" has been one of the pivotal findings in support of propositional theory (Kamin, 1969). Participants see repeated co-occurrences of a cue (CS 1) and an outcome (US), after which they make a judgement regarding the relationship between these two events. Typically, the presentation of the cue (CS 1), will lead to a conditioned response (CR) that is a result of anticipation of the US. Blocking occurs when a second cue (CS 2) is then also paired with CS 1 and the US, participants do not learn to associate it with the US; i.e. they do not see CS 2 as predictive of the US (despite its co-occurrence with the US), and thus presentation of CS 2 by itself does not provoke a conditioned response. Consequently, prior knowledge (although other blocking studies have shown intervention, experience, and instruction effects) can directly impact contingency learning (De Houwer, 2009).

Given the presence of such internal propositions, non-automatic processes, such as those listed above, are able to intervene on the process of learning. This bears relevance to the thesis topic at hand – providing a theoretical grounding for the link between communicated knowledge (which can be brought under the banner of propositions, whether in the form of instruction, prior knowledge, intervention, or a combination of these categories) and the learning process through experience.

Propositions have been shown to not only indicate the *presence* of a relation, but also provide an indication of how such a relation is *structured* (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Work in causal reasoning by Steyvers, Tenenbaum, Wagenmakers and Blum (2003) demonstrated this structurally indicative capacity by assessing whether a causal model of a set of events (i.e. causal structure) could be inferred by participants through statistical covariation alone (as would be the sole requirement for associative learning). In their experiments, in which participants needed to infer which mind-reading aliens can send messages (causes) and which can receive (effect), observational data alone resulted in performance just above chance level. Furthermore, this performance was even assisted by prior knowledge concerning which causal structures were possible candidates. It was only with the addition of interventions (whereby the truth value of proposed possible structures could be tested), that performance significantly improved (Lagnado & Sloman, 2004). Thus, through the formalisation of a propositional statement (in this case indicating the causal structure at hand), learning occurs through the capacity to test the truth value of such a proposition.

Similarly, work in the advice taking literature has investigated the role of propositional statements (termed “description”) regarding risky versus safe choices (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997; Siegrist, Gutscher, & Earle, 2005; Twyman, Harvey, & Harries, 2008; Yaniv, 2004). Typically, in such paradigms participants are exposed to a task / question (often a judgement coupled with a confidence rating), at which point they make an initial estimate. Following this, participants then receive “advice” from either the experimenter (Harvey & Fischer, 1997; Newell & Rakow, 2007) or other participants (Yaniv, 2004), indicating an opinion / preference judgement (the advisor’s interpretation of the problem – often quantitative), along with the participant’s own previous judgement. The participant then makes a final judgement incorporating the advice (or not). Using the distance between the participant’s initial and final judgements, it is possible to determine how much (or little) the participant incorporated the advisor’s information, along with measures such as changes in confidence and accuracy (Bonaccio & Dalal, 2006). Results have shown that not only can participants use advisor information to improve the accuracy of their own judgements, but in multiple advisor systems participants can also assign values to these advisors based on the validity of each advisor’s statements (propositions) over time (Sniezek & Van Swol, 2001; Yaniv & Kleinberger, 2000). Taken together, these effects demonstrate both the capacity to incorporate, and advantage of using, second-hand information.

The aim of this section has been to give the reader a brief overview of how some of the theory of learning in psychology has developed to include the role of propositions and second-hand information. In doing so, literature from both social and cognitive fields has been brought in to both lay a broad foundation for the study of communicated beliefs, and provide some foreshadowing of methodological techniques and mechanistic explanations which form part of the backbone for this thesis. Further, the synthesis of work from social and cognitive domains will be a common occurrence throughout, as both provide insights into the mechanisms involved with belief uptake, adherence and biased evidence integration. To frame this work within only one of these two domains would result in a short-sighted approach to what is, in essence, both a social and cognitive phenomenon.

2.1.2 Belief-Evidence Mismatches: Advantageous Adherence

Determining the true nature of an environment is an inherently difficult task for an imperfect observer. This difficulty results in an obvious consequence: communicated knowledge is not always accurate. It is important to note when trying to define the voracity, or accuracy, of a piece of knowledge, that there is a logical question that arises: accuracy relative to what? The true state of the environment? Is this even a realistic or pragmatic baseline for comparison in the first place? Perhaps one could draw comparison to the behaviour of an agent in that same environment without the encumbrance of this "erroneous" prior knowledge? In either case, there is a cost / benefit to each action within the environment, and the utility of such outcomes may not line up neatly with knowledge of the "true state" of the environment. One such example is the common conception that "every vote counts". In a population of millions, the personal costs to going out and voting (leaving aside time spent considering who to vote for) can outweigh the value of the casting of the single vote among the other millions. However, even though from the point of view of a single individual, the belief that every vote counts might appear to be "erroneous", other benefits – such as to the functioning of a democracy as a whole, or to the prospective sense of vicarious victory of the chosen candidate through a sense of group membership – outweighs this error. Furthermore, as illustrated by this example, irrespective of the chosen accuracy "baseline" and the potential uncoupling of "truth" and agent outcome utility, there is room for technically "deviating" beliefs that carry advantage to the believer.

The difficulty in determining accurate beliefs stems from two domains; opacity of the environment itself, and limited capacity of the agent within it. These two domains are in fact two ends of the same line – the greater capabilities of the agent, the less opaque the environment it inhabits, and conversely, the more complex and opaque an environment, the less capable the agent appears. Given this relationship, there exist instances in which the agent must engage with situations wherein the "correct" course of action is not feasibly calculable, and so must resort to more "quick-and-dirty" alternatives.

Originally described by Simon (1955), bounded rationality – the idea that an agent is limited by biological and temporal constraints –put forward the notion in psychology that the human mind, as a computational apparatus, is imperfect. An example of the difficulty agents can have in discerning workable rules from an environment comes from the superstitious responding literature (Catania & Cutts, 1963; Ono, 1987; Rudski, Lischner, & Albert, 2012). Typical paradigms involve asking the participant to discern what rules govern their receipt of a reward (indicated by the illumination of a light), with a series of switches, levers and buttons in front of them. Participants then spend time coming up with various "rules" based on actions/behaviours that coincided with the light turning on, such as "If I press the blue button three times, then pull the right-hand lever and switch the middle switch at the same time, I get a reward.". In truth, the receipt of a reward is governed solely by a 30 second timer, and would occur irrespective of the action taken by the participant. One potential criticism of this work is that the environment the participant is placed within, i.e. an experiment in a lab, where there has been some apparatus put in front of them *implies action*, making the necessary way to uncover the true rule for rewards – *inaction* – unlikely (Harris & Osman, 2012). This is further reinforced by the role instructions provided to participants by the experimenter play in shaping expected behaviour (as mentioned in the previous section), and provides an example of the way in which an appeal to authority (Walton, 1997) works.

Appeals to authority are a form of logical fallacy wherein an argument rests on being true because an authority / expert source says it to be true – which does not necessarily equate to truth. Although logically fallacious, there is some rationality to this form of thinking (Goodwin, 2011; Harris, Hahn, Madsen, & Hsu, 2015). Especially in instances where the agent has little to no experience of an environment, there is a benefit to adopting communicated beliefs. Given that the agent has a finite amount of time and experience, depending on the efficiency and accuracy of the communication process, receiving information from others can be of great benefit. This benefit also aligns with the potentially large downsides in having to explore first-hand (e.g., listening to the village elder about the dangerous current in the nearby river, versus finding this out first-hand). Accordingly, as a form of heuristic or mental shortcut, adhering to communicate beliefs as a general rule makes evolutionary sense (Abbott & Sherratt, 2011). For example, when out foraging in the woods a nearby bush starts rustling, a member of the group has been told that bushes rustling means a bear might be inside and the group should run immediately. Upon communicating this to the group, there are two choices, each entailing different cost / benefit outcomes. Either the group believes the communicator and runs, or doubts the communicator – having never seen a bear in the bushes themselves. In the former case, if the rustling was actually the wind, the group incurs the small cost of forgone time and energy that could have been spent foraging. If the rustling was in fact a bear, the group has run away and survives. In the latter case, if the rustling was wind, the group has not incurred the small cost. However, if there was in fact a bear in the bushes, the group incurs a substantially higher penalty. Thus, even if the likelihood of a bear is relatively low, the vast asymmetry in the prospective costs of adhering versus ignoring the belief makes the fact the belief may be mostly incorrect (in most cases, there is no bear) an unwise reason to refute it.

One example of this type of reasoning, wherein the uptake and adherence to a communicated belief is advantageous despite its falsehood, is in placebo effects. Typically, a placebo effect is defined as a beneficial effect produced by an intervention (e.g., taking a pill) that is due to the patient’s belief in the intervention, rather than the properties (i.e. medically active ingredients) of the intervention itself. Put another way, a placebo effect is a beneficial response to the context in which a treatment is delivered, rather than the specific actions of the drug (Wager & Atlas, 2015). Part of this “context” is the formulation of the treatment’s description to the patient, along with the authoritative source of this information (i.e. the medical practitioner). The trust in, and adherence to the patient’s belief that the treatment will cure them, despite this belief’s inaccuracy (the causal explanation of the treatment – it’s medically active components – is false), can result in a net benefit to the patient.

Dependent on the medical condition the placebo is being prescribed for, not only does the rate of placebo responses differ, but the prospective costs of the falsehood of the belief also change. For example, conditions in which there are low rates of placebo responses (e.g., most forms of cancer), and there are empirically supported medical treatments (e.g., chemotherapy), the cost of adhering to a placebo-based treatment (especially given the forgone option of evidence-based treatment) far outweigh potential benefits (Blanco, 2017). However, conditions known to be highly placebo responsive, such as back pain (Evans, 2004), which can be difficult to treat medically – or at least in doing so can carry large costs (e.g., liver damage from high dosage painkiller regimens), present net benefits for adherence. Thus, in this latter case, despite the absence of treatment-supporting evidence, the belief that the treatment may help carries advantage.

Returning once again to the notion of the perceived costs and benefits of adhering to a belief, there is a final factor that must be incorporated into the discussion – motivation. This factor has purposefully been left until last, as there are arguments for motivated reasoning behind belief adherence being both a benefit and a cost, thus placing it somewhat in limbo between this current section and the costly adherence section that follows. Explicitly, these arguments can be synthesised onto a single dimension – to who are costs and benefits referring: the agent as an individual, or society as a representative of the "average" agent.

In the former case, motivations typically provide an additional weighting framework that alters the costs and benefits of belief adherence, even when such a belief may not be "accurate". For example, in the case of acting within an uncertain environment, people are motivated to reduce their uncertainty (Curley, Yates, & Abrams, 1986; Keinan, 1994; Kusec et al., 2016; Woolley, Bigler, Markman, Reeves, & Whitson, 2015) and exert control over the environment, with research showing that there are both individual differences in this need (Webster & Kruglanski, 1994), and that people actively seek out information to resolve it (Becker & van der Pligt, 2015; Choi, Koo, Choi, & Auh, 2008; Strojny, Kossowska, & Strojny, 2016). Further, the inability to resolve this need results in a "learned helplessness" (Matute, 1994; Seligman, 1972); a form of choice paralysis, whereby an agent no longer attempts to be active within the environment due to perceived inefficacy. In the case of communicated beliefs, the adherence to a belief that provides a rule for acting within the environment (especially in cases where there is no obvious rule for action) provides a subjective benefit to the agent through the resulting reduction in uncertainty (Matute, 1995; Rudski & Edwards, 2007) – despite the potential inaccuracy of such a rule.

Similarly, agent motivations alter the costs and benefits of belief adherence in situations where the belief forms one part of a larger context, such as in self-concept preservation (Heller, Komar, & Lee, 2007). Specifically, the cost of abandoning a belief is increased due to the damage it may cause to associated aspects of the agent's belief system or self-concept, avoiding cognitive dissonance (Festinger, 1962). For example, if you derive your identity as being a fundamental Christian, then being told that evolution is a lie by your pastor now forms part of your belief structure. Accordingly, the cost of admitting such a belief is false now impacts on your faith in the pastor, and possibly your faith as a whole, which includes further motivations such as a fear of death (King, Hicks, & Abdelkhalik, 2009; Schmeichel et al., 2009), commonly associated with religious belief. Given this increased cost, based on the motivations of such an agent, the relative costs of believing, on an *individual level* (i.e. denying evolution), are lower than the prospective cost of abandonment, thus making adherence to the belief, despite its inaccuracy, beneficial to the agent. One can argue similarly for political ideologies that involve the scapegoating of minorities, wherein despite no evidence for adhering to such a belief (e.g., "Immigrants are taking our jobs!"), for the individual concerned, such a belief provides a psychological benefit in that worries / concerns / fears about the economy or society can be neatly explained and uncertainty reduced.

However, in these latter cases, there is an obvious cost to society as a whole for adherence to such beliefs: denying science is a regressive and dangerous tendency that hampers progress and can harm others; scapegoating minorities leads to persecution and fails to address the true causes of a society's problem. One could argue similarly for global warming denial : although there is an obvious cost to the world in ignoring evidence on such a matter, an individual may adhere to the belief as the long-term consequences are unlikely to affect them on an individual level.

Consequently, in the last few paragraphs, we have started to lay the groundwork for instances in which beliefs are not supported by evidence, but are adhered to nonetheless. So far, we have focused on those beliefs that can still carry a net benefit to the believer, despite their inaccuracy. Latterly, this has led us to a category of beliefs in which adherence may carry a net benefit to the individual agent, but also carry a net cost to society as a whole. It is these latter beliefs, along with instances where even a net benefit to the individual is questionable, that we turn to next.

2.1.3 Belief-Evidence Mismatches: Costly Convictions

At the time of writing, there have been over 1,400 cases of measles (a previously borderline eradicated disease) in the United States since 2010. The reason for this rise is attributed to the unprecedented number of unvaccinated children, which is a direct result of the fallacious (and repeatedly debunked) belief that vaccines cause autism[4]. In South Africa, 1,175 rhinos were poached, and across Africa and Asia, illegal poaching has pushed Western black rhinos into extinction (International Union for Conservation of Nature, 2011) and the remaining 5 rhino species into endangered or critically endangered.[5] Rhinos are poached for their horns, made valuable by large Asian markets that use the horn in traditional Chinese medicine to cure diseases such as gout, typhoid, food poisoning, rheumatism, and many more. Such a belief is based on an over 2000-year-old tradition, which has no scientific or medically proven basis.

The above are a couple of examples where the cost-benefit relationship between belief adherence and refutation should favour refutation. In many cases this is because the relative costs of adherence are greater when there is a readily available, better (in the sense of an objective criteria, e.g., medical proof) alternative to the maintained belief. Unlike situations in which an alternative option is not readily apparent, or otherwise is costlier than the belief option (e.g., an alternative that yields a better outcome than the belief, but comes at such a cost that on balance the low cost of the belief makes it more

[4] Taken from the Centers for Disease Control and Prevention, http://www.cdc.gov/measles/cases-outbreaks.html

[5] Taken from the conservationist NGO Save the Rhino, https://www.savetherhino.org/rhino_info/poaching_statistics

worthwhile), such as those mentioned in the previous section, belief adherence in this case can generally be considered harmful.

One category of belief generally associated with this form of cost is superstition. Within psychology, superstition has been studied in several different ways. Early research on superstition took an anthropological stance, with seminal work by Bronislaw Malinowski (1948), who investigated magical beliefs amongst an indigenous population of pacific islanders. Within the population, there were two types of fishermen: close shore, who stayed relatively close to the safety of the atolls, catching small but regular amounts of fish, and deep sea, who risked the dangers of long journeys in perilous waters for the chance of more lucrative catches. Malinowski found that the former group, which operated in safety and with minimal uncertainty, had relatively few superstitions, whilst the latter deep sea fishermen exhibited high levels of superstitious beliefs surrounding their occupation. Such work led to the study of superstition in correlational, individual difference frameworks. This focused on specific groups, such as athletes (Bal, Singh, Badwal, & Dhaliwal, 2014; Gaudreau, Blondin, & Lapierre, 2002) and exam-takers (Rudski & Edwards, 2007), correlating the propensity to hold so-called “irrational” beliefs with various traits, anthropological factors, as well as situational factors. Leaving aside definitional issues regarding the nature of whether such beliefs in each circumstance are as negative or costly as the researchers often imply, this research proposes superstitions are a natural reaction to establish control over one’s environment and reduce uncertainty (Keinan, 2002; Keinan, 1994).

Experimental work on superstition has typically been in the aforementioned superstitious responding literature, which focuses on “erroneous” beliefs that occur as a consequence of flawed first-hand learning (Catania & Cutts, 1963; Ono, 1987; Rudski et al., 2012). There are, as alluded to previously, issues with this work regarding the artificiality of the experimental paradigms typically deployed, and consequent claims of irrationality may not be justified. Namely, the set up and instructions of the experiment imply that action is necessary to gain points, when in truth the points accrue based on a timer outside of the participant’s control. Consequently, the participant is biased towards responding, and as a result is in the process of action when a reward event occurs, concluding a meaningful relationship from the co-occurrence of action and reward. It is thus questionable how valid a conclusion of consequent beliefs being irrational is given the context provided to participants.

Importantly, when regarding first-hand learning error accounts of superstition, a similar cost-benefit assessment can be made regarding the supposed “irrationality” of the belief’s adoption. Following on from the conclusions made regarding superstitions in correlational research, regarding the role of superstitions as a response to uncertainty (Keinan, 1994; 2002), this has been expanded to view superstitions as a form of cognitive “gap filling” (Serbin et al., 1960), caused by the difference in distance between the objective “true” state, and the subjective, “known” state. In this way, the agent seeks to “fill” the distance with some form of rule or belief that provides a satisfactory degree of [subjective] control over the ambiguity the perceived distance causes (Curley et al., 1986; Keinan, 1994). In fulfilling this role, the adoption of the belief in question (whether this is generated first-hand, or adopted from a second-hand source) satisfies the evolutionary need to detect causal relationships (Foster & Kokko, 2009). Importantly, whether investigated as a belief generated first-hand, or as a pre-existing belief-set within an individual, beliefs typically take the form of a directional hypothesis (e.g., “Pressing the left lever hard gives me the reward”, “Breaking mirrors leads to bad luck”, “My lucky charm improves my performance”), but are void of further quantification (e.g., “My lucky charm led to me getting 90% on this exam”). This generalised, verbal rule satisfies the necessary explanation required for control, facilitates easy transmission, but consequently makes falsification more difficult (Gilovich, 1993).

In the study of superstition, research has therefore focused on the generation of first-hand beliefs (Catania & Cutts, 1963; Ono, 1987; Rudski et al., 2012), and the correlational factors associated with those who possess them (Bal et al., 2014; Gaudreau et al., 2002; Rudski & Edwards, 2007). However, it is the process by which a belief can be communicated and then taken up by a naïve agent that is of key interest to this thesis, and, I would argue, is the more dangerous form of erroneous belief, and consequently worthy of investigation. The reason for this assertion is simple; on an individual level, a first-hand learning error can be costly, but such costs are typically limited to the individual in question. Conversely, when an erroneous belief is communicated, if it is believed / adhered to, then the cost increments with each transmission and can even follow power law like curves (Castellano, Fortunato, & Loreto, 2009). The notion of error propagation through erroneous belief transmission has been tangentially investigated in research investigating social influence in information cascades (Schöbel, Rieskamp, & Huber, 2016) and the correction of misinformation (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012).

The former work by Schöbel et al. (2016) demonstrated the cascading effect of information in a sequential decision making task. As a consequence of receiving information from an authoritative source – participants overweighted this "public" information above their own "private" information regarding the outcomes of a task. This leads to only public information being passed on, despite its errors, to the next participant – without correction. The latter work by Lewandowsky et al. (2012) on the correction of misinformation further describes the dangers of such erroneous belief propagation, from the role governments and vested interests play in generation and dissemination, to the ever-expanding role of the internet in spreading beliefs at an unprecedented rate. As Jonathon Swift presciently once put it in 1710: “Falsehood flies, and the truth comes limping after it.”

Cognitive literature has started to investigate this previously social psychology based phenomenon. Work in causal learning (Yu & Lagnado, 2012) has shown that giving participants an incorrect prior regarding the distribution of possible outcomes in a one-armed bandit gambling task can result in distorted interpretations of the outcome space, even after direct experience with the outcomes in question. Similarly, the aforementioned advice taking literature in decision making has been investigating the effect of what can be described as second-hand information in the judgement process (Harvey & Fischer, 1997; Siegrist et al, 2005; Twyman et al, 2008; Bonaccio & Dalal, 2006; Yaniv, 2004). However, this work has only recently started to look at when description and experience disagree (Collins, Percy, Smith, & Kruschke, 2011; Lejarraga & Müller-trede, 2016; Weiss-Cohen, Konstantinidis, Speekenbrink, & Harvey, 2016), with some promising effects regarding the potential dominance of described (second-hand) over experienced (first-hand) evidence.

Over the course of this section, both the prevalence and potency of communicated beliefs have been discussed, along with empirical findings in several branches of literature that indicate the power of such beliefs in seemingly overriding first-hand experience. The natural question this final section provokes is, given the potential costs to the believer, or society as a whole, for what reasons – whether motivational, cognitive, or a combination of factors – does an agent uptake and adhere to an erroneous belief?

2.2 Mechanisms of Transfer and Maintenance

So far, we have determined that humans have the propensity to both seek out and learn from communicated knowledge, with the capacity to then use this knowledge to develop working rules for an environment without the requirement of first-hand experience. This process, depending on the degree of accuracy in the belief and the costs / benefits associated, results in three main categories of outcomes: an accurate belief, resulting in effective and efficient learning; an inaccurate, but not costly belief, resulting in an overall either neutral or beneficial consequence; and an inaccurate and costly belief, which should result in abandonment. It is the latter of these outcomes that provokes the most interesting question: how and why are these erroneous and costly beliefs adhered to?

In answering this question, we turn once again to the cognitive psychology literature for a discussion of mechanisms of bias, along with a look at the social psychological literature to understand where conflicting motivations can exacerbate such erroneous tendencies.

2.2.1 Cognitive Account

Normatively, when second-hand information is received, this should inform an agent’s prior belief about the given environment. However, if such a belief is incorrect, it should be updated by evidence to reflect its inaccuracy, and, depending on the factors that might inform the initial confidence in this prior (such as the perceived expertise and trustworthiness of the source), the belief should be abandoned. Such a notion follows a broadly Bayesian account of learning, where the posterior belief of a hypothesis being true is based on a weighted likelihood, known as Bayes Rule (see equation 3 below). In effect, the probability of the hypothesis being true, given the evidence (P(H|E)) is equal to the probability of the evidence occurring given the hypothesis (P(E|H)), multiplied by the prior likelihood of the hypothesis being true (P(H)), which in turn is divided by the probability of the evidence occurring (P(E)) irrespective of any hypothesis. This method allows for the incremental updating of a hypothesis as evidence is gathered, taking into account both the initial likelihood of such a hypothesis (P(H)), and the base rate probability of such evidence occurring (P(E)). This model has been used in statistics (Lee, 1989; Wagenmakers, 2007) and has been gaining popularity in areas of psychology, notably in the study of causal and active learning (Fernbach & Sloman, 2009; Fischhoff & Beyth-Marom, 1983; Holyoak & Cheng, 2011; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) and in the development of Bayesian Network Models (Pearl, 2000, 2014). It typically provides a normative account of what an optimal learner should do in a given environment.

$$P(\mathrm{H}|\mathrm{E}) = \frac{\mathrm{P}(\mathrm{E}|\mathrm{H}).P(\mathrm{H})}{P(\mathrm{E})} \qquad (3)$$

Accordingly, to indulge in a short thought experiment, one could apply the premise of an erroneous belief communication to a Bayesian learner. For the sake of simplicity, one could assume a positive prior regarding the belief ("Of two options – A or B, Option A is the best.") being true, based on various factors such as the trustworthiness of the source (e.g., P(H) = .6). But, if A is in fact the worse of the two options (i.e. it only wins 40% of the time, whilst option B wins the remaining 60%), then over time, the (lower) probability of supporting evidence, P(E) = .4, drives down the posterior (P(H|E)) to reflect the impact of evidence. As a consequence, the likelihood of the belief being true eventually drops below .5 and moves towards falsification (in this example, if P(H|E) – the probability of A being best given the evidence – drops below .5, then the anterior (P(¬H|E)) – the probability of B being best given the evidence, then goes above .5).

Given that the normative model above indicates that an erroneous belief should result in falsification when presented with enough evidence, but as we have indicated earlier, such beliefs are often believed and even prevalent in society, it is clear that humans deviate from this norm. The question remains as to where.

One possibility is that humans update optimally, but simply have not experienced enough evidence yet to reach the threshold for refutation. However, research has shown that learning optimally in such a way may be too computationally difficult (Gigerenzer & Goldstein, 1996; Hahn & Harris, 2014; Simon, 1955; Tversky & Kahneman, 1975). This sub-optimality first gained traction in psychology in the aforementioned bounded rationality proposed by Simon (1955), who put forward the notion that humans as learners and reasoners are burdened with biological and physical constraints. These limitations, whether physiological (thinking requires energy; humans can get tired) or psychological, result in a deficit when taken *in reference to a normative baseline*. As we have already touched upon, questions remain as to whether such a comparison is valid when discussing said limitations in terms of irrationality. However, the notion of systematic deviations from a normative baseline, or biases, provides insight to the question at hand, irrespective of conclusions regarding the rationality of such deviations.

Cognitive biases, and how they relate to the possible effects of communicated beliefs on the integration of evidence, are typically classified as such by the outcome: typically, choice or judgement deviations. However, it is best to distinguish these biases by the point in the learning process during which a biasing effect is incurred: information *input*, or evidence *interpretation* mechanisms (see Hahn & Harris, 2014, for a review).

2.2.1.1 Input (First Order)

In the case of first order biases, evidence available to the participant is skewed, resulting in a biased assessment of the environment. Importantly, such an outcome would occur even if the agent was perfect at interpreting and incorporating the evidence. A simple analogy for this would be a robot with an imperfect sensor, or placed facing down in a garden. The robot is asked to calculate the variety of vegetation and their proportions within the garden, and has the perfect mathematical equipment within it to make this calculation. But, if it has a short-sighted sensor, or is placed facing down, it will conclude that the garden only contains grass, seemingly ignoring the flower beds behind it. Irrespective of the reasons behind it, first order biases are thus defined by an asymmetry of input into the agent (MacDougal, 1906, but for a meta-analysis of motives behind confirmatory search, see Hart et al., 2009).

One of the most famous examples of an input asymmetry comes from work on rule induction. In seminal work by Peter Wason (1960), participants were given number triplets and asked to infer the underlying rule governing their relationship to one another (e.g., 2, 4, and 8, with a rule of "doubling"). Participants were then asked to generate "query-triplets" for the experimenter, from which participants were then informed whether the proposed triplet is correct (conform to the "true" underlying rule). Results showed that most participants created triplets that would confirm their currently suspected rule, rather than create triplets that would provide a falsification. As a consequence of this tendency, participants were rarely *exposed* to more informative outcomes, resulting in spurious and inefficient learning. This form of selective exposure is similar to work on positive test strategies (Klayman, 1988; Klayman & Ha, 1987; Matute, 1994; Navarro & Perfors, 2011), wherein participants seek out events and information which confirm the current hypothesis. For example, if an agent believes that taking a medicine will cure a disease, then when the disease is present the agent seeks to take the medicine. Given this choice of action, the agent is not exposed to the counterfactual outcome of whether the disease would have gotten better without the medicine, skewing the input of evidence in favour of confirmation (only instances when the cue – taking the medicine – is present are seen).

Such an asymmetrical exposure pattern has been demonstrated in explicit hypothesis testing (Doherty & Mynatt, 1990; Doherty, Mynatt, Tweney, & Schiavo, 1979; Fischhoff & Beyth-Marom, 1983), known as pseudo-diagnosticity. In this instance, the agent is particularly insensitive to the implications of the probability of the evidence occurring given an *alternative* (i.e. not focal) hypothesis (P(E|¬H)). Work in the illusions of control and causality have advocated the importance of attention to the null as critical to refuting harmful beliefs (Blanco, Barberia, & Matute, 2014; Matute, Yarritu, & Vadillo, 2011; Rudski, 2001; Yarritu, Matute, & Luque, 2015; Yarritu, Matute, & Vadillo, 2013a). The end result of these forms of bias, whether in a medical diagnosis (Doherty & Mynatt, 1990), or rule generation (Doherty et al., 1979; Wason, 1960, 1968) paradigms, is a sub-optimal (i.e. biased) way of updating that occurs as a consequence of selective attention, or filtering of inputs (MacDougall, 1906).

Similarly, work on information search paradigms (Choi et al., 2008; Hendrickson, Perfors, & Navarro, 2014; Jonas, Schulz-Hardt, Frey, & Thelen, 2001), task participants with deciding on a particular moral or social issue (e.g., the death penalty). When presented with various arguments for or against their own position, participants typically take arguments that support their own position at face value, whilst scrutinising (searching for additional information about) opposing arguments until invalidating evidence is found. The resulting asymmetry in evidence exposure, much like the above forms of filtering, leaves the agent with a false sense of reasoned consensus, i.e. the agent falsely believes the conclusion reached is a result of seeing all the necessary evidence.

Evidence filtering has also been shown in tasks during which participants must purchase information with the aim of coming to an accurate judgement. Participants tend to deviate from the amount of information they normatively “should” purchase, by chronically under-sampling; wherein agents reach a conclusion too quickly, purchasing less information than needed (Pitz, 1969). Of course, when variation within an environment is small, then the sample size required may be smaller, much in the same way as a statistical power calculation will yield a lower required sample size if the effect size is larger [clearer]. However, in such paradigms, participant choices can be thought of as tending towards under-powering their conclusions, leading to errors (Fiedler & Kareev, 2006; Hertwig, Barron, Weber, & Erev, 2004). This form of deviation relates to the notion of the law of small numbers (Benjamin & Raymond, 2012; Tversky & Kahneman, 1971), wherein the diagnosticity of small samples are overweighted. For example, having witnessed 5 flips of a coin, an individual infers that the coin is biased or loaded in some way, as heads comes up 4 times out of 5; however, were the individual to wait longer and witness more coin flips (such as 100 or more), then they would find that these misleading short-term deviations tend to average out towards the true mean (P(heads) = .5).

This brief overview touches upon the biasing effects of various forms of selective evidence exposure. Leaving aside arguments regarding how (mal)adaptive such selectivity is, the effects discussed have ready application to the maintenance of erroneous beliefs. To explain further, if one considers a communicated belief as a working hypothesis for the recipient, then the various forms of selectivity above provide a filtering mechanism through which a false belief could be sustained. The umbrella term, under which the first order biases described above have been categorised, is confirmation bias (Klayman, 1995; Nickerson, 1998), a category of biases that are generally concerned with the preservation of a hypothesis, whether through evidence selection, as discussed here, or through evaluation and interpretation, as addressed below.

2.2.1.2 Integration (Second Order)

To return to the robot in the garden analogy, for second order biases, the fault is assumed on the part of the mathematical processes within the robot. In this case the robot has adequate sensors and is placed so that it can see the entire garden, and thus does not receive skewed information. Instead, the calculation inside is imperfect (the programmer made a mistake, or had designed the mechanism for a different environment or question); resulting in an asymmetry in how the evidence is *weighted*. Such an asymmetry could arise, for example, in the form of weighting the early evidence it sees as more important than latter evidence known as primacy – the robot sees grass first and factors this into the calculation more heavily than the flowers it sees later. Alternatively, the robot might overweight evidence that supports a pre-conceived hypothesis; the robot starts with the assumption that the garden is entirely grass in its calculation, overweighting grass inputs, and underweighting flower inputs.

Research investigating these sub-optimal forms of evidence integration extends back over a hundred years, from MacDougal's (1906) work on what he termed "secondary biases" in judgement. He noted various judgement distortions in general knowledge 'rankings' questions (such as the lengths of rivers and populations of cities) as a consequence of availability or association (e.g., state capitals tend to have their populations overestimated, whilst lesser known towns in the same state have their populations chronically underestimated). This lead MacDougal to conclude that the way in which information is integrated into a judgment is prone to asymmetries and errors.

Over half a century later, integrative biases started to be explored in probabilistic evidence paradigms, known as "bag and poker chip" studies. These paradigms typically use two bags (or some form of opaque container), and within each container is a different proportion of red and blue poker chips, for example one container may have 60% blue poker chips, the other having 60% red poker chips. Participants are usually aware of what these two proportions are, but as the experimenter / task produces a new piece of evidence – an extracted red or blue poker chip (which is then returned to the bag) – the participant is required to estimate from which bag (and thus which distribution) the poker chips are being extracted from (e.g., Peterson & Miller, 1965).

An alternative variation has participants pick one of the two bags each trial, and on the basis of what colour the ball is, update their assessment of which bag contains which distribution (Slovic & Lichtenstein, 1971). Regardless of minor procedural variations, the strength of these paradigms is the capacity to compare the participant's *quantified* judgements to an optimal baseline for how the learner *should* update, as each new piece of evidence is experienced.

Conservatism bias, wherein the agent underweights new information received, resulting in insufficient changes in judgements over time, has been demonstrated using various forms of "bag and poker chip" style paradigms (Phillips & Edwards, 1966; Peterson & Beach, 1967). To elucidate further, as new information is received, the degree to which the participant updates their belief (whether positively or negatively) is less than the amount prescribed by a normative, or Bayesian, standard. Conservatism has been linked to under-sampling (as mentioned above) – for example through a common undervaluing of prospective evidence (Phillips & Edwards, 1966) – and the "inertia effect" (Pitz, 1969), the latter of which using a variation of the bag and poker chip design. These designs used a probabilistic reversal, wherein the distributions in the two bags were reversed at a certain point during the task, so that the overall outcomes (and thus the normative account of the belief concerning the two bags) were equal, i.e. the distribution difference in the second half, cancels out the inverse distribution of the first half (Peterson & DuCharme, 1967). Thus, although the learner should return to their starting point, the participant's beliefs instead fail to return to their point of origin.

This form of inertia in updating led to research into order effects in how evidence is weighted (Dennis & Ahn, 2001; Hogarth & Einhorn, 1992; Mantonakis, Rodero, Lesschaeve, & Hastie, 2009; Peterson & DuCharme, 1967; Pitz, 1969; Pitz, Downing, & Reinhold, 1967). Primacy, the overweighting of early evidence and underweighting of later evidence, has been found when participants are required to make on-line (or trial-by-trial) judgements in probabilistic learning (Peterson & DuCharme, 1967), among other sequential decision making paradigms (Curley, Young, Kingry, & Yates, 1988; Hogarth & Einhorn, 1992). Inversely, when making a judgement at the end of a sequence, recent evidence (i.e. evidence experienced towards the end of the sequence) plays an overweighted role relative to earlier evidence, known as a recency effect (Ayton & Fischer, 2004; Barron & Leider, 2010; Canic, 2014; Hogarth & Einhorn, 1992; Mantonakis et al., 2009). In either case, the deviation in weighting is a function of the point in a sequence, rather than providing equal weighting irrespective of order. Such tendencies indicate a susceptibility to misinterpretations of evidence that could result in erroneous belief maintenance. However, such secondary biases discussed so far do not provide a formalised link to the *confirmatory* aspect of evidence integration in belief preservation.

Concerning the role of a hypothesis in integrative biases, there are a plethora of interrelated phenomena that revolve around the concept of overweighting confirmatory evidence. The differences between these lines of research have involved both the reference point for confirmation, and the reasons behind it. For example, work in confirming expectancies (Billman, Bornstein, & Richards, 1992) has found that people tend to overweight evidence that confirms the prior expectancy regarding an outcome. Whilst work on optimism bias (Sharot & Garrett, 2016) has found positive evidence in relation to self-concepts is overweighted, even if this runs counter to expectancy, for example overweighting the significance of a surprisingly high test result (relative to a surprisingly low test result). In this way, positive self-concepts are reinforced, much in the same way that people seek to maintain belief-consistency (Lybbert, Barrett, McPeak, & Luseno, 2007), and thus differ from classic expectancy work by demonstrating the importance of valence in biasing effects. However, recent work has underlined the difficulties inherent to studying systematic biases, finding that studies taken as demonstrating such optimism biases can equally be explained by a statistical / methodological artefact inherent to the design of such studies (Shah, Harris, Bird, Catmur, & Hahn, 2013). Studies investigating social judgments (Pyszczynski & Greenberg, 1987), impression formation (Anderson, 1965), causal reasoning (Blanco et al., 2014; Garcia-retamero, Hoffrage, Müller, & Maldonado, 2010; Yarritu & Matute, 2015), one-shot argument evaluation (Lord, Ross, & Lepper, 1979), and special pleading for hypothesis incongruent information (Baron, 1995; Gilovich, 1991) have all shown a similar impact of integrating evidence in a manner that confirms the focal hypothesis or belief.

Such deviations have been shown to occur through a discounting, or underweighting of hypothesis-inconsistent information (Buchy et al, 2007; Woodward et al., 2007; as cited in Whitman et al 2015), as in information distortion literatures, where trailing hypotheses are weakened because of pre-decisional information (Nurek, Kostopoulou, & Hagmayer, 2014). Work using sequential gambling tasks (Gilovich, 1983), and reinforcement learning paradigms (Decker, Lourenco, Doll, & Hartley, 2015; Doll, Hutchison, & Frank, 2011; Doll et al., 2009; Staudinger & Büchel, 2013), have also investigated this skewed interpretation of evidence over time, through probabilistic forced-choice tasks. Such studies have similarly found overweighting of confirmatory instances.

Regarding possible cognitive mechanisms behind such effects, there are two, which are intertwined, of critical interest. The first of these is the availability of an alternative hypothesis (Navarro & Perfors, 2011; Nickerson, 1998). If it is difficult to generate an alternative, which can be especially problematic when evidence is ambiguous (Klayman, 1995), then there is only one hypothesis that is updated by evidence, with the only alternative being the null (that the hypothesis is untrue). This is problematic, as if the null is not a viable alternative (e.g., “I think action A leads to outcome E” goes to “I do not think action A leads to outcome E”) it becomes difficult to then establish any control over the environment (Curley et al., 1986; Ha & Hoch, 1989; Matute et al., 2011). Whereas, if the null is accompanied by a viable alternative (e.g., “I think action A leads to outcome E” goes to “I do not think action A leads to outcome E, because action B instead leads to outcome E”) then there is no such asymmetry in the prospective amount of control and ambiguity (and thus capacity for action), as a consequence of updating.

This leads naturally to the second mechanism implicated by integrative accounts of confirmation bias. Pattern matching (Shermer, 2008; Whitman et al., 2015) explanations of bias focus on a learning under uncertainty perspective of belief maintenance. The asymmetry inherent to the bias occurs as a result of an intrinsic asymmetry between pattern (i.e. belief or hypothesis) and evidence *matches* (confirmation), and pattern-evidence mismatches. A large body of neuroscience literature has shown a strong link between dopamine and learning, with the dopaminergic system critical to updating beliefs in light of evidence (Aggarwal, Hyland, & Wickens, 2012; Behrens, Woolrich, Walton, & Rushworth, 2007; Boureau & Dayan, 2011; Burke & Tobler, 2011; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Schultz, 2010), an effect that is impaired in Parkinson’s patients, who suffer from deficiency in this system (Peterson, Elliott, & Song, 2009; Swainson, Rogers, & Sahakian, 2000). When applied to the notion of evidence as either matching or mismatching a focal hypothesis, recent research has shown that underlying neural networks favour the acceptance and consolidation of a coherent pattern (evidence-hypothesis match, i.e. confirmation), whilst evidence that should lead to hypothesis rejection involves only the *lack* of a pattern (Whitman et al., 2015). Whitman and colleagues found significantly stronger neurological markers of updating in cases of acceptance, rather than rejection. To phrase this asymmetry another way; there is no corresponding neurological pattern to strengthen in the case of refuting information.

These integrative biases and the mechanisms underpinning them will be discussed in more detail in the theoretical framework section below. Before that, it is necessary to discuss the motivated reasoning account; an integral aspect of bias that has already been alluded to in work mentioned regarding valence (Sharot & Garrett, 2016), belief and self-concept preservation (Kusec et al., 2016; Lybbert et al., 2007; Yarritu, Matute, & Vadillo, 2014). An overview of confirmation bias that focuses solely on cognitive mechanisms misses a critical aspect of biases as a phenomenon; as will be shown below, rarely are there cases when motivations do not skew the values associated with selecting and integrating evidence in some manner.

2.2.2 Motivated Reasoning & Social Psychology Account

So far, we have looked at the cognitive account for biases as a consequence of communicated beliefs, and primarily confirmation bias as the central point of this process. However, such an account typically leaves the role of motivations and goals out of the reasoning process. Accordingly, to provide a more complete picture of *why* unsupported beliefs may be upheld, one needs to incorporate elements of motivated reasoning (Kunda, 1990). As has already been addressed, when determining the utility of a belief, one needs to understand the potential costs and benefits to the agent. In making this assessment, one must appreciate that there is a subjective element – one agent may treat a particular outcome as much costlier than another. These differences incorporate the motivations and goals of said agent. For example, an agent that has just eaten will have less of a need for food, and thus value it lower than a hungry agent. This applies to the same agent, just at different points in time. Alternatively, differences in the desire for social consensus could lead to orthogonal value functions between two individuals when it comes to deciding whether to cooperate with or compete against others. In this way, the utility function of an agent can both vary between agents, and even within an individual agent across time points. Consequently, when determining the potential costs and benefits of adhering to a belief, motivations, goals and needs can shift what would otherwise (based on a normative value function) be irrational or “costly” beliefs into the [subjectively] beneficial domain.

Work on motivated reasoning (Kunda, 1990) has focused on the relationship between an individual’s motivations and resultant behaviour, and has been explored within the field of Social Cognition. Although early work showed an impact of motivation on attitudes (Festinger, 1954), attributions (Heider, 1958) and even perception (Erdelyi, 1974), such work came under heavy criticism primarily on the grounds that effects purported to be a consequence of motivated reasoning, could simply be reinterpreted in cognitive terms (Nisbett & Ross, 1980), as a combination of prior beliefs and expectancies. However, in the 1980’s theorists proposed an interaction between cognition and motivation, whereby the former’s processes of integration and representation could be affected by the latter’s selection and use of said processes (Kruglanski & Freund, 1983; Kunda, 1987; Pyszczynski & Greenberg, 1987). In doing so, the foundation was laid for a motivated reasoning account of biases. To place this in terms of motivations altering value functions, there is a distinction between motivations that result in optimal process selection, whereby the value function derived by the motivation is in line with the normatively expected value function, as opposed to directional motivation which are in comparison “misaligned”, that lead to sub-optimality.

This distinction has been raised previously in discussions of motivated reasoning accounts of biases (Hahn & Harris, 2014; Kunda, 1990) and in particular regard to confirmation bias (Klayman, 1995; Nickerson, 1998). One can rephrase the distinction as a motivation to be *correct* (i.e. objectively accurate) and a motivation to be *right* (i.e.

reach the *desired* outcome, irrespective of any objective accuracy). As pointed out by Kruglanski and colleagues (Kruglanski, 1980; Kruglanski & Klar, 1987; Kruglanski & Ajzen, 1983), both these goals require a form of motivated reasoning, but this does not mean the two share the same mechanisms, and are therefore potentially distinct in more ways than just outcome category.

2.2.2.1 Accuracy Driven Reasoning

In determining how to act within an environment, an agent must expend some degree of effort in coming to a conclusion. As a consequence, there is a balance to be struck between the effort required (cost) and the benefits gained from an accurate (in the sense that it maximises the values of the agent, whether subjectively or objectively derived) assessment. This balance was formalised by Simon (1955) in his proposal of "satisficing"; reaching the desired threshold of assessment, for the lowest possible cost. Extending this principle, this effort-accuracy trade-off implies that it is not only the potential benefits of certain strategies that people are aware of, but that there is a consideration of the required cost (Beach & Mitchell, 1978; Stigler, 1961). Consequently, if an agent is motivated to be accurate, then they should be more willing to engage in costlier (in terms of both time and effort) strategies and reasoning processes.

Work on impression formation, which has found primacy effects whereby an initial judgement of an individual is overweighted relative to subsequent information in final judgements (Anderson, 1965; Freund, Kruglanski, & Shpitzajzen, 1985; Mann & Ferguson, 2015; Tetlock, 1983b), has been used to demonstrate the impact of accuracy motivations on this form of bias. By explaining to participants that they would have to publicly justify their judgements, or have their evaluations affect the target person's life, the motivation for accuracy was manipulated. Researchers found significantly fewer primacy effects, less use of ethnic stereotypes, and less anchoring in probability judgments (Freund et al., 1985; Kruglanski & Freund, 1983) in increased accuracy motivation conditions. Similar effects of accuracy motivations have been found in legal reasoning contexts (Tetlock, 1985b), subjective probability estimates (Pruitt & Hoge, 1965) and control deprivation manipulations have been shown to reduce fundamental attribution error (Pittman & D'Agostino, 1985).

The proposed mechanism through which accuracy motivations can reduce these biases and errors is by eliminating the mistakes that are incurred by *hasty* reasoning. Such a proposition has been supported by work on time pressures in reasoning processes, which has typically found exaggerated bias effects when under pressure (Kruglanski & Freund, 1983), as well as the requirement for at least moderate cognitive resources to be available for an incorrect implicit evaluation to be corrected (Mann & Ferguson, 2015). However, for accuracy motivations to have any impact on reducing these biases, manipulations must occur prior to exposure to evidence (Tetlock, 1983b, 1985a) to allow for deeper processing, and such deeper processing strategies must be readily available to the participant, be better suited to achieving accuracy, and be *known* to the participant as superior to other strategies (Kahneman & Tversky, 1972; Lord, Lepper, & Preston, 1984).

Finally, work on need for closure (Kruglanski & Ajzen, 1983; Webster & Kruglanski, 1994) has suggested that accuracy goals not only lead to the choice of more effortful, deeper strategies, but also play a role in the processes of hypothesis generation and consideration. In essence, the consideration period for the generation and deliberation over possible hypotheses is variable as a function of the motivation for accuracy (an orthogonal concept to the need for closure). This results in not only more hypotheses being brought into consideration, but the lengthier process also results in a greater sampling of evidence against which to validate hypotheses. In bringing both these mechanisms (strategy selection and evaluation periods) together, the motivation for accuracy thus influences not only the quantity, but quality of inferences (Kunda, 1990). Unfortunately, outside of experiments and the scientific method, seldom is it the case that accuracy motivations are unaccompanied by other, directional motivations.

2.2.2.2 Directional Reasoning

Work on self-concept preservation (Festinger, 1954), social conformity (Asch, 1955; Cialdini & Goldstein, 2004; Rodriguez, Bollen, & Ahn, 2015), and various studies of confirmation bias using participants' prior opinions (Allahverdyan & Galstyan, 2014; Lord et al., 1979; Pitz, 1969) all demonstrate biased reasoning to favour a *directional* goal. Such effects are not necessarily the consequence of unconstrained, conscious reasoning towards one's own goals, but can instead be a flawed attempt at rational evaluation of, and justification for, the outcome (Kunda, 1990). This "illusion of objectivity" (Pyszczynski & Greenberg, 1987) still requires the agent to gather enough evidence to support the desired conclusion (Darley & Gross, 1983), but simultaneously leads to an overconfidence in the objectivity of the outcome (Pronin, Gilovich, & Ross, 2004). Accordingly, it is not the absence of a reasoning process (which might be the case in the explicit conclusion of the desired outcome irrespective of any evidence), but rather the *selection* of particular reasoning processes / rules / beliefs / strategies and thus affect the information taken into consideration (whilst giving the sense of fair reasoning). Support for selective implementation has been found in the use of statistical rules (Kunda & Nisbett, 1986), social judgements (Higgins & King, 1981), confirmatory search (Hart et al., 2009), and attitude endorsement (Snyder, 1982). Further, the work on accuracy motivations has indicated that directional motivations play a role in both the selection and duration (depth) of reasoning (Kruglanski, 1980; Kruglanski & Ajzen, 1983), whilst self-esteem models (Pyszczynski & Greenberg, 1987) suggest directional motives can influence all stages of hypothesis generation and evaluation, including biasing the retrieval of information from memory (Frost et al., 2015).

Research investigating dissonance effects, wherein counter-attitudinal behaviour leads to subsequent changes in attitudes (Festinger, 1962; Harmon-Jones, Peterson, & Vaughn, 2003), provides some insight into the mechanisms behind the impact of directional motivations on reasoning. Dissonance is an arousal caused by a threat to self-image, such as engaging in actions that are known to have negative consequences (Cooper & Fazio, 1984), or more generally acting against one's own positive self-image (Aronson, 1968; Greenwald & Ronis, 1978). Accordingly, to resolve this state of dissonance, the agent is motivated to disconfirm the negative view of themselves as incongruent, for example by modifying their attitude so that it is congruent with the preceding behaviour, i.e. a directional goal. For example, studies in which participants are required to endorse an attitude they themselves do not originally condone, such as limiting free speech (Linder, Cooper, & Jones, 1967) or police brutality (Greenbaum & Zemach, 1972), or having to re-describe a "boring" task as interesting (Festinger & Carlsmith, 1959), show consequent attitude shifts away from their initial (now dissonant) position. This coherence-based motivation has been recently reformulated in terms of a Meaning Motivated Model (Heine, Proulx, & Vohs, 2006; Proulx & Inzlicht, 2012), which argues meaning-reaffirming perceptions or behaviours follow from perceived psychological "threats" (typically to self, such as self-esteem, feelings of uncertainty, rejection, or mortality). However, such changes in beliefs or attitudes are constrained by prior beliefs; attitudes can only change so far. This constrained change does lend support to the notion of a biased memory search, wherein without the directional motivation, minimal change would occur, but it is not an unrestrained redressal of attitude, the change must fit the existing knowledge constructs (Kunda, 1990).

Similar biased memory search effects have been found for the evaluation of the abilities of others (Klein & Kunda, 1989), wherein a participant that finds out they are either dependent upon, or in competition with, a target individual results in the participant rating the targets abilities as higher (after a perfect test score in either condition) in the former case. Beyond biases in memory search, such biasing effects of directional motivations have been found in the interpretation of athletic events (Gilovich, 1983), medical conditions (Ditto, Jemmott, & Darley, 1988) and even judgements of event likelihoods (Arrowood & Ross, 1966; Irwin, 1953; Pruitt & Hoge, 1965). Directional goals have also been found to influence the evaluation of arguments (Kassajian & Cohen, 1965; Kunda, 1987; Lord et al., 1979). In such cases the presence of the directional motivation precedes exposure to evidence, but finds, for example, decreased ratings of validity for and greater scrutiny of disconfirming evidence, even holding prior beliefs constant (Kunda, 1987).

This latter point regarding prior beliefs raises the difficult issue of how one can disentangle motivation from typically cognitively interpreted features such as prior beliefs. Further, such an issue becomes even more problematic when directional and accuracy motivations are not investigated in isolation, with little research done on the possible competition between the two. In principle, an accuracy motivation should push more towards optimal, but “imperfect” cognitive mechanisms mean bias is still a possible by-product (Hahn & Harris, 2014). In the case of directional motivations, deviation is intrinsic to the goal at hand. It is important to remember though, that the continuum or distinction between these two is typically undetectable to the agent themselves. In the theoretical framework that follows, the purpose is to delineate between the impact of these different motivations (directional and accuracy) on subsequent biases, and the impact of cognitive processes in and of themselves, and in conjunction with motivations.

2.3 Theoretical Framework

Before focusing on the proposed mechanisms behind the biasing effects of communicated beliefs, it is worth first synthesising what has been covered regarding cognitive and motivational components.

Firstly, there is an important distinction to be made between first (input) and second (integrative) order cognitive biases (Hahn & Harris, 2014; MacDougall, 1906). The former of these can provide an explanation for deviations that are a direct consequence of environments in which exposure is outside of the individual's control (Matute et al., 2011). However, when the individual does have control over evidence exposure, several theorists have posited a hierarchical causal relationship between second and first order biases (Klayman, 1995; MacDougall, 1906; Nickerson, 1998). To elucidate further, the biased selection of evidence is a natural consequence of overweighting confirmatory evidence and/or underweighting contrary evidence. As such, second order mechanisms of confirmation biases, such as hypothesis restriction (Klayman, 1988, 1995; Nickerson, 1998) and asymmetry in pattern-updating (Doll et al., 2011; Shermer, 2008; Whitman et al., 2015), result in a value function for evidence that favours confirmation, which in turn results in search strategies that favour this preference. Furthermore, selective exposure does not rule out integrative accounts, as although there is a correlation between selective exposure and confirmatory instances, it is rarely a perfect correlation. As such, non-confirmatory instances will still be encountered despite selectivity, and thus integration of such instances will still be required.

Secondly, motivations play a role in the selection of these cognitive processes (both second and first order), as well as the duration and degree of hypothesis generation and evaluation. Motivations for accuracy, which have been associated with greater task engagement (Tetlock, 1983a, 1985b), selection of more suitable processing strategies (Lord et al., 1984), greater numbers of alternative hypotheses considered (Kruglanski & Ajzen, 1983; Webster & Kruglanski, 1994) and longer deliberation periods for the evaluation of said hypotheses, may lead to improved accuracy and performance in experiment tasks relative to an un-motivated baseline (Kunda, 1990). However, such improvements are generally seen as resolving instances of biases that are a consequence of hasty reasoning, rather than a universal solution to bias (Kunda, 1990). To put this another way, although motivations for accuracy can improve performance by aligning the value function for confirmatory versus dis-confirmatory evidence closer to a normative standard, this cannot absolve the agent of biases resulting from more fundamental sub-optimal cognitive processes.

Conversely, directional motivations, *ceterus paribus*, lead to greater use of sub-optimal (sub-optimal as defined by an optimal learner with a "normal" value function) cognitive processes, and consequently greater likelihood of biased conclusions (Hahn & Harris, 2014; Kunda, 1990). The use of such cognitive mechanisms, although leading to greater degrees of bias in the sense of an "optimal / normative" learner comparison, may instead be well suited to favour the value function *given the agent's motivation*. It is accordingly important to once more reiterate the difficulties in classifying such behaviours as "irrational". On one end of the spectrum, when intention is to be *correct* (i.e. accurate), then deviation from optimal more readily fits an "irrational" definition, in the sense of a failure to maximise value. However, there are other constraints that the agent must account for, such as time and effort, which complicate the classification of irrationality for any deviations from maximisation, given satisficing constraints (Simon, 1955). Finally, when dealing with directional motivations, which have a value function that prioritises being *right* (e.g., internally consistent) rather than being accurate to a normative standard, then defining subsequent deviations in accuracy as irrational is wrong-headed (whether the possession of such a directional motivation is irrational is another matter).

So far, we have covered propositional learning as a broad framework for the communication of beliefs in a general sense, as well as the limited nature of the study of fallacious belief adherence within this domain (notably the under-studied consequences of *communicable* false beliefs within a cognitive or social psychological framework). Finally, we have covered some of the general cognitive and motivational processes behind biased reasoning, which are argued to play a role in such belief adherence. Accordingly, these elements are unified into the theoretical framework that underpins this thesis:

Premise 1. A communicated belief is a pattern / rule / hypothesis for an agent to act within an environment.

A communicated belief, in essence, acts as a rule, pattern or hypothesis for acting within an environment. Such beliefs, for example in the case of superstitions, are a verbal description of an action-outcome, and as such rarely contain quantified elements (Gilovich, 1993). As far as the recipient of the belief is concerned, beyond the directional pattern or hypothesis laid out by the belief, there are no further specifications, leaving interpretation somewhat subjective – especially in ambiguous evidence environments (Chambers, Graham, & Turner, 2008; Cimpian, Brandone, & Gelman, 2010; Leslie, Khemlani, & Glucksberg, 2011). Such a distinction of using verbal statements indicating only a rule or hypothesis provides a closer external validity to the phenomenon under scrutiny, moving this work away from the more traditionally studied, quantified "advice" in judgement and decision making literatures (Harvey & Fischer, 1997; Newell & Rakow, 2007).

Such verbal statements can be termed “generic”, in that they express an unquantified generalisation about “the way things are” (Prasada, 2000). Work in linguistics, cognitive psychology, and philosophy has previously looked at generic in comparison to statistical (quantified, or numerical claim based) beliefs, typically in terms of categorization (Leslie, 2008; Prasada, Khemlani, Leslie, & Glucksberg, 2013). Generic statements have been shown to resist falsification more so than their statistical equivalents (Chambers et al., 2008; Gilovich, 1993), and require little evidence for uptake and maintenance (Brandone, Gelman, & Hedglen, 2015; Cimpian et al., 2010). It is this “generic statement” form that is applied in the current work to the conceptualisation of beliefs.

It is this belief content, along with source factors, that together play a role in evaluation of the strength of a belief (Hahn, Harris, & Corner, 2009). Initially, we seek to hold these two factors (content and source) constant, to focus on the role of evidence factors (such as order and ambiguity) in belief uptake and maintenance.

Premise 2. In cases where the source of the belief is unqualified, early evidence acts as a “gatekeeper” in the uptake of a belief, acting to validate both belief and source.

Following from premise 1, work in positive test strategies (Klayman & Ha, 1987; Navarro & Perfors, 2011; Nickerson, 1998), pseudo-diagnosticity (Addario & Macchi, 2012; Doherty et al., 1979), rule discovery (Wason, 1960, 1968) and illusion of causality (Yarritu & Matute, 2015) indicates the presence of such a belief (or hypothesis) may result in confirmatory search strategies. In other words, upon receiving a hypothesis, the individual seeks to confirm it. An additional wrinkle contained within this premise, is the added factors that are incorporated into reasoning when a hypothesis is transferred second-hand, rather than self-generated. There are two key elements to this. Firstly, the verbalisation of a hypothesis requires interpretation on the part of the receiver, and is thus susceptible to noise. Secondly, the source of the belief contains cues to the possible validity (or persuasive capability) of the transmission (Briñol & Petty, 2009). Factors such as the perceived trustworthiness (Siegrist, Cvetkovich, & Roth, 2000; Siegrist et al., 2005; Twyman et al., 2008), expertise (Goodwin, 2011; Sniezek & Van Swol, 2001), similarity to recipient (Petty & Cacioppo, 1979), and attractiveness of the source (Kelman, 1958) can all critically impact the degree of confidence in the belief being transmitted (Briñol & Petty, 2009). However, when such cues are not readily obtainable (i.e. the source is "unqualified") then the agent is in a position of uncertainty regarding the belief's validity (Priester & Petty, 1995).

There are two implications that can be drawn from this, the first being that in the case of a belief unaccompanied by cues to the source's credibility, initial evidence acts not only in confirming the belief itself, but also the credibility of the source. The second implication is that when a belief is communicated by a source that is credible or trustworthy in the eyes of the recipient, for example if the belief comes from an authority figure, then the impact of initial evidence is reduced, as the recipient already has confidence in the belief's validity.

Premise 3. "Confirmed" beliefs result in the biased integration of evidence.

Once uncertainty regarding the belief (as a consequence of an unknown source) has been resolved via initial evidence based confirmation, the subsequent integration of evidence is biased in favour of confirming it. This claim extends from work in confirmation bias on self-generated hypotheses, and their consequential overweighting of confirmatory instances (Gilovich, 1983; Klayman, 1984; Klayman, 1995) and underweighting of non-confirmatory events (Nurek et al., 2014). Further, recent work in reinforcement learning (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013) and in neuroimaging (Whitman et al., 2015) have found hypothesis-consistent information is overweighted relative to hypothesis-inconsistent information. Such a process is explained by Whitman and colleagues as a natural reflection of the asymmetry between evidence that fits an established set of connections, or pattern (hypothesis), which results in stronger learning signals, relative to evidence which does not fit any pattern (i.e. the null; therefore, having no established connections to update). At this point, we have left motivations and non-integrative (i.e. first order / selective input) cognitive process out of the model in question, but this is not to waiver their likely influence, and so it is these we incorporate next.

Premise 4. Stripping away the capacity for first order biasing mechanisms, such as selective exposure, does not eliminate bias. However, where such strategies can be employed, biasing effects may be amplified.

Although we propose that the confirmation bias effect of a communicated belief in an uncertain environment is, at its heart, a result of a fundamental overweighting of confirmatory evidence, as we have mentioned previously, first order biasing effects may also play a role. This means that when first order biasing mechanisms, such as positive test strategies (Klayman, 1995; Klayman & Ha, 1987; Navarro & Perfors, 2011), are not possible, then confirmation bias outcomes may still occur. To assess this methodologically, it is necessary to include counterfactuals in any forced-choice based paradigms. Conversely, if counterfactual feedback is not included then it becomes difficult to discern whether any subsequent bias is due to biased selection (first order) or integration (second order). However, when it is possible for an agent to be selective or otherwise engage in first-order bias strategies, then such strategies are, based on research on confirmatory search (Jonas, Schulz-hardt, Frey, & Thelen, 2001; Klayman & Ha, 1987), likely to be employed, and will consequently result in higher levels of bias (Klayman, 1995; MacDougall, 1906; Nickerson, 1998).

Premise 5. Directional motivations, when present, should exacerbate the use of sub-optimal strategies and lead to greater bias. Inversely, when bias is in part due to hasty reasoning, accuracy motivations should reduce biasing effects through increased engagement and deliberation, along with selection of more optimal cognitive strategies.

In the discussion of the motivational account for biased reasoning, two pivotal motivations were described; the motivation for accuracy, which typically results in greater and more prolonged engagement with evidence and the generation and evaluation of hypotheses (Freund et al., 1985; Kruglanski & Freund, 1983; Kruglanski & Ajzen, 1983), and directional motivations, which value a *desired* outcome, rather than the *correct* (i.e. accurate to an objective standard) outcome (Kunda, 1990). Synthesising these two motivations into the model laid out so far, they are likely to have orthogonal impacts on the bias process. The motivation for accuracy, for example, should reduce biasing effects – assuming a degree of bias is due to hasty reasoning processes. Given the premise of a communicated belief transferred to an individual, accuracy motivations should result in more effortful evaluation of the belief's validity as evidence is gathered (Tetlock, 1983a, 1983b). This process, in the case of an erroneous or unsupported belief, should lead to lower levels of belief adherence.

Conversely, directional motivations, such as the communicated belief also being related to the agent's self-esteem or self-concept (Festinger, 1954; Pelham & Swann, 1989), or coming from a source that provokes additional motivations (e.g., a close friend vs. a stranger; Frost et al., 2015), results in increased use of sub-optimal cognitive processes (Kunda, 1990). For example, if evidence is seen that runs against confirmation, and accompanying directional motivations result in an additionally imposed value for said confirmation, then contradictory evidence is more likely to be dismissed as anomalous (Klayman, 1995; Kunda, 1990).

Accordingly, the purpose of the thesis is to address the validity of the premises set out above. To do so it is first necessary to try to excise the additional layer of conflicting motivations, as when both cognitive and motivational elements are involved, ascertaining which is critical to biases that allow for unsupported communicated belief adherence becomes difficult. The following empirical chapter seeks to engage with these initial premises, before subsequent empirical chapters then build upon this work to both re-affirm the initial premises and extend to encompass latter aspects of the model.

CHAPTER 3: COMMUNICATED BELIEFS ABOUT ACTION-OUTCOMES: THE ADOPTION AND MAINTENANCE OF UNSUPPORTED BELIEFS

Human beings, like the majority of animals, have the capacity to learn how to interact with an environment through first-hand experience of action-outcome relationships. Although some animals have developed the limited ability to communicate these relationships, such as primates, dolphins and bees (Bradbury & Vehrencamp, 1998; Frisch, 1950), humans have taken this ability to much higher levels. This transfer of knowledge can be highly adaptive – we can for instance be informed that having a coffee will cause us to feel more awake, and from this information choose to have a coffee to realize this outcome, without having to start from scratch in working out what might reduce our tiredness. Hence, the development of language has allowed us to transfer information about action-outcomes with an unparalleled capacity and flexibility.

However, this communicative capacity can result in some beliefs being passed on that are not supported by evidence, whether due to misinterpretations or perceptions of evidence in the communicator, or wilful deception. This combination of erroneous or unsupported beliefs, and the capacity to transfer (a capacity that is ever-increasing with the development of technology, from the printing press to most recently the internet) creates dangerous, viral effects (Lewandowsky et al., 2012), such as the dis-proven belief that vaccines cause autism[6]. Such phenomena provoke an obvious and critical question; why are such fallacious beliefs adopted and maintained?

[6] According to 2013 US polling data, 20% of Americans sampled believe vaccines cause autism. Sourced from: http://www.publicpolicypolling.com/pdf/2011/PPP_Release_National_ConspiracyTheories_040213.pdf

In the present chapter, we seek to answer this question. We demonstrate that evidence order and ambiguity are critical in the evaluation (and maintenance) of a communicated belief. Specifically, we demonstrate that if a belief regarding an action-outcome with an unclear truth value (such as an anonymous tip) is communicated to a naïve participant, early experiences dictate whether such beliefs are then adopted. Given this adoption, “believers” then go on to maintain such a belief, despite repeated evidence indicating such a belief is erroneous, demonstrating a confirmation bias. The reliance on early experiences is found given the *absence* of source cues such as perceived expertise (Goodwin, 2011; Harris et al., 2015; Walton, 1997) and trustworthiness (Briggs, Burford, De Angeli, & Lynch, 2002; Metzger & Flanagin, 2013; Schul & Peri, 2015; Siegrist, Gutscher, & Earle, 2005; Sniezek & Van Swol, 2001; Twyman, Harvey, & Harries, 2008). Consequently, the demonstration in the present work of an unsupported communicated belief’s long-lasting biasing impact on learning processes as a function of evidence order *alone*, has particular relevance to literature on persuasion (Briñol & Petty, 2009; Petty & Cacioppo, 1984; Wood, 2000), impression formation (Anderson, 1965; Kruglanski & Freund, 1983; Mann & Ferguson, 2015; Neuberg & Fiske, 1987), and the investigation of instruction effects (Doll et al., 2009; Mertens & De Houwer, 2016; Roswarski & Proctor, 2003; Van Dessel et al., 2015; Van Dessel, Gawronski, et al., 2016). Importantly, in maintaining such beliefs, previous work investigating confirmation bias has suggested motivated reasoning (Kunda, 1990) and skewed / selective evidence exposure (Doherty et al., 1979; Jonas et al., 2001; Wason, 1960) explanations. We instead obtain these results when such explanations are *absent*. As a result, we propose a confirmation bias in *integration* as the more fundamental mechanism behind erroneous belief acquisition and maintenance.

3.1.1 Truth Values

Before delineating the proposed mechanisms behind such belief-uptake processes, it is first worth providing a theoretical context to beliefs and the route towards evaluation. While information about action-outcome relations has been widely regarded to be represented in terms of associations (Hommel, Müsseler, Aschersleben, & Prinz, 2002), it has been pointed out that this does not necessarily mean that these representations are formed by slow associative processes (i.e., Hebbian learning) that are, for instance, thought to underlie habit formation (Custers & Aarts, 2010). They can also result through propositional processes (Mitchell et al., 2009). These allow for fast and flexible changes in associations as these propositions are hypotheses about the state of the world that have a "truth value" and can therefore be confirmed or disconfirmed. Hence, while people may form action-outcome representations slowly through repeated experiences, they may also evaluate the truth value of communicated beliefs about these relations by others.

In the context of the present chapter, beliefs are defined as generic statements (Cimpian et al., 2010), as a relational hypothesis, akin to an action-outcome association. These statements lack quantified parameters (Leslie, 2008), and as such are typically more difficult to falsify (Cimpian et al., 2010; Leslie et al., 2011). For example, when comparing “Drinking snake-oil will cure your ailment” to “Drinking 250ml of this snake-oil will cure your ailment within 3 hours.” the former benefits from the added flexibility of interpretation. Both versions do however imply the same action-outcome, but it is the former that bears a more appropriate real world parallel to natural communications of (erroneous) beliefs (Gilovich, 1993).

3.1.2 Communication and Confirmation Bias

Normative accounts suggest a communicated belief and evidence should interact so that the belief gradually updates as more evidence is experienced. This updating should lead to a final belief that matches the evidence (Fischhoff & Beyth-Marom, 1983). However, failures to integrate evidence objectively are not uncommon in the psychological literature. Psychological research into cognitive biases has instead shown systematic misinterpretations of evidence (Bar-Eli, Avugos, & Raab, 2006; Gilovich, 1983; Gilovich, Vallone, & Tversky, 1985; Tversky & Kahneman, 1971), and failures to adjust beliefs accurately (Abbott & Sherratt, 2011; Dave & Wolfe, 2003; Dennis & Ahn, 2001; Rozin, Millman, & Nemeroff, 1986; Tversky & Kahneman, 1973) across many domains of learning (Pohl, 2004).

One explanation for the persistence of communicated beliefs it that people fall prey to the aforementioned confirmation bias (Klayman, 1995; Nickerson, 1998), wherein (in this case) belief-congruent evidence is overweighted. However, what complicates delineating possible mechanisms difficult is the fact that communication of belief rarely occurs in a vacuum. Work in persuasion, argumentation, and risk communication literature has demonstrated that when determining the strength of communicated arguments (aimed to provoke, for example, a change in attitude), individuals will incorporate cues associated with the source (Briñol & Petty, 2009; Hahn et al., 2009), including perceived source trustworthiness (Siegrist et al., 2005; Twyman et al., 2008), attractiveness (Kelman, 1958), and expertise (Goodwin, 2011; Sniezek & Van Swol, 2001). These factors invoke a myriad of both cognitive and motivational reasons for maintaining an erroneous belief.

We now briefly highlight some of these (at times competing) motivational and cognitive explanations, with a view to demonstrating the importance of assessing the impact of beliefs in the absence of such motivations and cognitive strategies. In doing so, we forward an account of confirmation bias in (erroneous) belief maintenance that is at its heart a consequence of an asymmetry in the way evidence is *integrated*. This integrative bias occurs irrespective of directional motivation (Kunda, 1990) or skewed evidence exposure (Klayman & Ha, 1987; Nickerson, 1998) explanations commonly associated with erroneous belief acquisition. Such effects are instead shown to be dependent upon evidence order in the immediate attempted validation of both belief and (by proxy) source.

3.1.2.1 Motivational Explanations

Motivations behind confirmation bias vary from social conformity (Asch, 1955; Cialdini & Goldstein, 2004) to cognitive dissonance (Festinger, 1962). For example, when asked to evaluate the effectiveness of arguments either in favour of, or opposed to the death penalty (Lord et al., 1979), participants pre-existing political, ethical, and social motivations behind their particular opinion, led to more positive evaluations of arguments that favoured their prior opinion. This was taken as evidence that people are motivated to uphold their personal beliefs when evaluating arguments, and can be thought of as *directional* motivations (Kunda, 1990).

When focusing on the effects of communicated beliefs regarding action-outcome relationships, many of these directional motivations contribute to the confirmation bias effect (Klayman, 1995; Pyszczynski & Greenberg, 1987) in a complex fashion that raises problems for an experimental setting. That is, a communicated belief (e.g., a homeopathic medicine works) may bias evidence integration because it interacts with other needs (such as self-preservation). In other words, the resulting confirmation bias may not directly reflect the communicated belief, but be motivated by the individual’s associated needs. Although it is difficult to remove all elements of motivated reasoning from real world situations (Yarritu et al., 2013), the question of whether merely hearing about a belief is enough to bias evidence integration, has important ramifications for our understanding of when such biases would occur.

3.1.2.2 Cognitive Explanations

How could a communicated belief lead to confirmation bias effects even in the absence of these motivations? The removal of directional motivations can help clarify the remaining mechanisms at the heart of belief biasing effects. Such a removal has been posited, through work investigating the interaction between motivated reasoning and cognitive processes (Hart et al., 2009; Kunda, 1990), to result in less use of sub-optimal cognitive processes, which might otherwise be selectively employed to favour the motivated outcome. These (biasing) processes can be divided into two camps, first order (or *input* based) and second order (or *integration* based) accounts (MacDougall, 1906).

First Order. As an individual learns action-outcomes from experiences, if the evidence *seen* favours confirmation (whether through purposeful strategy, or a naturally skewed environment), any resultant bias could be in part (or entirely) due to this asymmetry in evidence exposure. In other words, if selective information intake is possible within an environment, one cannot discern whether the biasing effect of a communicated belief is due to an asymmetry in the *valuation* of confirmatory evidence over contradictory (Klayman, 1995), or due to the asymmetrical *exposure* to confirmatory evidence (or a combination of the two).

Selective intake of evidence can fall into two broad categories. The first of these, positive test strategy (Doherty et al., 1979; Klayman & Ha, 1987), involves an asymmetry in *choices* – selecting events that are likely to confirm the hypothesis, avoiding exposure to evidence that might refute it. For example, if someone holds the belief that taking a homeopathic medicine will cure an illness, when afflicted with that illness, the person will *make the choice* to take the medicine. As such they are not exposed to the outcome that the disease would have cleared up without the medicine – the reason for including a "no treatment" group in clinical trials. The second form of selective information intake mechanism results in an asymmetrical exposure to evidence through selective *search* (Lord et al., 1979; Nickerson, 1998). Commonly associated with work on opinion rather than action-outcome beliefs, selective search typically stems from an asymmetry in how arguments are evaluated. Arguments that confirm what the person already holds to be true (e.g., that homeopathy is effective) will be taken at face value, whilst arguments that contradict the belief are scrutinised in much more detail, until some evidence is found that discredits the argument (such as seeking out a possible conflict of interest in a paper refuting homeopathy, whilst not affording the same scrutiny to an article supporting it).

Both of these forms of selective information intake have the same consequence: an asymmetry in evidence exposure that favours confirmation, resulting in a biased interpretation of the environment. Accordingly, lines of research investigating the damaging effects of biases often recommend remedial interventions that draw attention to this asymmetry (Blanco et al., 2014) as a way of removing the consequent bias, based on the premise that this selective exposure is the key mechanism behind it. As such, the question of whether participants will still show a bias when such a mechanism cannot take place has important implications for the role of communicated action-outcome beliefs leading to biases through the skewed *integration* of evidence alone.

Second Order. Seen by theorists in confirmation bias as hierarchically responsible for selective strategy use (Klayman, 1995; MacDougall, 1906), the integrative account of confirmation bias posits that confirmatory evidence is *valued* asymmetrically over contradictory evidence. Put another way, despite both confirmatory and contradictory evidence being integrated, the former is systematically overweighted relative to the latter. Evidence for the overweighting of confirmatory and underweighting of contradictory evidence has been found in work on confirmation bias (Gilovich, 1983; Klayman, 1984; Klayman, 1995) and information distortion (Nurek et al., 2014). Such work has found support from reinforcement learning (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013) and neuroimaging studies (Whitman et al., 2015). The latter of which has demonstrated such an integrative bias is a natural consequence of the asymmetry between the updating signal that occurs when evidence matches an established set of neural connections (taken as the representation of a pattern or hypothesis) versus evidence that does not fit the pattern. The smaller updating signal of this mismatched evidence has no equivalent "null" pattern to update in kind.

This work fits with the demonstrated impact of focal hypotheses on the restriction of considered alternatives (Klayman, 1995), otherwise known as a restriction of the hypothesis space (Fischhoff & Beyth-Marom, 1983; Pyszczynski & Greenberg, 1987). In problem solving this has been termed the Einstellung effect (Bilalić, McLeod, & Gobet, 2008; Luchins, 1942), where holding a hypothesis hampers the capacity to find alternative solutions. In this way, the presence of the focal hypothesis not only provides the pattern upon which updating may occur, but at the same time restricts the availability of alternative patterns for updating.

However, such a process in the case of self-generated hypotheses, suggests an initial critical period of sensitivity when the hypothesis is formed / selected. This leads to primacy effects, as individuals are initially more sensitive to evidence as a hypothesis is formed / selected. Evidence for this been found in judgements of causal strength (Dennis & Ahn, 2001; Fugelsang & Thompson, 2003) and information distortion (Blanchard, Carlson, & Meloy, 2014; DeKay, Miller, Schley, & Erford, 2014; Nurek et al., 2014).

As the research discussed above only pertains to self-generated hypotheses, the mechanism is forwarded with a minor adaptation to account for the case of communicated beliefs. The key difference being that instead of the hypothesis originating from first-hand experience, it is instead introduced to the actor by a third-party. As we have mentioned previously, when evaluating such a communication, there is an implied relationship to the credibility of its source (Hahn et al., 2009), in that cues indicating the reliability of a source impact the perceived validity of the communication. However, when a belief is communicated in the absence of source cues, then we posit that early experiences, previously demonstrated to be pivotal in hypothesis *formation* processes (Anderson, 1965; Dennis & Ahn, 2001), are instead required to validate both the belief, and by inference, its source (Bovens & Hartmann, 2003).

Further, by removing source cues such as affiliation with the source (Frost et al., 2015), and the perceived expertise (Goodwin, 2011; Harris et al., 2015; Walton, 1997), then the resulting motivational and cognitive explanations for confirmation bias (e.g. the belief was communicated by a friend, whom one is motivated to agree with) are excised. Such an excision is necessary to focus on an integrative confirmation bias account of erroneous belief maintenance, distancing this work from literatures in which source cues are themselves taken as evidence with a truth value (i.e. a cue indicating trustworthiness not only impacts the evaluation of the truth value of the communication, but in itself possesses a truth value of either being true, or not). Attitude change, argumentation, and persuasion literatures, which have looked at the impact of this form of evidence (Briñol & Petty, 2009; Hahn et al., 2009; Harris et al., 2015; Petty & Cacioppo, 1984; Priester & Petty, 1995) indicate the efficacy of an argument (or belief) as dependent upon other truth value assessments of "evidence" (i.e. source cues), which interact with an individual's priors (e.g., a prior attitude or opinion).

The present work, by using a novel environment and probabilistic evidence (which in itself carries no truth value), allows for not only the investigation of how *learning over time* is impacted by a prior belief, but restricts possible explanations of deviation to how evidence is integrated to validate, and then maintain, the singular belief / hypothesis in question. As a consequence, once such a validation occurs, we posit that a communicated belief will similarly lead to the perception of evidence as being either supporting or contrary to it (Abbott & Sherratt, 2011), with overweighting of the former / underweighting of the latter resulting in an integrative confirmation bias that maintains the communicated belief.

3.1.3 Communicated Beliefs as Advice

The Judgement and Decision Making literature has focused on the roles of communicated (termed ’description’) and experienced evidence as advice taking (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997). Typically, these paradigms use binary choices between risky (probability of high reward or nothing) and safe (guaranteed low reward) gambles, in which participants must either rely on their own experience or on descriptions of the choices provided by the experimenter (Ludvig & Spetch, 2011; Newell & Rakow, 2007; Rakow & Newell, 2010), investigating competition between the two forms of information.

However, as opposed to action-outcome beliefs, and the present research consequent use of generic, unquantified statements (Cimpian et al., 2010; Gilovich, 1993; Leslie, 2008), advice taking research typically communicates probabilistic information (e.g., "60% chance to win $2") to the participant with every trial of experienced evidence. However, as the current research investigates the role of a communicated belief as a starting point (initial hypothesis) for possible biases in subsequent evidence integration, repeated exposure to the communicated information is not appropriate. Further, given the cognitive explanation put forward in this work, re-exposure to the belief may lead participants to overweight it (if people keep seeing the same communicated information from the experimenter, they may infer a greater value to this information, than if they only saw information initially), which tampers with the way in which a belief might otherwise be updated (Epley & Gilovich, 2001, 2006).

From a motivational standpoint, the fact that in advice taking the participant is provided with a description from the experimenter (for a notable exception, see Yaniv, 2004), motivationally may bias towards description, as participants may be subject to experimenter demand effects (Hertwig & Ortmann, 2008). To place this in terms of the aforementioned work on source credibility, the knowledge that the information comes from the experimenter, through the perceived expertise (and trustworthiness) of such a source, impacts the perceived validity of the information (Briñol & Petty, 2009; Hahn et al., 2009) These issues reflect the potential impact of social context even when focusing on the more purely cognitive aspect of advice taking typical in the literature (Collins et al., 2011). An important exception to this tendency is work by Collins and colleagues (2011) that focused the role of social influence (in this case the inclusion of non-social cues regarding advice sources, such as their political orientation) within an attentional cue learning paradigm typical to advice taking. Their findings that when non-social cues and advice are encountered together (a more realistic learning scenario in daily life), blocking – the prevention of an additional cue being associated with an outcome through repeated occurrence if the outcome is already associated with a similarly correlated cue – does not occur. The inclusion of both forms of information, which extant cognitive models consider redundant, instead lead to different patterns of learning. Although the focus of Collins et al.'s research was corroborative information across sources, it nevertheless neatly demonstrates the fruitfulness of research that incorporates both social and cognitive psychological theory.

Taking the above issues on board, it appears necessary that adaptations are required for a method that assesses the fundamental effects of a communicated belief on the integration of evidence, stripping away the noise of additional motivations and selective evidence exposure.

3.1.4 Present Research

The line of research developed here looks to implant a communicated belief in a manner that allows for uptake in the absence of additional motivations (such as experimenter or authority effects). Through Amazon's Mechanical Turk (MTurk), an on-line lottery context was used in which participants were told they would be choosing between two lotteries repeatedly, trying to generate the greatest overall payment. A novel manipulation was designed that communicated beliefs through an on-line "comment section" prior to the task. Participants were shown anonymous comments from "previous players" regarding the task, under the guise that they could add their own comments once they had finished making their choices, with the aim of providing information to both the task developers and other players. The "previous players" comments were in fact generated by the task itself, with most being neutral in nature, whilst those in the belief conditions contained a directional hypothesis. Participants were taken through filter questions at the end of the experiment to ensure they had not seen through the manipulation. Such a manipulation provides a novel, ecologically current form of communicating a belief that avoids aforementioned motivation pitfalls (Kunda, 1990), as participants believed they were simply playing a game with the goal of making as much money as possible.

Following this manipulation was a series of binary choices between the two lottery machines. One of the two machines (unknown to the participants) would start off as the probabilistically dominant option for a number of trials, before these probabilities then reversed, known as a probabilistic reversal (Peterson & DuCharme, 1967). Having a reversal of evidence lead to three between subject groups: a control group (that received no communicated belief), a belief group that received initial supportive evidence (BIS group), and a belief group that received initially undermining evidence (BIU group). All groups saw the same two-sided evidence, that is all participants saw exactly the same evidence for the two machines (which had equal outcomes overall), only the order of these outcomes was pseudo-randomised to create two "phases" in which one machine dominates over the other.

By using a reversal, it becomes possible to discern whether learning is still on-going and beliefs are updating (whether in a biased manner in the manipulation groups or at all in the control groups), rather than biased choices being due to a gradual lapse in attention over time to new evidence. In the latter case one would expect no effect of reversal on the proportion of choices made. Further, by providing a reversal-point, one can discern whether any bias is due to the different starting point (one would expect those told about an option to start with that option, whilst those told the alternative would start with the opposite choice). If group differences (bias) converge over the course of the first half of the task and then diverge upon reversal (when evidence changes), then participants are still actively integrating new evidence (assuming all groups change upon reversal, as in the prior point regarding the detection of ongoing learning), albeit in a biased manner in the case of manipulation groups. Alternatively, if the difference in the proportion of choices for each group remain the same throughout the entirety of the task (and thus differences in the second phase might simply be due to the pre-existing difference from the first phase), then conservatism (Dave & Wolfe, 2003; Phillips & Edwards, 1966; Pitz et al., 1967) is occurring. Perhaps most interestingly, if groups do converge on the initially dominant option, when the evidence then reverses, do biases re-emerge dependent on what participants had communicated to them prior to experiencing evidence?

Additionally, the three-group design allows for the investigation of whether a communicated belief would be abandoned (the BIU group pre-reversal, and BIS group post-reversal) when evidence did not favour it. Further, if a consequent bias occurs irrespective of the order of evidence (in other words the BIU group is just as biased in phase one, when it does not have supporting evidence, as the BIS group is in phase two, when it also does not have supporting evidence – but has had 100 trials of supporting evidence preceding it), then the communicated belief is acting as a strong prior (and thus dominating the effect of evidence). Alternatively, if a communicated belief is dependent on initial support – and the BIU group thus shows no bias as it has not been initially supported by evidence – then it is possible to rule out a strong prior explanation.

This method's capacity to assess whether learning is ongoing, whether there are conservatism effects in updating (irrespective of possible belief-evidence interactions), and detect "strong prior" explanations for belief effects, aims to provide a solid grounding for determining biasing effects of communicated beliefs in the integration of evidence. Furthermore, the design of both the belief manipulation and evidence presentation aim to improve upon previous research by removing aforementioned motivations and selective evidence exposure, both of which add unwanted alternative explanations for any subsequent bias.

3.2 Experiment 1

3.2.1 Method

Following the outline set out in the present research. Experiment 1 was designed using an 80-20 probabilistic reversal with 100 trials each side of the reversal, resulting in 200 trials in total. These trials were preceded by the aforementioned "comment section" which contained a communicated belief for the manipulation groups ("Machine A seemed luckier to me"), with the remainder (and for the control group, the entirety) of the comments of a neutral nature (e.g., "fun task", "seemed interesting").

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 500 prior HITs. Participants completed the experiment under the assumption the purpose of investigation was general gambling behaviours when using multiple lotteries. Participants were English speakers between ages 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

Design. Two lottery machines, labelled A and B, that both generated six number outcomes were used as choice context (see Appendix A.1.3 for an example trial output). The instructions given to participants explained that each machine uses a unique algorithm, based on the hyper-geometric distribution of outcomes intrinsic to most modern lotteries (Stern & Cover, 1989).

The order of outcomes between the two machines were structured into two phases of 100 trials. For the first phase 1 machine yielded 80% of the "wins" – classified as providing an outcome better than the alternative (e.g., machine A has one ball that matches the ticket, whilst machine B has two balls that match the ticket, so machine B has "won"), whilst the other yielded "wins" 20% of the time. During the second phase these probabilities reversed. Hence, the overall proportion of outcomes matched the natural probabilities of the hyper-geometric distribution found in 6 ball lotteries. Within each phase the order of outcomes was randomized. Which machine started off dominant was counterbalanced between participants. In line with the natural odds, 20% of trials were consequently uninformative as they involved draws (0-0, 1-1, 2-2) between the two machines, the remaining diagnostic trials (80%) followed the aforementioned probabilistic reversal.

Procedure. Before starting the trials, participants were shown a "comment section" in which previous participants had written their thoughts regarding the task. These comments were rigged to appear to be from other MTurk participants (complete with fake MTurk ID numbers), and were of a hypothesis neutral nature ("interesting task.", "good fun, thanks!"). For the belief conditions, instead of all comments being neutral (as in the control condition), the top comment was a communicated belief regarding the two machines ("I felt that machine A(B) was much luckier!").

On each trial participants pressed a button that generated a "ticket" of three numbers, and then chose a machine to gamble with on each trial. Participants were invited to earn as many points as possible, based on the number of matches between their "ticket" and their chosen machine. Each trial cost participants one point, so a failure to match any numbers resulted in a net loss of −1 point, whilst a single ball matched earned 2 points, 2 balls matches earned 8 points, and 3 balls matched earned 50 points, reflected the increasing rarity of these outcomes (see Appendix A.1 for an example feedback screen). Participants were aware of their current total points earned during the trials, and instructed that their total amount of points directly corresponded to an increasing bonus payment in dollars.

For each trial, once participants had generated their ticket of numbers and selected a machine with which to gamble, participants then pressed a button to generate the outcomes for that trial. Each machine drew 6 balls, numbered from 1 to 49, with a new draw for each trial. Matches between the numbers of the participant and the selected machine were highlighted in green, whilst the forgone matches of the non-selected machine were highlighted in red.

Once participants had completed all gambles, demographics were filled out, along with questions regarding how often the participants felt they had chosen optimally, and how often they felt the other machine had provided a better outcome. Participants were also asked about the probabilities of each outcome for each of the machines, along with a brief questionnaire assessing various known correlates of superstition and gambling behaviours; locus of control (Levenson, 1973)[7], the revised Paranormal Belief Scale (Tobacyk & Milford, 1983)[8] items concerning luck, and neuroticism. Finally, participants completed a series of exit funnel questions to assess awareness of the comment section manipulation. Following completion of the task, participants were debriefed and given an email to contact if they had any further questions.

The main dependent variables under investigation were the proportion of choices made in favour of the initially dominant machine. We hypothesised that the BIS group (who receive initial support for the belief) would select the initially dominant machine in phase 1 (pre-reversal) significantly more than controls, and further, that this difference would persist into phase 2 (post-reversal). For the BIU group (who do not receive initial support for the belief), several aspects of communicated biasing effects became possible to test:

Firstly, phase 1 could demonstrate whether the communicated belief acts as a strong prior (i.e. the belief is given a very high value that takes a large amount of evidence to over-rule), which would result in the BIU group significantly differing in their proportion of choices relative to controls, favouring their belief indicated machine, in spite of its initial sub-optimality.

Secondly, for phase 2, two possible effects could occur in the BIU group: either the BIU group would favour the now dominant machine (which matches their

[7] Statements included were: “When I get what I want, it’s because usually I am lucky.” And “I have often found what is going to happen will happen.”. Response options were: “Strongly Agree”, “Agree”, “Slightly Agree”, “Slightly Disagree”, “Disagree” and “Strongly Disagree”.

[8] Items included were: “I tend to worry about life.”, “Black cats bring bad luck.”, “The number 13 is unlucky.”, and “If you break a mirror you will have bad luck”. Response options were: “Strongly Agree”, “Agree”, “Undecided”, “Disagree” and “Strongly Disagree”.

communicated belief) more so than controls (which leads to a linear effect of condition in phase 2), or the BIU group would have refuted the belief and would be no different from controls in phase 2.

Table 3.2.1 below summarises the key information from the above methodological description. This includes the phrasing of the task instructions, incentives scheme, and belief manipulation, as well as the measures taken (including manipulation check question phrasing).

Table 3.2.1: Experiment 1: Summary Table of Task Setup and Measures.

Setup / Manipulations	**Description**	**Details**	
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	"This task is designed to assess various gambling behaviours when playing lotteries. In this case you will be choosing between 2 lotteries. You will be paid based on your winnings. The more points you win, the **higher your total payment**."	
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.]	**Total Payment, based on points:** <50 points = standard $1, >50 points = $1.10, >100 points = $1.25, >150 points = $1.50, >200 points = $1.80, >250 points = $2.20, >300 points = $2.80	
Belief Manipulation	Online comment section of "previous participants". Shown to participants before trials under the guise of an interest in their thoughts. Example comments shown to participants controlled by experimenter, with 1 comment indicating a directional belief (manipulation), and remaining **9** comments of a neutral nature.	**Manipulation Comment (always top of screen)**: "I felt machine A was much luckier!" **Example Control / neutral comments**: "I tried to win as much as I could", "Good fun, thank you!", "took too long personally" (see Appendix A.1.1 for Sample comment screen)	
Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between two lotteries, A and B. 200 trials, 80% win rate A, 20% win rate B for first 100 trials, with reversal of probabilities for second 100.	n/a	A, B
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one machine?"	A, B, No

3.2.2 Results

Descriptives and Processing. The 400 participants recruited were US based, randomized into either the BIS (161), BIU (137) or control (102) conditions. The mean age was 35.52 years (48% female). The data were gathered in two stages: an initial run of 80 participants, with just the BIS and control groups allowed for an estimate of the power needed when also adding the BIU group (to test additional predictions) in run 2. This power analysis, using G*power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) was run using the smallest effect size of the dependent variables of interest, and indicated that to detected a significant effect of condition at the .05 level with 80% power would require an average group size of 91. For the two manipulation conditions this number was multiplied by about 1.5 to compensate for failures in passing the manipulation check (mentioned below), resulting in a total sample size of 360. Given the unknown nature of the additional BIU group, this was conservatively increased by 10% to 400. An ANOVA analysis was run, using experiment number as a covariate, finding no significant differences of experiment number on all dependent variables.

After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered[9]. If participants had no recollection of a manipulation comment (if one was presented in their condition), they were removed from subsequent analysis[10], leaving 110 in the BIS group (75% pass rate) and 81 in the BIU group (59% pass rate). The decision to remove those

[9] As a reminder, the exact phrasing of this question was “Do you think comments were biased towards one machine?”, with options of A, B, or No. If participants selected “No”, this was deemed grounds for removal. The question was preceded sequential exposure to first the question “Did you notice anything funny about the start of the experiment?” (open text response), to assess if participants suspected the cover story, followed by “Was there anything that influenced you regarding the previous participant comments?” (open text response).

[10] Significant differences were found with all participants included between conditions on the overall proportion of choices, $F(2, 399) = 4.291$, $p = .014$, $\eta^2 = .021$, and the subsequent contrast coding of the same analysis, $F(1, 396) = 8.185$, $p = .004$, $\eta^2 = .02$. This pattern of results is explained further in the main effects section.

who failed to remember was taken to reduce noise in further analysis break downs, specifically when breaking down the analysis into phases, participants who failed this check added a large amount of variance when analysing at a finer level. Furthermore, by removing those who fail the manipulation check, it is possible to better ensure possible differences between groups in remaining participants (notably for the BIU group) were not due to failures in memory or registering the manipulation in the first place.

The difference in the proportion of those who failed to remember the manipulation between groups was not significant. The remaining 293 participants (49% female), average age 35.12 years (*SD* = 12.08), were used for the analyses below. The results below, as for all experiments within this chapter, will not be presented exhaustively; for the sake of brevity, only those of central interest to the thesis questions are included.

Correlates. Locus of Control, age, gender, gambles per week and revised Paranormal Belief Scale (rPBS) variables were not correlated with any of the dependent variables[11].

Choice Data. The key dependent variables used in the analysis were the total proportion of choices made in favour of the initially dominant machine, and the proportion of these choices made broken down by phase (shown in grey in Figure 3.2.1).

[11] Due to a minor programming error, one counterbalance condition had a slight imbalance in the number of 2 ball matches in the second phase on one machine. To remove possible issues, counterbalancing was used as a covariate in all further analyses, as the counterbalancing factor was not exactly even across groups.

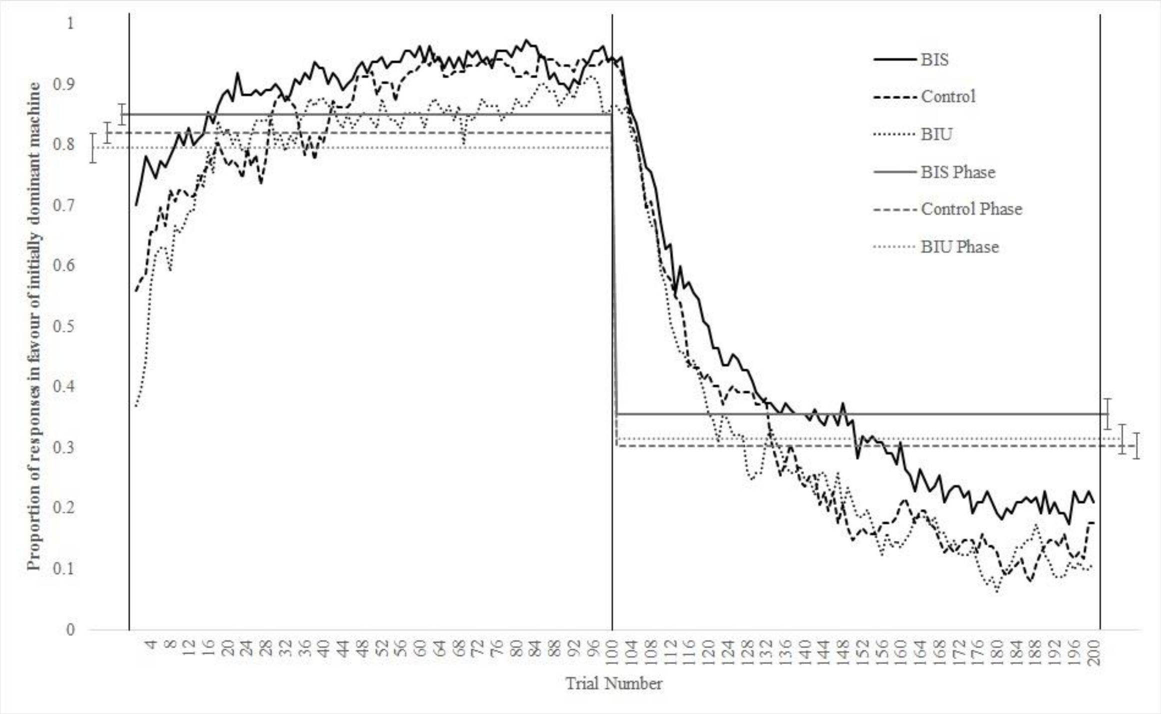


Figure 3.2.1. Experiment 1: Choice Data (Trend Lines with Phase Overlay). Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant machine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

A series of ANOVA tests were conducted to determine the main effects of condition and phase on the proportion of choices made. Significant effects of condition, *F*(2, 289) = 5.041, *p* = .007, and phase, *F*(1, 289) = 137.84, *p* < .001, were found on the proportion of choices made in favour of the initially dominant machine. The interaction term between phase and condition was not significant.

A series of pairwise comparisons between groups both overall and within each phase was conducted to break down the main effect of condition. These comparisons found the BIS group to be significantly higher than both controls, *F*(1, 211) = 4.804, *p* = .029, η^2 = .022, and BIU groups, *F*(1, 190) = 8.579, *p* = .004, overall. Furthermore, when broken down by phase, the BIS group was significantly higher than the BIU group in both phase 1, $F(1, 190) = 5.152$, $p = .024$, and phase 2, $F(1, 190) = 4.401$, $p = .037$, whilst the difference between BIS and controls did not reach significance for phase 1, $F(1, 211) = 1.697$, $p = .194$, $\eta^2 = .008$, or phase 2, $F(1, 211) = 3.733$, $p = .055$, $\eta^2 = .018$.

The differences between the BIU and control groups were not significant overall ($p = .329$), in phase 1 ($p = .318$) or phase 2 ($p = .632$)[12], a trend that is evidenced in the phase lines (grey) of Figure 3.2.1, suggesting that the undermined belief is abandoned. Visual inspection of the choice data suggested that the BIU group are no different from controls, whilst the BIS group is significantly different from the two, in line with the hypothesis that biasing effects of a second-hand belief relies on exposure to initially supporting evidence. This was further corroborated by the pairwise comparisons, leading to the development of a post-hoc contrast code analysis.

Post-hoc Contrast Code Formation. Accordingly, to test whether the BIU group were no different than controls in comparison to the BIS group, a contrast code ANOVA was conducted (BIS, Control, BIU: 2, −1, −1). The contrast code analysis of overall choice proportion was significant, $F(1, 289) = 9.535$, $p = .002$, $\eta^2 = .032$, demonstrating a significantly higher number of choices in the BIS group as compared to both control and BIU groups.

Breaking this down by phase, (as illustrated by the grey lines in Figure 3.2.1), the contrast code persisted in both phase 1, $F(1, 289) = 4.509$, $p = .035$, $\eta^2 = .015$, and phase 2, $F(1, 289) = 6.077$, $p = .014$, $\eta^2 = .021$, again demonstrating a significantly higher number of choices in the BIS group as compared to both control and BIU groups. The interaction between phase and contrast code was not significant ($p = .676$).

[12] Bayesian T-tests of these pairwise comparisons were conducted using the JASP statistical programme (Love et al., 2015), using a uniform prior across possible models (as used across all subsequent Bayesian analyses unless specified otherwise). Substantial support was found for the null for overall, $BF_{10} = .249$, phase 1, $BF_{10} = .259$, and phase 2, $BF_{10} = .178$, in accordance with the $<1/3^{rd.}$ cut off recommendation (Dienes, 2014).

Phase 1 Convergence. Visual inspection suggested that by the end of phase 1 participants had all converged towards the dominant machine. To test this, a mixed ANOVA of contrast coded condition (between subjects) x 10 trial epochs (10 in total) within-subjects was conducted. The consequent interaction between epoch and contrast code was significant, $F(2, 290) = 5.657$, $p = .004$, $\eta^2 = .038$, indicating a convergence of the contrast effect over the course of the phase[13].

3.2.3 Discussion

Several conclusions follow from these results. As can be seen from Figure 3.2.1, there is a strong effect of reversal, indicating that learning is ongoing in all groups. As such differences between groups are unlikely to be due to task disengagement and instead suggest participants were attentive to changes in the evidence. It should also be noted that this sensitivity to reversal was present in all three groups, which indicates the phase 2 contrast effect is not due to the BIS group no longer being attentive to any changes in evidence. In line with this, the convergence of all groups during the first phase suggests that the contrast effect during this phase could be due to starting point differences between the BIS and other groups. However, the subsequent re-emergence of the contrast effect in the second phase indicates that the effect of the belief seen in phase one has not been washed out by the evidence, and instead the BIS group's belief is still playing a role in biasing the integration of subsequent evidence, relative to the other groups.

This convergence is one indicator that a conservatism explanation of bias does not fit across the three groups as well as suggesting that a strong prior explanation for the effect of the communicated belief is not appropriate. This leads to the final, and

[13] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first epoch, and the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, $BF_{10} = 149.4$, whilst strong support was found for the null in accordance with the $<1/3^{rd.}$ cut off recommendation for the final epoch, $BF_{10} = .136$.

most important conclusion of Experiment 1: even with the presence of 2-sided evidence, and the reduction of motivational explanations of confirmation such as experimenter demand effects (along with the introduction of a competing (against belief adherence) accuracy incentive), it is initial evidence that dictates a belief's subsequent biasing effects. The immediate mapping of the initially undermined (BIU) manipulation group onto the control group suggests that initial evidence is needed to consolidate a communicated belief. It is important to further note that given the manipulation check criteria, the similarity between the BIU group and control group in the analyses was not due to participants in the BIU group having forgotten the communicated belief. This suggests that those in the BIU group have *refuted* their communicated belief, and are thus unaffected by it, even when the evidence changes to support it. Consequently, the most suitable explanation for the effect of a communicated belief on evidence is initially supportive evidence is required to consolidate the belief, but once consolidation has occurred, learning is still active, but biased to favour confirmation when interacting with subsequent evidence.

There are several limitations pertaining to Experiment 1. Firstly, the contrast analysis for Experiment 1 was post-hoc, and therefore requires replication. Secondly, the probability distributions used in the tasks may have been too strong. By having an 80/20 probability distribution, the dominance of one machine over the other was clear to all groups (as evidenced in Figure 3.2.1) demonstrated by their convergence during phase 1. All groups were also able to detect the reversal that occurred at trial 100, which is not unsurprising given the severity of the reversal, as the initially dominant machine becomes 60% worse in combination with the counterfactual showing the dominance of the alternative. This may have led to ceiling effects in the number of choices between manipulation groups and controls, that might have been teased apart better by a more uncertain environment.

Returning to the literature on confirmation bias effects, this may have led to an increased plausibility of alternative hypotheses that allowed manipulation groups to negate the validity of the second-hand belief (Klayman, 1995), muting the efficacy of a bias that might otherwise have persisted under uncertainty.

3.3 Experiment 2

Accordingly, the purpose of Experiment 2 was to replicate the exploratory contrast effect of Experiment 1 *a priori*, improving upon the design given prior limitations.

3.3.1 Method

The design and procedure followed that of Experiment 1, with the following changes outlined below.

Firstly, despite intriguing real world ramifications of the effect that such a small intervention can have on updating, a stronger manipulation (increasing the number of comments indicating the same directional belief) was constructed. The comment manipulation was increased in strength from one second-hand belief to three (all in the same direction). This was done to improve the rate of manipulation check failures experienced in Experiment 1 by improving the visibility of the manipulation to participants. Additionally, it increased the reliability of the communicated belief (Siegrist et al., 2000; Yaniv & Kleinberger, 2000), as there were now several (supposedly independent) sources all providing the same preference. As such, the belief could be interpreted as a trend (hence more likely to be valid), rather than a one-off occurrence. Relative to neutral comments, these manipulation comments were still in the minority, to avoid social conformity issues.

Secondly, the probabilistic reversal was altered from 80/20 – 20/80 to 70/30 – 30/70. This reduction in the severity of the reversal was introduced to help reduce possible ceiling effects discussed above due to the dominant machine (and reversal) being too obvious to both controls and manipulation groups alike. This change was aimed at teasing apart possible biasing effects further.

The third change was the introduction of three posterior measures at the end of the main task, in which participants chose which of the two machines they preferred ("Which machine do you think is better?"; binary preference), how confident they were (0-100%) in that preference, followed by their estimation of the distribution of better outcomes between the two machines ("What is the spread of better outcomes between the two machines?", from 100% A, through 50/50, to 100% B on a 100 point scale). The addition of posterior measures allows for the testing of whether the bias seen in the BiS group in Experiment 1 that converged with the other groups during phase 1, but then re-emerged following the reversal is a reflection of a truly consolidated belief.The measures were therefore included as a supplemental, exploratory measure to investigate if the contrast effects in the main (behavioural) dependent variables are found in end-of-sequence judgements as well (Hogarth & Einhorn, 1992).

Hypothesis. The hypotheses for Experiment 2 are primarily based on the contrast code effects found in Experiment 1. Those who receive a belief that is initially supported (BIS) will choose the machine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BIU) will not show such a bias and be no different from the control group. Furthermore, with the inclusion of the posterior probability measures, contrast effects were predicted to extend to these measures as well.

Table 3.3.1 below summarises the key information for Experiment 2. Importantly, entries that differ from Experiment 1 are italicised.

Table 3.3.1: Experiment 2: Summary Table of Task Setup and Measures.

Setup / Manipulations	**Description**	**Details**	
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	"*This task is designed to look at experiences when playing two lotteries. We are interested in the different randomization algorithms running each lottery.* You will be paid based on your winnings. The more points you win, the **higher your total payment**."	
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.]	**Total Payment, based on points:** <50 points = standard $1, >50 points = $1.10, >100 points = $1.25, >150 points = $1.50, >200 points = $1.80, >250 points = $2.20, >300 points = $2.80	
Belief Manipulation	Online comment section of "previous participants". Shown to participants before trials under the guise of an interest in their thoughts. Example comments shown to participants controlled by experimenter, *with 3 comments indicating the same directional belief (manipulation)* , and remaining *7* comments of a neutral nature.	**Manipulation Comments**: "I felt machine A was much luckier!", "the algorythm for B had better outcomes.", "seemed like B was better to me!" (last comment required scrolling down the screen). **Example Control / neutral comments**: "I tried to win as much as I could", "Good fun, thank you!"(see Appendix A.1.2 for Sample comment screen)	
Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between two lotteries, A and B. 200 trials, *70% win rate A, 30% win rate B for first 100 trials* , with reversal of probabilities for second 100.	n/a	A, B
Binary Preference	*Posterior Measure: Following trials, participants were asked which of the two machines they preferred.*	*"Which machine do you think is better?"*	*A, B*
Confidence in Binary Preference	*Posterior Measure: Having given their binary preference, participants were asked how confident they were in their preference.*	*"How confident are you in this preference?"*	*0-100 slider (default value of 0)*
Probability Estimate	*Posterior Measure: Participants were asked to give a probability estimate of the distribution of better outcomes between the two machines.*	*"What is the spread of better outcomes between the two machines?"*	*100% A, through 50/50, to 100% B (slider, default value 50/50)*
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one machine?"	A, B, No

3.3.2 Results

Descriptives and Processing. Based on the contrast analysis of Experiment 1, a second power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 2. Converting partial eta squared values for the contrast code analyses of the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 90 per group. Following the same procedure as Experiment 1, the groups sample sizes were increased to compensate for those failing manipulation checks, calculated from the failure rates of Experiment 1, resulting in a total sample size of 360.

Participants were recruited online using MTurk. Those who had taken part in the previous experiment were ruled out from participating. The 360 participants recruited were US based, randomized into either the BIS (121), BIU (122) or control (117) conditions. The average age was 34.7 years (*SD* = 11.47) and the sample was 49% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis[14], leaving 103 in the BIS group (85% pass rate, up from 75% in Experiment 1) and 96 in the BIU group (79% pass rate, up from 59% in experiment 1). The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

As expected, the change from showing one comment to several comments decreased the drop-out rate, although these differences (both between groups and

[14] Following the same protocol as Experiment 1, significant effects were found for both the effect of condition on the overall proportion of choices, *F*(2, 359) = 4.76, *p* = .009, η^2 = .026, and the contrast coding of the same analysis, *F*(1, 357) = 8.233, *p* = .004, η^2 = .023, regardless of manipulation check removal.

between studies) were not significant. The following analyses were conducted using the remaining 316 participants, with an average age of 34.71 years (*SD* = 11.41) and 50% female.

Correlates. Locus of Control, age, gender, gambles per week and revised Paranormal Belief Scale (rPBS) variables were not correlated with any of the dependent variables.

Choice Data. As can be seen in Figure 3.3.1, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant machine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant machine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

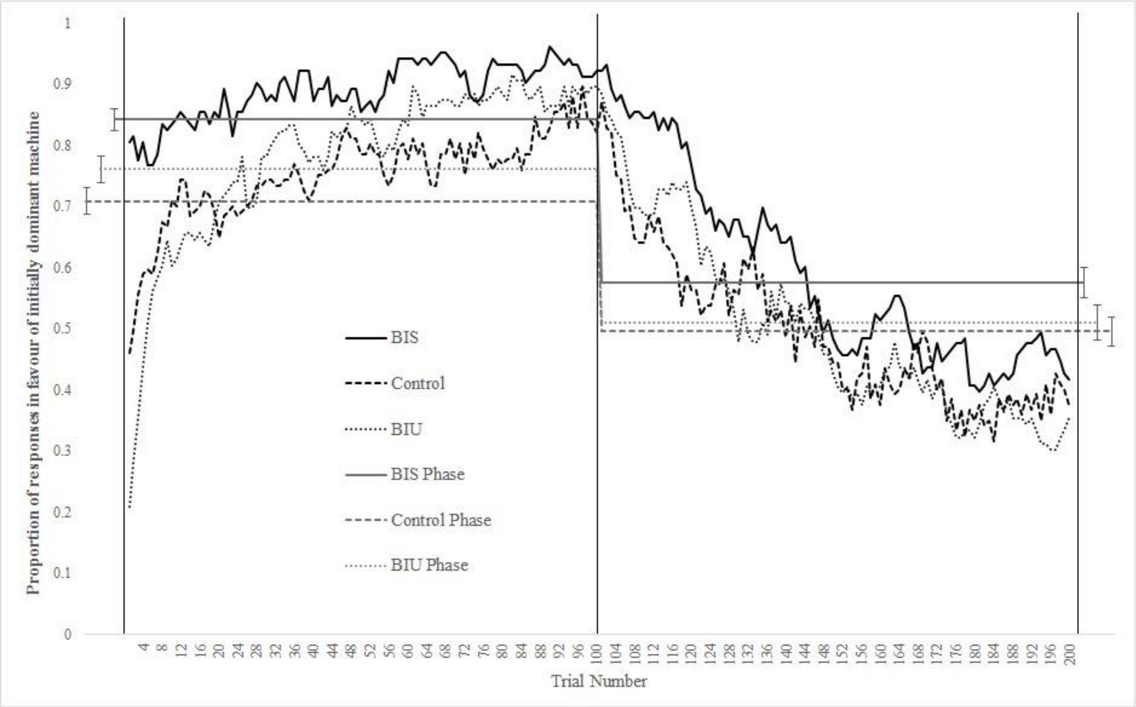


Figure 3.3.1. Experiment 2: Choice Data (Trend Lines with Phase Overlay). Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially

dominant machine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

To assess whether Experiment 2 replicated the key findings of Experiment 1, the same contrast code analysis procedure was conducted for the three conditions (BIS, Control, BIU: 2, −1, −1), along with pairwise comparisons between groups. The contrast code analysis of condition on the overall proportion of choices made was significant, $F(1, 313) = 13.637$, $p < .001$, $\eta^2 = .042$. Along with corroborating pairwise ANOVA analyses which show the BIS group was significantly higher than controls, $F(1, 219) = 15.192$, $p < .001$, $\eta^2 = .065$, and BIU, $F(1, 198) = 7.111$, $p = .008$, $\eta^2 = .035$, groups, whilst the difference between controls and BIU was not significant ($p = .318$).

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Figure 3.3.1, prior to reversal point), the contrast code was significant, $F(1, 313) = 16.69$, $p < .001$, $\eta^2 = .051$. This was corroborated by pairwise ANOVA showing the BIS group was significantly higher than controls, $F(1, 219) = 19.993$, $p < .001$, $\eta^2 = .084$, and BIU, $F(1, 198) = 8.188$, $p = .005$, $\eta^2 = .04$, groups, whilst the difference between controls and BIU was not significant ($p = .129$).

These effects continued into the second phase (grey lines in Figure 3.3.1, post reversal point), as the contrast code was again significant, $F(1, 313) = 5.046$, $p = .025$, $\eta^2 = .016$. This was corroborated by pairwise ANOVA showing the BIS group was significantly higher than the control group, $F(1, 219) = 4.739$, $p = .031$, $\eta^2 = .021$, whilst the difference between controls and BIU was not significant ($p = .805$). This difference between the BIU and BIS groups was not significant in phase 2 ($p = .077$), however as the two key comparisons (BIS is different from controls, whilst BIU is not) remain significant, the position of the BIU group proportion (on the BIS group side of the control group) further supports the notion of the BIU group refuting their belief.

Phase 1 Convergence. Following the analysis protocol of Experiment 1, a mixed ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1. The contrast code was the between-subjects grouping factor, and 10 trial epochs were within-subjects. The consequent interaction between epoch and grouping was significant, $F(2, 313) = 8.941$, $p < .001$, $\eta^2 = .054$, indicating a convergence of the contrast effect over the course of the phase[15].

Posteriors. The posterior measure of estimated probability distribution (see Figure 3.3.2) showed a significant contrast code effect of condition, as found in the above behavioural measures, $F(1, 313) = 4.192$, $p = .041$, $\eta^2 = .013$.[16]

[15] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast code in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first block, $BF_{10} = 4.73*10^5$. In the final epoch, there was no longer support for the contrast effect, but this did not reach the $<1/3^{rd.}$ to be classified as substantial support for the null, $BF_{10} = 1.731$. However, it should be noted that a Bayesian ANOVA on the subsequent epoch (first epoch post-reversal) does indicate a reversal in trend from convergence to divergence upon reversal, as strong evidence is once again found for the contrast effect, $BF_{10} = 10.41$.

[16] A Bayesian T-test was conducted to confirm the lack of a difference between controls and BIU was not due to insufficient power by looking for strong support of the null (a Bayes Factor of $< 1/3^{rd.}$). Substantial evidence was found in support of the null, $BF_{10} = .186$.

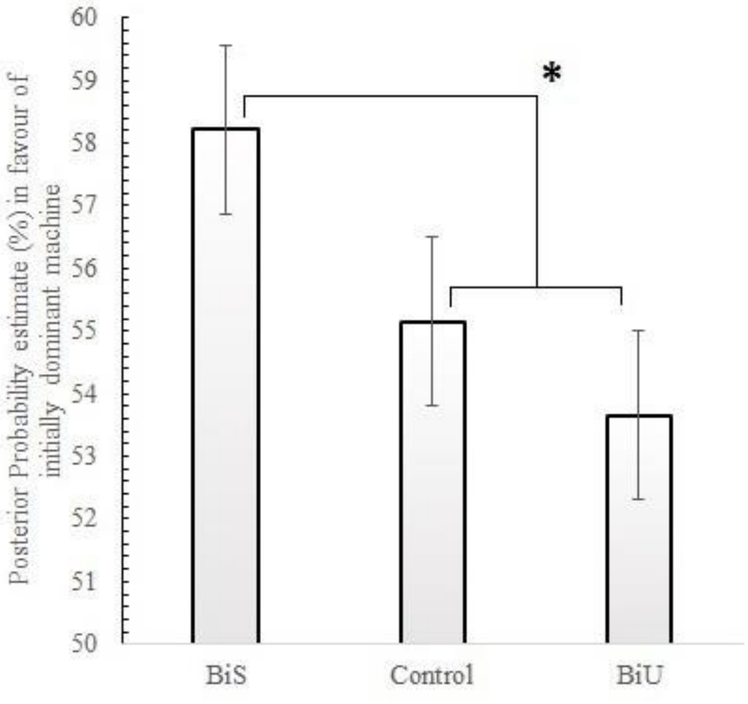


Figure 3.3.2. Experiment 2: Posterior Probability Estimates as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant machine, less than 50% indicates a preference for the initially sub-optimal machine. Outcomes split by group (* p <.05).

A further series of t-tests comparing the probability estimates of each group to a 50% expected value revealed a primacy effect, wherein the initial evidence favouring one machine dominated the reversed evidence in the latter half of the task, in the BIS, *t*(102) = 5.862, *p* < .001, 95% CI [5.43, 10.99], control, *t*(116) = 3.586, *p* < .001, 95% CI [2.3, 7.99], and BIU, *t*(95) = 2.129, *p* = .036, 95% CI [.25, 7.05], groups.

Participants' binary preference posterior measure did not yield a main effect of condition, but did show strong primacy effects, wherein the initial evidence favouring one machine dominated the reversed evidence in the latter half of the task (using a Chi squared comparing proportion of preferences to .5), in BIS (*M* = .69, *SD* = .465), X^2(1, N = 103) = 14.767, *p* < .001, control (*M* = .64, *SD* = .482), X^2(1, N = 117) = 9.308, *p* = .002, and BIU (*M* = .68, *SD* = .47), X^2(1, N = 96) = 12.042, *p* < .001, groups. The confidence measure for the binary preference did not show any significant effects of condition, or order effects.

3.3.3 Discussion

Replication of the findings of Experiment 1 demonstrates the necessity of initial supporting evidence in validating the second-hand belief, leading to subsequent biases in updating. Similar to the effects discussed in Experiment 1, phase 1 differences can be seen as due to adjustment from initial starting points (as dictated by communicated belief). However, despite all participants (regardless of group) converging beyond a probability matching (Edwards, 1961) level (choosing the dominant option 70% of the time) by the end of phase 1, the contrast code once again re-emerges in phase 2.

Importantly, when investigating the posterior measures, despite a general tendency towards primacy amongst all groups, which is not surprising for an end-of-sequence judgement (Hogarth & Einhorn, 1992), an asymmetry between the primacy in the BIS group as compared to the other groups existed. This was corroborated by a replication of the contrast analysis for the posterior probability estimate. Such an extension lends credence to the argument that the biasing effect due to consolidating a communicated belief has persistent effects beyond trial-by-trial choices and into end-of-sequence judgements. Furthermore, such a bias occurred despite all groups witnessing the same evidence (due to the presence of counterfactual feedback), which suggests this difference is not simply a product of a differential in evidence exposure.

Finally, Experiment 2 replicated the effects found in Experiment 1 using a 70/30 probability reversal, which reduced ceiling effects in the proportion of choices made within each phase. The second change to paradigm – including multiple comments indicating the communicated belief – resulted in a lower proportion of participants failing the manipulation check and likely attributed to a stronger contrast effect, in line with models of source trust in advice taking literature (Siegrist et al., 2000; Yaniv, 2004), as perceived trust is increased when several sources indicate the same information, up from a single source.

3.4 Experiment 3

Experiment 3 set out to extend the effects found in Experiments 1 and 2 into the domain of health decision making. The first reason for this change was to ratify possible claims of generalizability of the effects in question, by moving outside of a gambling context. Secondly, in making the transition away from gambling, it was hoped that improvements could be made in short-term "streak-shooting" noise in the choice data. In other words, given the aim of the experiments was to have participants focus on the long-term, overall quality of the two options available to them, the context of lotteries is conducive to short-term, fallacious strategies such as gambler’s fallacy (Ayton & Fischer, 2004; Barron & Leider, 2010; Jessup & O’Doherty, 2011). Such strategies are based on recent performance, such as assuming that because one machine has not won recently, then it is "due" for a win. These strategies are hence, in relation to the overall effects under investigation, a possible distraction.

Furthermore, by moving into the domain of health, outcomes are defined as a medicine either "curing" or "failing to cure" the disease in a particular patient. In Experiments 1 and 2, where outcomes were variable (number of ball matches) for each machine, and thus comparisons between machines had to be based on a calculation of the relative number of matches between options (e.g., an assessment of whether a two-ball match is better than 3 sets of one-ball matches).

By simplifying this to a 1 (cure) or 0 (fail to cure) outcome for each option, one can better assess the role of outcome probability as intrinsic to the biasing effects in question, having eliminated alternative elements of computational difficulty and ambiguity. In doing so, it is possible to therefore answer whether probabilistic ambiguity alone is sufficient to sustain previously found effects. The change to binary outcomes bares a closer parallel to more social implications of these effects, such as impression formation and stereotyping (Anderson, 1965), wherein evidence is often categorically present or absent.

The final advantage of extending these effects into the health domain is the necessary implication for belief manipulation generalizability. By altering the context in which evidence is integrated, the belief itself also necessarily needs to be adapted to fit the new context. The consequent change in the content of the belief not only speaks further to the generalizability and robustness of the effects in question, but further allows for initial inferences in the degree to which beliefs are processed.

Accordingly, the methods below are a direct extension of Experiment 2, with the following changes:

3.4.1 Method

The context for the task was changed from a lottery task in which participants were required to assess the relative strength of two lottery machine algorithms, to a health domain in which the participant played the role of a physician prescribing medicines to patients. In this way, each trial was a new patient, presenting with the anonymised disease "Q". Participants were tasked with assessing the overall efficacy (i.e. across different patients) of two new medicines, anonymized to "K" and "Z". The cover story stated that each patient varied by genotype, and such variance may result in different responses to the two medicines. Such a change in context also required the comment section manipulation to change to reflect the health domain, so instead of a manipulation comment of "Machine A seemed luckier to me" as used in the lottery experiments, comments instead reflected the medicine options (e.g., "I think medicine Z was the most effective" and "medicine Z was better than K."). In this way, communicated beliefs regarding the options still reflected the unquantified, directional hypotheses used previously.

As mentioned above, this change in context also allowed for the response format to change to a stricter, binary set of outcomes (instead of variable numbers of ball matches). A successful trial was defined as when the selected medicine "Cured" the disease (with the participant winning 3 points as a reward), and an unsuccessful trial as when the selected medicine had "No Effect" (costing the participant 1 point). As in previous experiments, participants could see the counterfactual outcomes for each trial (what the outcome for the patient would have been had they selected the other medicine), and were incentivized with increasing monetary bonuses for each 50-point boundary they crossed in earnings throughout the task, as in Experiments 1 and 2. Similarly, both the number of trial pre- and post-reversal remained the same, with 100 each side, and the probabilities of each option were again 70/30 pre-reversal, and 30/70 post-reversal.

Finally, in the demographics and questionnaire section following the main task and posteriors, Locus of Control and Revised Paranormal Belief Scale measures, which had previously failed to yield any relationships to both behavioural and judgement data, were replaced by an abbreviated Need for Closure scale (Roets & Van Hiel, 2011). Given prior literature associating Need for Closure with the propensity towards engagement, deliberation and entertainment of alternative hypotheses (Webster & Kruglanski, 1994) in individuals, this was hypothesised to have a potential impact on the biasing effects under investigation.

Hypothesis. The hypotheses for Experiment 3 are to replicate the contrast effects found in Experiment 1 and 2, extending these effects into a health context. Those who receive a belief that is initially supported (BIS) will choose the medicine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BIU) will not show such a bias and be no different from the control group. These effects are also hypothesised to extend to posterior probability estimates, replicating Experiment 2.

Table 3.4.1 below summarises the key information for Experiment 3. Importantly, entries that differ from Experiment 2 are italicised.

Table 3.4.1: Experiment 3: Summary Table of Task Setup and Measures.

Setup / Manipulations	**Description**	**Details**	
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	*"Your objective is to try and cure as many patients as you can by learning the effectiveness of the medicines. Each successful cure will earn you points. The medicines both cost 1 point to take. At the end* ***your score will determine your bonus in dollars as follows:*** "	
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.]	**Total Payment, based on points:** <50 points = standard $1, >50 points = $1.10, >100 points = $1.25, >150 points = $1.50, >200 points = $1.80, >250 points = $2.20, >300 points = $2.80	
Belief Manipulation	Online comment section of "previous participants". Shown to participants before trials under the guise of an interest in their thoughts. Example comments shown to participants controlled by experimenter, with 3 comments indicating the same directional belief (manipulation), and remaining **7** comments of a neutral nature.	**Manipulation Comments**: "*I think medicine Z was the most effective* ", "*medicine Z was better than K.* ", "seemed like *Z* was better to me!" (last comment required scrolling down the screen). **Example Control / neutral comments**: "I tried to win as much as I could", "Good fun, thank you!"(see Appendix A.1.4 for Sample comment screen)	
Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between *two medicines, Z and K*. 200 trials, 70% cure rate *Z*, 30% cure rate *K* for first 100 trials, with reversal of probabilities for second 100.	n/a	*Z, K*
Binary Preference	Posterior Measure: Following trials, participants were asked which of the two machines they preferred.	*"Which medicine do you think is better?"*	*Z, K*
Confidence in Binary Preference	Posterior Measure: Having given their binary preference, participants were asked how confident they were in their preference.	"How confident are you in this preference?"	0-100 slider (default value of 0)
Probability Estimate	Posterior Measure: Participants were asked to give a probability estimate of the distribution of *cures between the two medicines*.	*"What is the distribution of cures between the two medicines?"*	100% *Z*, through 50/50, to 100% *K* (slider, default value 50/50)
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one medicine?"	*Z, K, No*

3.4.2 Results

Descriptives and Processing. Based on the contrast code analysis of Experiment 2, a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 3. Converting partial eta squared values for the contrast code analyses for the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 80 per group. Following the same procedure as Experiment 1, the groups sample sizes were increased by 33% to compensate for those failing manipulation checks, calculated from the failure rates of Experiment 1 and conservatively increased to 360.

Participants were recruited online using MTurk. Those who had taken part in the previous experiment were ruled out from participating. The 360 participants recruited were US based, randomized into either the BIS (103), BIU (119) or control (138) conditions. The average age was 35.52 years (*SD* = 11.181) and the sample was 47.2% female.

After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis[17], leaving 77 in the BIS group and 93 in the BIU group. The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

As expected, the change from showing one comment to several comments decreased the drop-out rate, although these differences (both between groups and

[17] Following the same protocol as Experiment 2, a significant effect was found for both the contrast code analysis of the overall choices, *F*(1, 356) = 5.341, *p* = .021, η^2 = .015, and posterior probability estimates, *F*(1, 356) = 4.941, *p* = .027, η^2 = .014, regardless of manipulation check removal.

between studies) were not significant. The following analyses were conducted using the remaining 299 participants, with an average age of 35.37 years (*SD* = 10.903) and 49.5% female.

Correlates. Need for closure did not correlate or interact with any of the effects under investigation, but was found to have an impact on the speed of learning across epochs in the first phase. Accordingly, Need for Closure was assessed in the convergence analysis. Age and gender variables were not correlated with any of the dependent variables.

Choice Data. As can be seen in Figure 3.4.1, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant medicine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant medicine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

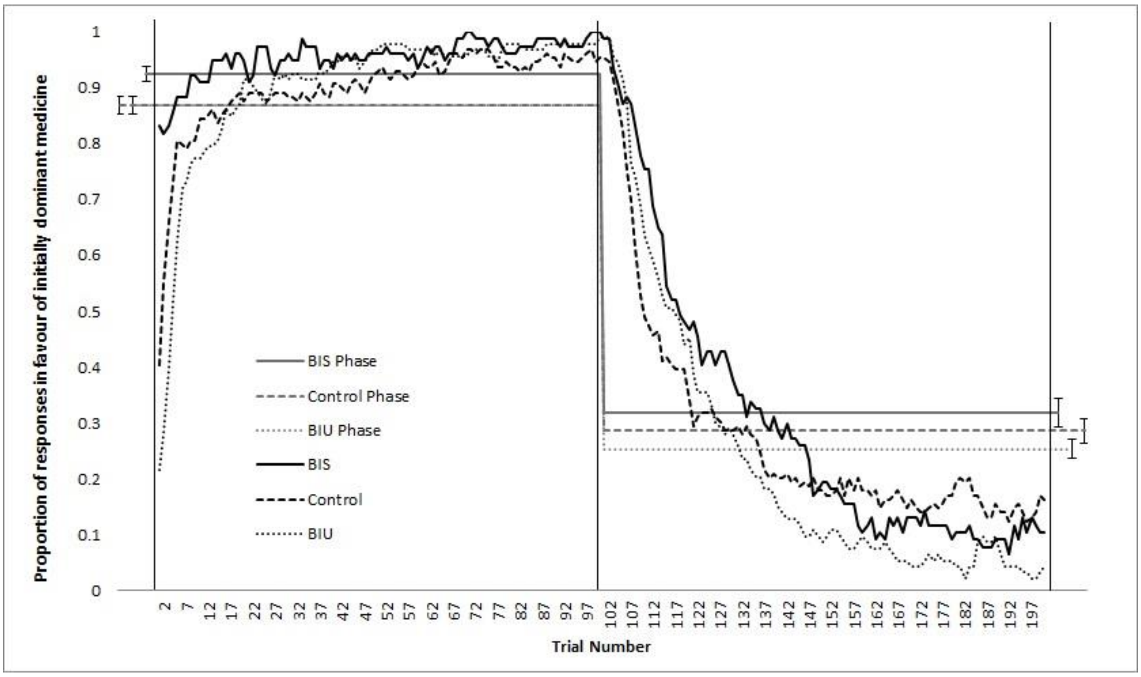


Figure 3.4.1. Experiment 3: Choice Data (Trend Lines with Phase Overlay). Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant medicine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

To assess whether Experiment 3 replicated the key findings of Experiment 2, the same contrast code analysis procedure was conducted for the three conditions (BIS, Control, BIU: 2, −1, −1), along with pairwise comparisons between groups.

The contrast code analysis of condition on the overall proportion of choices made was significant, *F*(1, 296) = 8.462, *p* = .004, η^2 = .028. Along with corroborating pairwise ANOVA analyses which show the BIS group was significantly higher than controls, *F*(1, 205) = 4.859, *p* = .029, η^2 = .023, and BIU, *F*(1, 169) = 10.189, *p* = .002, η^2 = .058, groups, whilst the difference between controls and BIU was not significant (*p* = .407).

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Figure 3.4.1, prior to reversal point), the contrast effect was significant, *F*(1, 296) = 7.441, $p < .001$, η^2 = .025. This was corroborated by pairwise ANOVA showing the BIS group was significantly higher than controls, *F*(1, 205) = 7.044, *p* = .009, η^2 = .034, and BIU, *F*(1, 169) = 7.337, *p* = .007, η^2 = .042, groups, whilst the difference between controls and BIU was not significant (*p* = .975).

These effects were not continued into the second phase (grey lines in Figure 3.4.1, post reversal point), as the contrast effect was not significant, *p* = .095. This was explained by a pairwise ANOVA showing the BIS group was not significantly different from the control group (*p* = .37). Whilst the difference between controls and BIU was not significant (p = .293), the difference between the BIU and BIS groups was however significant in phase 2, F(1, 169) = 4.73, p = .031, η^2 = .028.

Phase 1 Convergence. Following the analysis protocol of Experiment 2, a mixed ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1, controlling for Need for Closure. The contrast code was the between-subjects grouping factor, and 10 trial epochs were within-subjects. The consequent interaction between epoch and grouping was significant, F(2, 295) = 15.471, $p < .001$, η^2 = .095, indicating a convergence of the contrast effect over the course of the phase[18].

Posteriors. The posterior measure of estimated probability distribution (see Figure 3.4.2) showed a significant contrast effect of condition, as found in the above behavioural measures, F(1, 296) = 6.641, p = .01, η^2 = .022[19].

[18] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, BF_{10} = 9732.711. In the final epoch, there was no longer support for the contrast effect, reaching the $<1/3^{rd.}$ to be classified as strong support for the null, BF_{10} = .265. Both epoch analyses accounted for the effect of Need for Closure as mentioned in the correlates section.

[19] A Bayesian T-test was conducted to confirm the lack of a difference between controls and BIU was not due to insufficient power by looking for strong support of the null (a Bayes Factor of $< 1/3^{rd.}$). Substantial evidence was found in support of the null, BF_{10} = .252.

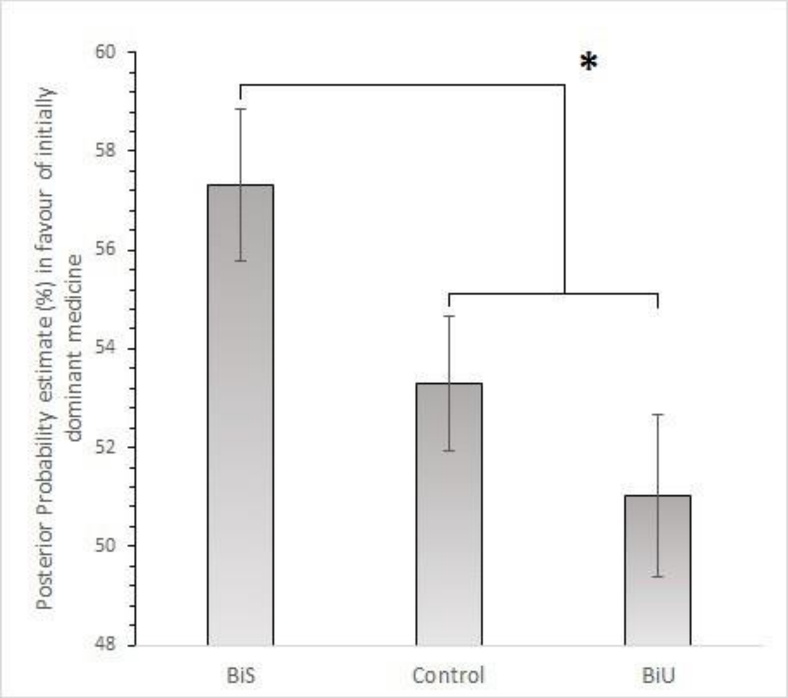


Figure 3.4.2. Experiment 3: Posterior Probability Estimates as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant medicine, less than 50% indicates a preference for the initially suboptimal medicine. Outcomes split by group (* p = .01).

This was corroborated by pairwise ANOVA showing the BIS group was significantly higher than controls, *F*(1, 205) = 4.3, *p* = .039, η^2 = .021, and BIU, *F*(1, 169) = 7.597, *p* = .006, η^2 = .044, groups, whilst the difference between controls and BIU was not significant (*p* = .272).

A further series of t-tests comparing the probability estimates of each group to a 50% expected value revealed a primacy effect, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in the BIS, *t*(76) = 4.788, $p < .001$, 95% CI [4.27, 10.35], and control, *t*(128) = 2.43, *p* = .016, 95% CI [.61, 5.96], groups, but not the BIU group (*p* = .533).

Participant’s binary preference posterior measure yielded a main effect of condition, X^2(1, N = 299) = 6.023, *p* = .049, and showed strong primacy effects, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task (using a Chi squared comparing proportion of preferences to .5), in BIS (*M* = .73, *SD* = .448), X^2(1, N = 77) = 15.909, $p < .001$, and controls (*M* = .6, *SD* = .492), X^2(1, N = 129) = 4.845, *p* = .028, whilst BIU (*M* = .55, *SD* = .5)showed no such effect (*p* = .351). The confidence measure for the binary preference did not show any significant effects of condition, or order effects.

3.4.3 Discussion

The successful replication of the overall contrast effect found in both Experiments 1 and 2 despite the change of domain from gambling to health, speaks to the generalizability of the effects in question. The successful use of novel belief manipulation content, novel use of binary outcome types for experienced evidence, and the medicine prescribing context further validate the robustness of the biasing effects under investigation. However, the authors take pains to note that this is not a perfect replication. Although both overall and phase one choices reflect the patterns seen previously (including the convergence of choices across phase one), the re-emergence of the contrast effect does not quite reach significance in phase 2. It is however equally important to note that the contrast effect re-emerged in the posterior probability estimate, indicating the final judgement regarding the options is (just as in Experiment 2) a consequence of the interaction between belief and initial evidence.

One likely reason for this failure to replicate in phase 2, is the change from variable (and thus more computationally complex) outcomes in the gambling context, to the more obvious binary outcomes of curing or failing to cure in the health context. This can be seen in visual inspection of the trend data shown in Figure 3.4.1, wherein the point of convergence occurs earlier than in Experiment 2 (which is matched on both probability distribution and trial number), within the first half of the phase, rather than by the end of the phase. Further, the proportion of choices in all groups is substantially closer to maximisation (with the latter half of phase one showing all three group averages at about 90% optimal choices). Both of these trends are indicators that learning which option is optimal is easier, a situation that was present in Experiment 1 (which suffered similar ceiling effects likely due to unambiguous probability distributions), and successfully remedied via a subsequent increase in probabilistic ambiguity in Experiment 2. In other words, when first-hand learning is easier (re: less ambiguous evidence, whether through a reduction in computational complexity by going from variable to binary outcomes, as in Experiment 3, or through unambiguous probabilities, as in Experiment 1), the need for and use of communicated beliefs decreases (Ha & Hoch, 1989; Hoch & Ha, 1986).

Need for Closure was found to play a role in explaining some of the variance in choice behaviours, primarily in the first phase of the experiment. This finding fits with the Need for Closure literature and the role it plays in the length and degree of engagement with the process of consolidating on a single hypothesis (Webster & Kruglanski, 1994). To put this another way, those high in Need for Closure more quickly came to a conclusion regarding which option to choose for the remainder of the phase. This finding provides an innovative demonstration of the impact individual differences in Need for Closure can have on extended evidence integration processes. It should further be noted that such effects were independent of manipulation condition, suggesting the consequent biasing effects run independently of this dimension.

3.5 Experiment 4

Experiment 4 set out to replicate the principal findings of Experiments 1, 2, and 3, whilst extending them into a reduced trial format. By reducing the number of trials in each phase, Experiment 4 sought to test the potential limits or thresholds for the consolidation / refutation effects on belief as consequence of the first phase in previous experiments. By seeking to extend such effects to instances where there is substantially less evidence to either refute or consolidate the communicated belief pre-reversal, it becomes possible to address questions regarding potential memory issues in the longer tasks contributing to previously found effects. Finally, such a change opens the avenue for future research into the limitations of previously found belief-initial evidence interactions.

3.5.1 Method

The design and procedure of Experiment 4 followed that of Experiment 3, wherein participants made choices between two medicines in an attempt to discern which is the more effective in curing a disease, some having been pre-exposed to a belief that indicates one of the options as superior. This belief is either supported (BIS) or undermined (BIU) by the first phase of evidence. The probability distribution of outcomes between the two options starts with one dominant (cures 70% of the time) and the other suboptimal (cures 30% of the time), before these probabilities then reverse halfway through the task. These choices are then followed by posterior measures assessing participants end-of-sequence, overall judgements regarding the two options. Finally, these are followed by demographics and the Need for Closure scale.

The principle change in methodology from Experiment 3 to Experiment 4 was that the total number of trials pre- and post-reversal were halved, resulting in two phases of 50 trials. Such a change also meant the pay scale for the task needed to be adjusted downwards to reflect a bonus scheme proportional to the reduced length of the task.

Hypothesis. The central hypothesis follows that of Experiments 2 and 3, namely a replication of the contrast effects found in both choice data and posterior judgements. To once again spell this out explicitly, those who receive a belief that is initially supported (BIS) will choose the medicine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BIU) will not show such a bias and be no different from the control group. Accordingly, these effects are hypothesised to extend to this reduced trial format.

Table 3.5.1 below summarises the key information for Experiment 4. Importantly, entries that differ from Experiment 3 are italicised.

Table 3.5.1: Experiment 4: Summary Table of Task Setup and Measures.

Setup / Manipulations	**Description**	**Details**
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	"Your objective is to try and cure as many patients as you can by learning the effectiveness of the medicines. Each successful cure will earn you points. The medicines both cost 1 point to take. At the end **your score will determine your bonus in dollars as follows:**"
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.] *Note: This was rescaled to reflect the shorter number of trials.*	**Total Payment, based on points:** *<50 points = standard $.50,* *>50 points = $.60,* *>100 points = $.75,* *>150 points = $.90,* *>200 points = $1.05,* *>250 points = $1.20,* *>300 points = $1.40*
Belief Manipulation	Online comment section of "previous participants". Shown to participants before trials under the guise of an interest in their thoughts. Example comments shown to participants controlled by experimenter, with 3 comments indicating the same directional belief (manipulation), and remaining **7** comments of a neutral nature.	**Manipulation Comments**: "I think medicine Z was the most effective", "medicine Z was better than K.", "seemed like Z was better to me!" (last comment required scrolling down the screen). **Example Control / neutral comments**: "I tried to win as much as I could", "Good fun, thank you!"(see Appendix A.1.4 for Sample comment screen)

Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between two medicines, Z and K. *100 trials*, 70% cure rate Z, 30% cure rate K for *first 50 trials*, with reversal of probabilities for *second 50*.	n/a	*Z, K*
Binary Preference	Posterior Measure: Following trials, participants were asked which of the two machines they preferred.	"Which medicine do you think is better?"	*Z, K*
Confidence in Binary Preference	Posterior Measure: Having given their binary preference, participants were asked how confident they were in their preference.	"How confident are you in this preference?"	0-100 slider (default value of 0)
Probability Estimate	Posterior Measure: Participants were asked to give a probability estimate of the distribution of cures between the two medicines.	"What is the distribution of cures between the two medicines?"	100% *Z*, through 50/50, to 100% *K* (slider, default value 50/50)
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one medicine?"	Z, K, No

3.5.2 Results

Descriptives and Processing. Based on a reduced phase contrast code analysis of Experiment 3, a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 3. Converting partial eta squared values for the contrast code analyses for the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 80 per group. The group sample sizes were increased to 90 to compensate for those failing manipulation checks (adjusted on the basis of the change to a shorter paradigm), resulting in a total sample size of 270.

Participants were recruited online using MTurk. Those who had taken part in the previous experiment were ruled out from participating. The 270 participants recruited were US based, randomized into either the BIS (105), BIU (98) or control (68) conditions. The average age was 32.44 years (*SD* = 10.54) and the sample was 39.1% female.

After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis[20], leaving 91 in the BIS group and 76 in the BIU group. The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

The following analyses were conducted using the remaining 229 participants, with an average age of 32.24 years (*SD* = 10.194) and 38.9% female.

[20] Following the same protocol as Experiment 2, a significant effect was found for both the contrast code analysis of the overall choices, $F(1, 268) = 24.114$, $p < .001$, $\eta^2 = .083$, and posterior probability estimates, $F(1, 268) = 8.024$, $p = .005$, $\eta^2 = .029$, regardless of manipulation check removal.

Correlates. Need for closure, age and gender variables were not correlated with any of the dependent variables.

Choice Data. As can be seen in Figure 3.5.1, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant medicine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant medicine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

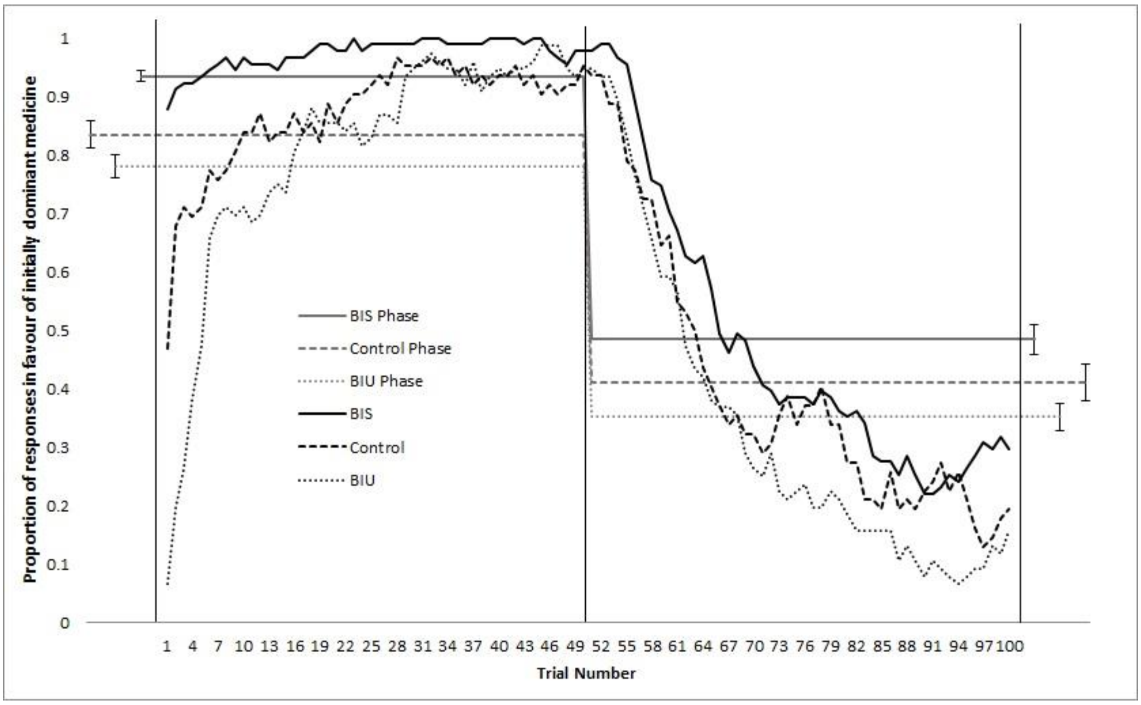


Figure 3.5.1. Experiment 4: Choice Data (Trend Lines with Phase Overlay). Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant medicine as participants move through the 100 trials (with reversal occurring at 50 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

To assess whether Experiment 4 replicated the key findings of Experiment 3, the same contrast code analysis procedure was conducted for the three conditions (BIS, Control, BIU: 2, −1, −1), along with pairwise comparisons between groups.

The contrast code analysis of condition on the overall proportion of choices made was significant, $F(1, 226) = 34.675$, $p < .001$, $\eta^2 = .133$. Along with corroborating pairwise ANOVA analyses which show the BIS group was significantly higher than controls, $F(1, 152) = 12.833$, $p < .001$, $\eta^2 = .078$, and BIU, $F(1, 166) = 46.12$, $p < .001$, $\eta^2 = .218$, groups, whilst the difference between controls and BIU was also significant to a lesser degree $F(1, 137) = 4.769$, $p = .031$, $\eta^2 = .034$.

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Figure 3.5.1, prior to reversal point), the contrast effect was significant, $F(1, 226) = 41.796$, $p < .001$, $\eta^2 = .156$. This was corroborated by pairwise ANOVA showing the BIS group was significantly higher than controls, $F(1, 152) = 20.682$, $p < .001$, $\eta^2 = .12$, and BIU, $F(1, 166) = 60.302$, $p < .001$, $\eta^2 = .268$, groups, whilst the difference between controls and BIU was not significant ($p = .073$).

These effects were continued into the second phase (grey lines in Figure 3.5.1, post reversal point), as the contrast effect was again significant, $F(1, 226) = 10.354$, $p = .001$, $\eta^2 = .044$. This was explained by a pairwise ANOVA showing the BIS group was significantly higher than the BIU group $F(1, 166) = 13.688$, $p < .001$, $\eta^2 = .077$. However, the difference between controls and BIS, and controls and BIU groups were both not significant ($p = .078$, and .131 respectively).

Phase 1 Convergence. Following the analysis protocol of Experiment 3, a mixed ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1. The contrast code was the between-subjects grouping factor, and the five 10-trial epochs were the within-subjects factor. The consequent interaction between epoch and grouping was significant, $F(2, 226) = 33.191$, $p < .001$, $\eta^2 = .0227$, indicating a convergence of the contrast effect over the course of the phase[21].

Posteriors. The posterior measure of estimated probability distribution (see Figure 3.5.2) showed a significant contrast effect, as found in the above behavioural measures, $F(1, 226) = 8.460$, $p = .004$, $\eta^2 = .036$[22].

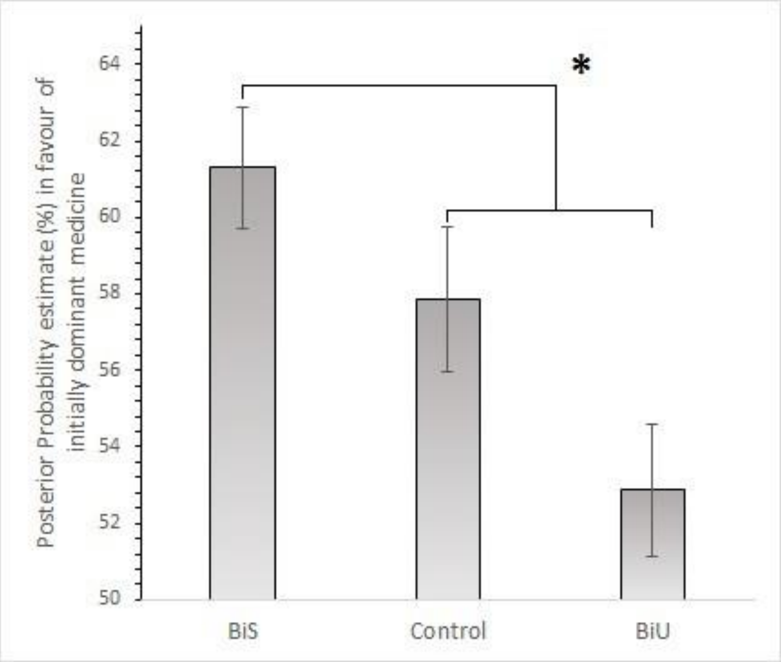


Figure 3.5.2. Experiment 4: Posterior Probability Estimates as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant medicine, less than 50% indicates a preference for the initially suboptimal medicine. Outcomes split by group (* p <.01).

[21] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, $BF_{10} = 1.581*1011$. In the final epoch, there was no longer support for the contrast effect, but this did not reach the <1/3rd to be classified as strong support for the null, $BF_{10} = 2.058$. However, it should be noted that a Bayesian ANOVA on the subsequent epoch (first epoch post-reversal) indicated a reversal in trend from convergence to divergence upon reversal, as strong evidence is once again found for the contrast effect, BF = 6.951

[22] A Bayesian T-test was conducted to assess the null effect between controls and BIU. Although there was a null effect, $BF_{10} = .997$, this did not reach the $< 1/3^{rd.}$ recommended Bayes Factor for strong support of the null.

Pairwise ANOVA showed the BIS was significantly higher than the BIU group, $F(1, 166) = 12.849$, $p < .001$, $\eta^2 = .072$. The differences between controls and BIS, and controls and BIU groups were not significant ($p = .167$ and .055 respectively).

A further series of t-tests comparing the probability estimates of each group to a 50% expected value revealed a primacy effect, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in the BIS, $t(76) = 4.788$, $p < .001$, 95% CI [.814, 14.44], and control, $t(128) = 2.43$, $p = .016$, 95% CI [4.04, 11.64], groups, but not the BIU group ($p = .533$).

Participants' binary preference posterior measure yielded a main effect of condition, $X^2(1, N = 229) = 10.626$, $p = .005$, and showed strong primacy effects, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task (using a Chi squared comparing proportion of preferences to .5), in BIS ($M = .84$, $SD = .373$), $X^2(1, N = 91) = 40.89$, $p < .001$, controls ($M = .77$, $SD = .422$), $X^2(1, N = 62) = 18.645$, $p < .001$, and BIU ($M = .62$, $SD = .489$) groups, $X^2(1, N = 76) = 4.263$, $p = .039$.

The confidence measure for the binary preference showed a significant contrast effect (BIS $M = 63.25$, $SD = 18.533$; BIU $M = 56.01$, $SD = 19.708$; Controls $M = 56.55$, $SD = 22.746$), $F(1, 226) = 6.549$, $p = .011$, $\eta^2 = .028$, which was corroborated by pairwise analyses.

3.5.3 Discussion

The results of Experiment 4 replicate the effects found in Experiments, 1, 2, and 3, across all choice measures (overall, phase 1, and phase 2) and posterior probability estimates. This pattern of results demonstrates once again the importance of initial evidence in the validation of a communicated belief, and the subsequent integrative confirmation bias effects that occur as a consequence of consolidation. Furthermore, the replication of this pattern in Experiment 4 demonstrates that the threshold amount of evidence for consolidation or refutation of a communicated belief is lower than previous experiments have indicated. Such a finding raises questions regarding the limitations of these initial evidence effects for further research. In particular, the differences in biasing effects when reducing *amount* of evidence versus *clarity* of evidence.

However, as can be seen from visual inspection of Figures 3.5.1 and 3.5.2, for the second phase choices and posteriors measures, although the contrast effect is significant, the difference between controls and the BIU group, although non-significant, does not show strong support for the null. Consequently, this finding suggests the potential proximity of the aforementioned threshold for belief refutation.

A further novel finding from Experiment 4 is the extension of the contrast effect to posterior confidence. Interestingly, this pattern demonstrates that those who show the most pronounced bias, i.e. the BIS group, also retain the highest degree of confidence in their [biased] choices and judgements.

3.6 General Discussion

Through the use of on-line paradigms, participants played various lottery gambles and medical prescription tasks, choosing between two options, experiencing outcomes of both (chosen and counterfactual). Prior to this, participants were exposed to an on-line comment section that either contained entirely neutral comments (control group), or additionally contained a communicated belief regarding one of the two options being better. These beliefs were either initially supported (BIS) or undermined (BIU) by the subsequent probabilistic evidence (probabilities then reversed halfway through the task, rendering the options equally profitable overall). As such, the focus of this research has been to investigate the role of communicated beliefs in biasing the integration of evidence. Exploratory analysis from Experiment 1 found that a communicated belief required initially supporting evidence to procure a bias: When evidence changed, only those with a supported belief showed a re-emergence of confirmation bias even though evidence no longer favoured the belief. This reluctance to abandon the belief was not evident in participants who receive a belief that was initially undermined. This effect was then replicated in Experiment 2, and further shown to extend beyond choice data to posterior judgements. This pattern of results was then replicated using a health context (Experiment 3) and with shorter phases (Experiment 4). Although both binary preference and probability estimate posteriors showed strong overall primacy effects, in the latter case the interaction between belief and initial evidence lead to an exacerbation of the influence of initial evidence on the final judgements regarding the two options in the group for which the belief was initially supported. Furthermore, the design of the tasks makes it unlikely that biasing effects are driven by directional motivations to confirm the belief (e.g., demand characteristics). Moreover, the presence of counterfactual evidence assists in reducing selective exposure explanations.

The extension into the health domain (Experiment 3) highlights the robustness of the effects given changes to belief content, outcome types (variable to binary) and decision context (curing patients versus winning gambles). Such a finding strengthens the generalizability of the belief biasing mechanism proposed. Furthermore, the successful replication of the contrast code in the reduced trial number format of Experiment 4, speaks to the strength of the effects found. However, the pattern of results also raises interesting questions regarding the limitations of a refutation period in belief-evidence interaction effects. In other words, the finding in Experiment 4 that a linear contrast code is also a good fit for the pattern of results (i.e. there is starting to be a difference from controls in the choice / judgement data in the Belief Initially

Undermined group) suggests the reduced amount of phase 1 evidence (100 down to 50 trials) is starting to impact belief refutation.

In contrast to most studies investigating confirmation bias effects, a novel manipulation was used in which additional motivations for confirmation (such as self-concept preservation and authority effects) were absent. Such an absence was due to the anonymous, on-line peer source of the directional beliefs regarding the two options, along with the seemingly incidental presentation of such information (no explicit instruction was given regarding the content of the comment section, and filter questions were used to identify and remove any participants that saw through this manipulation). Given the reduction of these traditional motivational explanations for confirmation bias, the fact an effect was found regardless points towards an alternative explanation.

Rather than motivational explanations, the current findings are better explained by an extension of cognitive explanations taken from research investigating self-generated hypotheses (Fischhoff & Beyth-Marom, 1983; Pyszczynski & Greenberg, 1987). The proposed mechanism behind the subsequent bias is that the communicated belief orients subsequent evidence as either confirming or contradicting it, in line with work on the constriction of the hypothesis space (Fischhoff & Beyth-Marom, 1983; Klayman, 1995; Nickerson, 1998). To explain further, those who receive a communicated belief start out with the hypothesis: “either this belief is true, or not”, whilst controls have no specific hypothesis to entertain. Such an assertion is supported by the naïveté of participants to the context in which evidence is evaluated, distinguishing this finding from work in which either evidence exposure (Harvey & Fischer, 1997; Yaniv, 2004; Yaniv & Milyavsky, 2007) or prior opinion (DeMarzo, Vayanos, & Zwiebel, 2003; Lord et al., 1979; Pitz, 1969) precedes second-hand information. This naïveté results in controls (i.e. those who receive no belief) starting with a neutral prior, whilst belief-recipients are informed (but not bound by) the communicated belief. In this way, evidence-exposure alone both forms a new hypothesis in controls, and validates or refutes communicated beliefs in recipients. Thus, subsequent deviations in learning can be attributed to the interaction of a communicated belief with the same evidence (as controls receive), rather than the additional (unaccounted for) interplay of prior beliefs and opinions which both communicated information and evidence integration will both be informed by (i.e. beliefs are not having to persuade the recipient away from a pre-existing informed prior).

Further, given the absence of cues to the credibility of the source (Briñol & Petty, 2009; Hahn et al., 2009), this hypothesis validation is, by inference, a validation of the source as credible or not. As such, initial evidence updates both of these likelihoods (the likelihood the belief is true, and in relation, the likelihood the source is credible), acting as a consolidation period. As such, those entertaining a hypothesis who experience initially undermining evidence for the belief not only consider the belief invalid, but the source (by proxy) as unreliable. In doing so, such individuals match the control group (i.e. a belief that has not been consolidated is equivalent to starting without a belief). Alternately, those who receive a belief that is confirmed by initial evidence would now have a consolidated hypothesis, and evaluate subsequent evidence in the forwarded integrative account of confirmation bias (Decker, Lourenco, Doll, & Hartley, 2015; Doll, Jacobs, Sanfey, & Frank, 2009; Klayman, 1988, 1995; MacDougall, 1906; Staudinger & Büchel, 2013; Whitman et al., 2015). Such an evaluation stage of communicated knowledge fits within a propositional account of learning, in which the belief (or proposition) has its truth value either confirmed or refuted (Mitchell et al., 2009).

The effects outlined in the present work we propose to have occurred due to an active, prolonged, biased integration of evidence to favour confirmation. The methodological design and results found in these experiments allow us to rule out several alternative explanations. First of all, the probabilistic reversal has shown that across all experiments, all groups are actively learning throughout (see the sharp change in trend lines post-reversal in Figures 3.2.1, 3.3.1, 3.4.1 and 3.5.1 for Experiments 1, 2, 3 and 4 respectively). That is, as all groups remain sensitive to the change in probabilities, it is possible to rule out a differential in inattentiveness as an explanation of the bias found in the Belief Initially Supported (BIS) group. Secondly, the convergence of groups across phase one (the control and the Belief Initially Undermined; BIU groups, moving towards the BIS group) is then followed by a resurgence of the bias in phase two, indicating ongoing, active use of the communicated belief in the BIS group. To explain further, the convergence of groups over phase 1 to a null difference should have "washed out" the effect of the communicated belief (i.e. the BIS difference should be extinguished by the point of reversal) if the communicated belief solely acts as a starting point. Thirdly, this convergence-divergence pattern rules out a general conservatism explanation for group differences (Phillips & Edwards, 1966), where bias is dictated by the difference in starting point only (and all participants simply underweight all subsequent evidence). Similarly, such a pattern, which indicates that the communicated belief does not act as a strong prior, is further corroborated by the asymmetry observed in the BIS and BIU groups. A strong prior account would predict the BIU group should be unaffected by early evidence and show a similar bias effect in the opposite direction to the BIS group. Instead the BIU group does not show any difference from the control group (and it should be highlighted here that all BIU group participants used in both experiment's analyses could still *remember* the belief at the end, and as such their similarity to controls was not a consequence of forgetting the manipulation), with an almost immediate drop off from the belief-dictated starting point in all experiments. Accordingly, our remaining explanation is of an active, ongoing bias in evidence integration, necessitated by initially supportive evidence.

The immediate refutation (or consolidation) of the communicated belief indicates the critical role of initial evidence, which rather than dominating or being dominated by the communicated belief, instead interacts with it (as in Weiss-Cohen et al., 2016). The presence of a bias in the BIS group (relative to the control group) and absence of bias in the BIU group indicates the belief must first be consolidated before biasing effects can occur beyond initial preferences. This role of initial evidence in either the consolidation or rejection of communicated beliefs has strong parallels to primacy effects found in other areas of psychology such as subjective probability learning (Peterson & DuCharme, 1967), causal judgements (Dennis & Ahn, 2001; Fugelsang & Thompson, 2003), and impression formation in social psychology (Anderson, 1965; Freund et al., 1985; Kruglanski & Freund, 1983; Mann & Ferguson, 2015), where early evidence plays a critical role in the formation of opinion. However, this work differs from traditional primacy (Hogarth & Einhorn, 1992) in one key way: Although in the current experiments all experimental groups showed traditional primacy effects (as indicated by the overall proportions favouring the initially dominant option, as well as the posterior measures in Experiments 2, 3, and 4), those who had received a communicated belief showed even greater susceptibility to early evidence (Staudinger & Büchel, 2013). This is demonstrated by the *re-emergence* of the bias in the initially supported belief condition (BIS) post-reversal, even after choices had converged prior to reversal. Further, at the point during the task of posterior measures, participants had seen the entirety of the evidence for both options (which were equal overall), yet the same pattern of bias was still found. This not only demonstrates the communicated belief had not been diluted by the first-hand evidence, but lends greater support to the potency of the interaction between the communicated belief and initial evidence, and the proposed consolidation, followed by integrative bias mechanism.

Once a belief had been consolidated, confirmation bias effects were found despite the presence of 2-sided evidence, ruling out selective evidence exposure as an explanation for these confirmation bias effects (Doherty & Mynatt, 1990; Doherty et al., 1979; Klayman, 1995; Klayman & Ha, 1987). It is expected however, that in environments in which the individual can be selective in what they see (e.g., always selecting to take the medicine, so that the alternative of the disease disappearing without taking a medicine is never seen), the consolidation threshold would likely be crossed quicker (Blanco et al., 2014; Doherty et al., 1979; Yarritu, Matute, & Vadillo, 2014). Similarly, it is expected that factors that increase participants' trust in the source of information (of which increasing the number of anonymous comments was one) will increase the degree of belief uptake (Yaniv & Kleinberger, 2000; Yaniv, 2004), along with factors that provide supplementary motivations for confirmation (Kunda, 1990).

Despite the presentation to 2-sided evidence, in principle we cannot rule out a selective attention explanation, wherein people attended more to evidence that supported their belief, while missing the contradictory information. Such selective attention could therefore be considered a form of positive test strategy (Doherty et al., 1979; Klayman, 1988; Navarro & Perfors, 2011). Nevertheless, if a form of attentional selectivity to evidence were the mechanism at work in our current experiments, one might expect delayed learning post-reversal in the BIS group, whereby it would take a larger number of trials before choice proportions started moving away from their phase 1 choice proportions. Instead, all groups rapidly moved away from their phase 1 proportion of choices upon reversal. Furthermore, similar delayed learning would be expected to amplify the effect of the prior in the BIU group during phase 1. To the contrary, we found a rapid negation of belief dependent on initial evidence. In other words, if participants were selectively attending to one of the options, negative outcomes for this option might, in isolation, be interpreted by the participant as short-term fluctuations (and thus not be worth switching away from). However, in the case of attending to both options, the negative outcomes in the selected option would be seen to co-occur with the counterfactual positive outcomes in the alternative, more rapidly indicating a change in the choice environment (resulting in a more immediate change of selected option), which is in line with the trends in choice data for all experiments. Although we cannot fully rule out selective attention in our current research, follow up lab studies could employ eye-tracking to assess what evidence participants are attending to, detecting possible asymmetries.

It is also important to note that effect sizes in the earlier experiments in the present paper are not large. However, it should be noted that as the methodology progresses through to latter experiments, effect sizes continue to increase. Such a trend is likely attributable to reducing sources of variance (e.g., from a gambling context that may invite more speculation and short-term fluctuations, to a more stable medical context with simpler outcomes, and from longer to shorter phases). Furthermore, given the purpose of the experiments presented within the present work has primarily been to make a theoretical point, and the paradigm itself is admittedly noisy, the small effect sizes may be an underestimation of real world instances of the biasing effects demonstrated. Further research is, however, needed to determine the extent of such a bias.

Two limitations pertaining to this work bear particular relevance for the extension of research investigating communicated belief effects. Firstly, when trying to ascertain whether an incidental belief has been noticed by participants, the use of a memory based manipulation check does carry a limitation: the criterion for passing applies greater scrutiny to the manipulation groups than the control group. In other words, an inattentive manipulation group participant would have been picked up by failing to remember the communicated belief, whilst those in the control group (who may be similarly inattentive), when answering that they had not seen anything, would have been marked as correct. This could result in an asymmetry in the number of (and quality) of participants in each group. However, as indicated within each experiment, when including all participants, main effects remain (just with greater variance), and further investigations on manipulation check failures found them no different from the control group. Importantly, as mentioned previously, this check (and removal of "forgetters") provided insight into the similarity between controls and the BIU group, showing this was not due to those in the BIU group forgetting the manipulation by the time the reversal had occurred.

Secondly, the use of MTurk has known shortcomings, notably the lack of experimental control (Goodman, Cryder, & Cheema, 2013). Although, several studies have shown strong replications of effects found in laboratory experiments using the site (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). The advantage of such a method is its greater ecological validity, as participants show better representation of demographics relative to the majority of psychological study populations (Henrich, Heine, & Norenzayan, 2010). From the context of demand characteristics as an additional, unwanted motivation for belief confirmation, the fact participants did not see the manipulation as experimentally driven added real world applicability, and distinguishes this work from more artificial, lab-based assessments of integrative confirmation bias (Staudinger & Büchel, 2013). Further, there is a growing relevancy in having a communicated belief manipulation that uses an on-line "comment section" in its natural setting, given such a context represents an ever-expanding and abundant source of communicated beliefs in the modern world.

The demonstration of the role of initial evidence as pivotal in the adoption and maintenance of a belief has important implications given the real-world consequences of such an effect. Research into placebo effects, for instance, relies on the patient's belief that a treatment will work, in spite of (unknown to the patient) the absence of a medically active ingredient. Typically, this research has investigated possible cues (setting, verbal suggestions), motivations (such as demand characteristics from doctor to patient) and "active agents" (such as caffeine) used to mimic a medical side effect (Wickramasekera, 1980) as highly impactful to the placebo effect (Wager & Atlas, 2015). The implication of the current research is to focus on providing initially supporting evidence to increase placebo efficacy. For example, by deploying treatments that start with a medically active component, initially supporting evidence (e.g., by decreasing symptoms) is provided for the belief of the patient that (s)he is in the treatment condition. Once the belief is consolidated the medically active component can be phased out of treatment in favour of a pure placebo, whilst retaining efficacy.

The effects discussed in this work have broad value for other research that typically focuses on either beliefs or evidence. For example, work in stereotyping and impression formation has shown sensitivity to early evidence in the formation of negative assessments of character (Anderson, 1965). The effects found here suggest that the combination of a communicated belief (for example, a negative stereotype) and random fluctuations in early evidence (i.e. there will be a portion of recipients who happen to receive initial confirmation of this belief), results in an asymmetry that may help explain such beliefs prevalence despite their inaccuracy. Homeopathy, for example, may benefit from such random fluctuations facilitating effective placebo responses, as any short-term fluctuations in symptoms upon receiving a treatment (more likely than any sustained improvement) will have a disproportionate effect on the believed efficacy. Such an effect can also be extended to other areas of research where there is an interplay of beliefs (or "priors") and evidence, including consumer research (Ha & Hoch, 1989; Hoch & Ha, 1986), decision making (Nurek et al., 2014), causal learning (Yarritu et al., 2015; Yu & Lagnado, 2012) and instruction effects (Doll et al., 2009; Mertens & De Houwer, 2016; Roswarski & Proctor, 2003; Van Dessel et al., 2015; Van Dessel, Gawronski, et al., 2016).

Conversely, to prevent the adoption and maintenance of potentially harmful beliefs (including various superstitions, whose deployment comes at a cost, but has no causal relationship to the desired outcome), the importance of initial evidence should be included alongside motivational and evidence exposure factors when determining interventions or preventative strategies. For example, in the realm of gambling, the combination of a directional belief regarding outcomes and various misconceptions regarding short-term evidence (Gilovich et al., 1985; Tversky & Kahneman, 1971) can lead to negative consequences. Early interventions during this critical period of belief consolidation, before sensitivity to contrary information is decreased and confirmation biases are embedded, are likely to incur the greatest change in belief outcomes relative to the cost of intervention. Taking this further, it is demonstrated here that interventions to prevent biases that are prefaced on improving the fairness of evidence exposure or removal of conflicting motivations, although likely reductive, may not be sufficient. For example, attempts to reduce racial bias have looked at re-categorization and positive exposure strategies (Gaertner & Dovidio, 2014), and although these interventions have found some success, this research suggests exploring interventions on *early / formational experiences*, such as during childhood, may prove more fruitful.

A notable avenue for further work is exploring how evidence ambiguity affects the interaction with communicated beliefs. The reliance on prior information has been found to be directly related to the degree of ambiguity in experienced evidence (Abbott & Sherratt, 2011). This bears particular relevance to research on superstitious responding (Ono, 1987; Rudski et al., 2012), wherein erroneous beliefs (termed "superstitious" due to their mismatch with objective evidence) are seen as a consequence of misinterpretations of first-hand evidence, such as pseudo-contingencies (Fiedler & Freytag, 2004; Yarritu et al., 2013, 2014). Our research suggests that unsupported beliefs can be transferred second-hand, dependent on initial evidence interactions, which lays the groundwork for describing the core mechanisms behind the adoption, maintenance and propagation of popular superstitious beliefs in an ecological model.

CHAPTER 4: FALSE BELIEFS IN MEDICAL DECISION MAKING: THE IMPACT OF INITIAL EVIDENCE AND COUNTERFACTUALS

The capacity to communicate knowledge from one individual to another has allowed humans to excel in the navigation and mastery of our environment. However, unlike many animals that share an – albeit limited – capacity to communicate, humans have finessed this capability through the development of formal language, and ever more interconnected technologies. Unfortunately, with the bounteous advantage of instantly accessible knowledge communicated by millions of other individuals, comes the cost that such knowledge is not always truthful. Whether this is due to ill-intent, or human error on the part of the communicator – such that their assessment of the environment in question is flawed – the ramifications, given such beliefs can be spread rapidly to countless recipients at ever increasing rates, are of grave importance. For example, the persistence of the belief that vaccines cause autism, despite an absence of evidence for such a link, is attributed as responsible for rising rates in preventable, damaging diseases such as measles across the western world. The natural question provoked by such phenomena is: given the unsupported nature of such beliefs, how are they adopted and maintained by recipients?

In answering this question, previous research has demonstrated that recipients of such communicated beliefs integrate subsequent first-hand evidence in an asymmetrical manner that favours confirmation of the belief (Chapter 3). These confirmation bias effects found in Chapter 3 were however dependent upon order effects, wherein an anonymously sourced belief required an initial period of supporting evidence for consolidation, with early undermining evidence resulting in refutation of the belief. However, whilst this previous work looked at extended periods of evidence exposure, questions remain regarding the amount of evidence required for biasing effects to occur. The present chapter builds off Chapter 3 by moving from sustained periods of evidence (which have allowed for greater scrutiny into the mechanisms behind the confirmation bias effects) to more generalizable, short-term fluctuations in evidence. Such an adaptation allows for implications to be drawn from associated literature including placebo research (Schafer, Colloca, & Wager, 2015; Schwarz & Büchel, 2015; Wager & Atlas, 2015), impression formation (Anderson, 1965) and implicit evaluation revision (Cone & Ferguson, 2015; Mann & Ferguson, 2015).

Whether given a belief about a new medicine being effective (but actually a placebo), an individual (incorrectly) possessing a negative character trait such as being quick to anger, or a (bogus) gambling strategy being effective, this research indicates the first few pieces of evidence are critical. If it supports such a belief, such as a decrease in symptoms during the first couple of days, an initial meeting in which the individual is angry, or some early success on the first attempts at using the gambling strategy, subsequent evidence is then interpreted in a confirmatory manner which maintains the beliefs – even if overall evidence should lead to refutation. Short-term fluctuations in evidence have been shown to lead to misinterpretations of an environment as a consequence of the law of small numbers (Bar-Hillel & Wagenaar, 1991; Tversky & Kahneman, 1971), wherein the deviating short-term pattern is thought to be representative of the long-term trend. These more common, short-term deviations have yet to be applied to how erroneous communicated beliefs are adopted and maintained, but in so doing provide a novel, real world generalizability to the effects in question.

Additionally, defining beliefs as generic (i.e. unquantified, but directional) statements (Cimpian et al., 2010; Leslie, 2008), allows for an account of beliefs as propositional statements regarding associations between actions and outcomes (Mitchell et al., 2009), whereby such a statement can have its “truth value” evaluated in terms of (dis)confirmation. Further, such an “unquantified” form of belief content fits with real world natural communication of (erroneous) beliefs (Gilovich, 1993). Work in argumentation, risk communication and persuasion literatures have illustrated the impact on argument strength from various source factors (Briñol & Petty, 2009; Hahn et al., 2009). These factors have included perceived source attractiveness (Kelman, 1958), expertise (Goodwin, 2011; Sniezek & Van Swol, 2001), and trustworthiness (Siegrist et al., 2005; Twyman et al., 2008), which invoke a myriad of both cognitive and motivational reasons for maintaining an erroneous belief. This led to the stripping away of such factors (and with them the associated alternative bias explanations) in Chapter 3, to focus solely on the first-hand evidence basis for belief maintenance in terms of order, quantity, context and clarity.

Regarding these evidence factors, in Chapter 3 initial evidence was manipulated using extended probabilistic reversal learning tasks, in which evidence initially favoured one of two options (gambles / medicines) – for example, a 70% (dominant) option vs. a 30% (sub-optimal) option. Halfway through the evidence integration period, the probabilities of the two options were reversed. Results showed that although communicated beliefs predicted initial choices, those who received a belief that was undermined by the first phase evidence distribution were not different from controls (who received no belief) across all choice and posterior judgements. Conversely, those who received a belief that was initially supported not only showed a significantly higher proportion of choices in favour of the belief-indicated option pre-reversal, but also – despite all groups converging in the proportion of choices by the point of reversal – showed similar deviations in favour of their belief-indicated option from the other groups when evidence changed to no longer support it. Such effects extended beyond choice data to posterior judgements of the options.

Critically, the effects found in Chapter 3 were derived from a methodology designed to rule out various explanations for confirmation bias effects. Confirmation bias is a broad category of biases (Hahn & Harris, 2014; Klayman, 1995; Nickerson, 1998) that may be broken down into two psychological "elements": cognitive mechanisms and motivations. The latter of these, termed here as motivated reasoning, can be distilled into two, somewhat opposing motivations when applied to biases. Firstly, the motivation for accuracy can be thought of as a motivation to find the *correct* answer, with associated effects including deeper processing (Tetlock, 1983a, 1985a) and higher quality of inferences (Kunda, 1990). However, the benefits of such a motivation are restricted to the avoidance of biases based on *hasty* reasoning (Kruglanski & Freund, 1983), and do not absolve the individual of other forms of faulty reasoning. The second motivation of interest is *directional*, which can be thought of as a motivation to find the *right* (i.e. most beneficial to the values of the individual) answer (Hahn & Harris, 2014; Kunda, 1990). Such motivations are associated with greater deployment of sub-optimal cognitive strategies, such as the use of selective search, so as to – for example – prevent dissonance and preserve a coherent or positive self-concept (Cialdini & Goldstein, 2004; Festinger, 1962; Kusec et al., 2016). By using incidentally communicated, anonymously sourced beliefs to avoid experimenter demand effects (Hertwig & Ortmann, 2001), and incentivising participants to be as accurate as possible via performance based pay (and using an incidentally communicated belief), we sought to rule out such motivation-based explanation of bias.

The cognitive account of biases can be similarly sub-divided into two camps: first order, or *input* / *exposure* based bias, and second order, or *integration* based bias (Hahn & Harris, 2014; MacDougall, 1906). First order strategies include selective search (Jonas et al., 2001; Lord et al., 1979) and positive test strategies (Hendrickson et al., 2014; Klayman & Ha, 1987; Navarro & Perfors, 2011) wherein the forgone, or counterfactual, event is never seen by the individual (e.g., if under the impression blowing on the dice yields a better roll, the individual not only then tends to blow on the dice before each roll, but in doing so forgoes the chance to see the outcomes without blowing). When such strategies are possible, it is difficult to discern whether the consequent bias is due purely to the input strategy, or if biased integration is also occurring (integration is still required even if the exposure is biased). To address this, we made counterfactuals present on each trial (if the participant selected option A, they still saw what they would have got had they chosen option B for that trial), in doing so worked to rule out first order explanations (Chapter 3).

Second order, or integration based cognitive biases have been studied within confirmation bias effects typically in the evaluation of ambiguous arguments (Klayman, 1995; Lord et al., 1979; Nickerson, 1998) and more recently in the evaluation of probabilistic evidence over time (Chapter 3). The mechanism can be described in terms of an *overweighting* of confirmatory information (MacDougall, 1906), or the *underweighting* of non-confirmatory information (Nurek et al., 2014). Such a mechanism has recently received support from neuroscience work on the natural asymmetry in updating signals when evidence is encountered that matches a focal hypothesis / pattern, as opposed to evidence that does not match (Whitman et al., 2015). Accordingly, the mechanism forwarded in Chapter 3, and further ratified in the current chapter, is that a communicated belief, once adopted by a receiver, acts as a focal hypothesis that invokes this asymmetrical (or "biased") updating to support it. As found previously however, order effects in evidence thus play a critical role in the adoption / consolidation of the belief, or its refutation.

4.1.1 Present Research

The present chapter seeks to build off previous work investigating the confirmation bias effects of communicated beliefs on evidence integration. Previous work sought to lay the theoretical groundwork for the role of initial evidence in the consolidation and refutation of a communicated belief, leading to sustained, biased integration as a result of the former, using a methodology designed to rule out first-order cognitive explanations, directional motivations and conservatism or strong prior effect (Chapter 3; Phillips & Edwards, 1966). The current chapter seeks to further test assumptions regarding counterfactuals, further explore the role of initial evidence as a "gatekeeper" to bias effects, and address the methodological limitations of previous work. Accordingly, a new paradigm was created, inspired by work in reinforcement learning using abstract shapes (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013), converting this into a medical decision making task and incorporating the communicated belief manipulation used in Chapter 3 prior to probabilistic first-hand evidence exposure.

The present research makes two important methodological changes from Chapter 3. Firstly, the control comparison to the "belief" condition is folded within-subjects. In other words, *all* participants receive a communicated belief. Such a belief is therefore directed towards one of the medicines of *one* of the two diseases. As such, because all evidential factors are equal between the two diseases (i.e. each has its own "dominant" medicine that works probabilistically more often than the "suboptimal" option, with the same number of trials and initial evidence manipulations), the impact of the belief can be measured as the difference between choices / judgements of these two diseases, avoiding between-subject variance in the comparison. Secondly, the manipulation of initial evidence, instead of lasting for an extended period of time (Chapter 3), is instantiated in the first couple of trials. By using such a short-term, immediate fluctuation of either supporting or undermining evidence, a more realistic, real world parallel can be drawn to other domains in which there is interplay between second and first-hand evidence, including placebo effects (Schafer et al., 2015; Schwarz

& Büchel, 2015; Wager & Atlas, 2015), impression formation (Anderson, 1965) and consumer research (Ha & Hoch, 1989; Hoch & Ha, 1986).

Finally, although the majority of research into confirmation bias has either allowed for first-order cognitive biases via selective exposure to evidence (Addario & Macchi, 2012; Doherty et al., 1979; Jones & Sugden, 2001; Matute et al., 2011; Navarro & Perfors, 2011), or has focused on integrative biases in short, often opinion based contexts (Allahverdyan & Galstyan, 2014; Del Vicario, Scala, Caldarelli, Stanley, & Quattrociocchi, 2016), we fixed the availability of counterfactual information to prevent selective exposure explanations in prolonged evidence integration (Chapter 3). However, such a methodological constraint does not fully rule out potential alternative explanations such as selective attention. Consequently, the present work explicitly manipulates the presence or absence of counterfactuals between participants, and in doing so looks to answer questions regarding their impact both on learning within an extended evidence integration paradigm, and more particularly their effect on the belief-initial evidence interaction.

Hypotheses. Given the emphasis on extending the effects found in Chapter3 concerning biased evidence integration as a consequence of communicated belief consolidation, the primary hypothesis is to demonstrate not only main effects of beliefs and initial evidence, but that they interact in a way that the combination of both a belief *and* initially supportive evidence resulted in significantly greater bias in both choice and posterior judgement data. Furthermore, given the manipulation of the presence or absence of counterfactual feedback, it is hypothesised, in line with work on selective exposure (Jonas et al., 2001; Wagner, 2016), that the absence of counterfactuals will impair effective learning.

4.2 Experiment 5

4.2.1 Method

Following the outline set out above, Experiment 5 was designed using a 60-40 probability distribution, in line with previous reinforcement learning literature (Staudinger & Büschel, 2013). Adapting this into a health context, participants had to cure a total of 100 “patients” (trials), with trials alternating between two different fictional diseases (“Lannixis” & “Deswir”). Each disease had its own fictional pair of medicines with which to attempt to cure the disease, “Mox” and “Nep” for “Lannixis”, and “Byt” and “Zol” for “Deswir”. In each pair of medicines, there was an optimal (60% cure rate) and sub-optimal (40% cure rate) medicine.

These trials were preceded by the aforementioned "comment section" for **one** of the diseases, which contained a communicated belief regarding its medicines ("I think the Zol medicine was the most effective"), with the remainder of the comments of a neutral nature (e.g., "nice task", "seemed interesting"). This belief indicated one medicine as superior, but this medicine was in fact (unknown to the participant) **always** the **sub-optimal** medicine within a pair. In this way, every participant was exposed to an erroneous belief for one of the diseases, whilst the other disease (which did not receive a belief) acted as a within-subject control. This allowed for a difference measure between “belief” and “control” diseases, as an assessment of the effect of the comment manipulation.

As in later reversal studies (Experiment 2 of Chapter 3), posteriors were included following trials; a binary preference, a confidence measure in that preference, and a probability estimate (see table 4.2.1 below). However, in the present paradigm, posteriors were included for each disease, so that the difference could be measured to assess the effect of belief.

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 100 prior HITs. Participants completed the experiment under the assumption that the purpose of the investigation was to improve medical decision making. Participants were English speakers between the ages of 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

Design. For each of the two diseases, their two medicine options ("Mox" and "Nep" for the "Lannixis" disease, and "Byt" and "Zol" for the "Deswir" disease), all generated either "Cured" of "No Effect" outcomes, with one medicine for each disease curing at a 60% rate (optimal option), and one medicine curing at a 40% rate (sub-optimal option). The instructions given to participants explained that each trial is a new patient presenting with one of the two diseases, and it is their job to prescribe one of the two medicine options for that disease for the patient, and assess the outcome. Successful cures earned them points (+3), whilst failures to cure cost them 1 point. Participants were further instructed that each patient may react differently to the medicines, so it is their job to discern the *overall* efficacy of the medicine options. Participants were further incentivized through a bonus scheme that earned them consecutively larger bonus payments based on the amount of points earned.

Before starting the trials, as mentioned above, all participants were exposed to a comment section from previous participants regarding **one** of the two diseases (counterbalanced), which contained comments regarding one of the medicines (always the **sub-optimal** option) being better. Which medicines were optimal and sub-optimal were also counterbalanced. In this way, the first factor of interest in the design is the within-subject difference in the choices and judgements for the "control" disease (the disease that did not have a comment section), and the "belief" disease. As the sole difference between the two diseases was the presence or absence of communicated beliefs, any subsequent difference could be discerned as due to this within-subject manipulation.

The second factor of interest followed both from previous experiments within this body of work, and the previous literature from which this paradigm draws inspiration (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013). Groups in the previous probabilistic reversal studies (Chapter 3) either received initial support (Belief Initially Supported; BIS) or were initially undermined (Belief Initially Undermined; BIU), which allowed for the assessment of the role initial evidence plays in belief uptake. This paradigm seeks to further focus on this interaction. Staudinger & Büschel's (2013) paradigm used a highly-localized implementation of initial evidence, which we apply to this current paradigm, the result being two between-subject conditions.

Having received the belief, the first few trials of evidence either support it (initially supportive condition; IE+), or undermine it (initially undermining condition; IE-). To explain further, given that the belief always (falsely) indicated the sub-optimal option as superior, in the IE+ condition, the sub-optimal options for *both* diseases received two positive trials (cures), followed by one negative (no effect), whilst the optimal options for *both* diseases received the opposite pattern; two negative trials (no effect), followed by one positive (cure). Conversely, the IE- condition followed the opposite pattern, with the *optimal* options now receiving two positive trials followed by one negative, and the *sub-optimal* options now receiving two negative trials followed by one positive. In this way, the belief received initially undermining evidence. It is important to highlight that in both conditions, the pattern of evidence (i.e. the distribution of cures and failures to cure across the two medicines for a disease) was identical for *both diseases* (i.e. the sole difference between diseases is the presence of absence of the *belief*). All trials following this three-trial manipulation followed the 60/40 probability distribution outlined above.

The final factor of interest in this design was based on previous research assumptions regarding the presence of counterfactuals (see Chapter 3). Whereas in previous experiments, the presence of counterfactual outcomes has been held constant, in this paradigm their presence or absence was manipulated as a between-subjects factor. This results in a two between-subject factor design: *initial evidence* (either supportive or undermining) and *counterfactuals* (either present or absent). Within-subjects, the difference between the “control” and “belief” diseases allowed for the assessment of the effect of *belief* in both choice and judgement dependent variables.

Procedure. Before starting the trials, participants were shown the "comment section" in which previous participants had written their thoughts regarding one of the diseases, and the task in general. These comments were rigged to appear to be from other MTurk participants (complete with fake MTurk ID numbers), and were mainly of a hypothesis neutral nature ("interesting task.", "good fun, thanks!"). However, three comments indicated a directional hypothesis regarding one of the medicines for the disease ("I think the Zol medicine was the most effective"). As previously mentioned, the medicine indicated as superior, was in fact always (unknown to the participant) the sub-optimal option.

On each trial participants selected which of the two medicines to prescribe to the patient, with “Mox” and “Nep” the two options for when the patient presented with “Lannixis” disease, and “Byt” and “Zol” for presentations of the “Deswir” disease. Which disease was the control and which was the “belief” condition was counterbalanced between participants, as were which medicine in each pair was set to be optimal or sub-optimal. Each trial, the side of the screen on which each medicine was shown was randomized. Participants were invited to earn as many points as possible, based on the number of cures. Each trial cost participants one point, so a failure cure resulted in a net loss of −1 point, whilst a cure earned 3 points. Participants were aware of their current total points earned during the trials, and instructed that their total amount of points directly corresponded to an increasing bonus payment in dollars.

For each trial, participants selected one of the two medicine options to prescribe to the patient, and this generated the outcomes for that trial. The outcomes were "Cured" written in green if the medicine led to a cure, "No Effect" in red if the medicine was unsuccessful. If the medicine was the selected option, then the outcome text was surrounded by a highlight box of the same colour (see Appendix A.2.2 for an example feedback screen with counterfactuals included).

Once all 100 trials (50 per disease, alternating each trial) were completed, participants then completed posterior measures for each of the diseases. This consisted of a binary preference for that diseases pair of medicines, the confidence in that preference, and a probability estimate for the distribution of cures between those medicines. Following this, participants could post an open text response "comment" in the comment section they had seen before the task. Based on the amount of time spent during the task, the comment section was updated with new neutral comments. Upon posting their comment, participants then completed a series of exit funnel questions to assess awareness of the comment section manipulation. This was finally followed by a demographics questionnaire and the Need for Closure measure (Roets & Van Hiel, 2011; Webster & Kruglanski, 1994). Following completion of the task, participants were debriefed and given an email to contact if they had any further questions.

Table 4.2.1 below summarises the key information from the above methodological description. This includes the phrasing of the task instructions, incentives scheme, and belief manipulation, as well as the measures taken (including posteriors and manipulation check question phrasing).

Table 4.2.1: Experiment 5: Summary table of task setup and measures.

Setup / Manipulations	**Description**	**Details**	
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	"*In this task, you will be attempting to cure cases of two different diseases. Each disease has a pair of medicines designed to cure it...* ...Your objective is to try and cure as many patients as you can by learning the **effectiveness** of the medicines. Each successful cure will earn you points. The medicines both cost 1 point to take. At the end **your score will determine your bonus in dollars as follows:**"	
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.]	**Total Payment, based on points:** <50 points = standard $.50, >50 points = $.60, >100 points = $.75, >150 points = $.90, >200 points = $1.05, >250 points = $1.20, >300 points = $1.40	
Belief Manipulation	Online comment section of "previous participants". Shown to participants before trials under the guise of an interest in their thoughts. Example comments shown to participants controlled by experimenter, with 3 comments indicating the same directional belief (manipulation), and remaining **7** comments of a neutral nature. *The comment section was specifically indicated as being for only* ***one*** *of the two diseases.*	**Manipulation Comments**: "I think the Byt medicine was the most effective", "Byt treatment was better than Zol.", "seemed like Byt was better to me!" (last comment required scrolling down the screen). **Example Control / neutral comments**: "I wouldn't want deswir!", "Good fun, thank you!"(see Appendix A.2.1 for Sample comment screen)	
Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between two medicines for each trial. *50 trials for each disease (alternating), to make 100 trials total*. Each disease had an optimal (60%) and suboptimal (40%) option.	n/a	*Byt, Zol / Mox, Nep*
Binary Preference	Posterior Measure: Following trials, participants were asked which of the two machines they preferred.	"Which medicine do you think is better?"	*Byt, Zol / Mox, Nep*
Confidence in Binary Preference	Posterior Measure: Having given their binary preference, participants were asked how confident they were in their preference.	"How confident are you in this preference?"	0-100 slider (default value of 0)
Probability Estimate	Posterior Measure: Participants were asked to give a probability estimate of the distribution of cures between the two medicines. **Note: All posterior measures were completed sequentially on two screens, split by disease.**	"What is the distribution of cures between the two medicines?"	100% Byt / Mox, through 50/50, to 100% *Zol / Nep* (slider, default value 50/50)
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one medicine?"	*Byt, Zol, Mox, Nep, No*

4.2.2 Results

Descriptives and Processing. The *Initial Evidence* (2) x *Counterfactuals* (2) between subject factorial design resulted in 4 groups for analysis. I+C+ (initially supportive evidence and counterfactuals present), I+C- (initially supportive evidence and counterfactuals absent), I-C+ (initially undermining evidence and counterfactuals present), and I-C- (initially undermining evidence and counterfactuals absent). An initial pilot of 20 participants was used to check that people still accepted the manipulation and understood the task. Using an estimated medium effect size based on Staudinger & Büschel (2013), a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 5. Accordingly, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 60 per group.

Following the same procedure as previous experiments using the comment manipulation, the groups sample sizes were increased by 30% to compensate for those failing manipulation checks, calculated from the failure rates of previous experiments, which was in turn conservatively increased a further 10% given the changes in paradigm and possible added complexity of tracking 4 medicines, resulting in a total sample size of 340.

Participants were recruited online using MTurk. Those who had taken part in previous experiments were ruled out from participating. Three participants were removed from analysis, as filter questions indicated they believed the comment section to be rigged. The remaining 337 participants recruited were US based, randomised into either the I+C+ (80), I+C- (79), I-C+ (98) or I-C- (80) conditions. The average age was 36.37 years (*SD* = 12.515) and the sample was 58.2% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment, they were removed from subsequent analysis, leaving 63 in the I+C+ group, 64 in the I+C- group, 60 in the I-C+ group, and 57 in the I-C- group. The decision to remove those who failed to remember a comment was taken following the same protocol and reasoning as in previous Experiments (see 3.2.1), whilst (through the use of a within-subject control) avoiding previous selectivity issues. The following analyses were conducted using the remaining 244 participants, with an average age of 35.42 years (*SD* = 12.087) and 54.9% female. The results below will not be presented exhaustively; for the sake of brevity, only those of central interest to the thesis questions are included.

Correlates. A correlation matrix was run to check for possible relationships between how variables such as Need for Closure, age, gender, and counterbalancing loaded onto the independent variables (i.e. did cells significantly differ in any of these variables). Further, the correlation matrix also checked for possible relationships between the potentially confounding variables above and the dependent variables. In both cases, no significant correlations were found involving potential confounds.

Choice Data. To assess the impact of the *belief*, *initial evidence* and *counterfactual* manipulations on choices, a mixed ANOVA was run using the total number of optimal choices for the belief disease and the total number of optimal choices for the control disease as the two-level within-subjects factor (hereafter termed *belief*). The between-subject factors included in the analysis were *initial evidence* and *counterfactuals*. As can be seen in Figure 4.2.1, there were significant main effects of *belief*, $F(1,240) = 57.201$, $p < .001$, $\eta^2 = .192$, and *initial evidence*, $F(1,240) = 47.116$, $p < .001$, $\eta^2 = .164$, whilst *counterfactuals* showed no main effect.

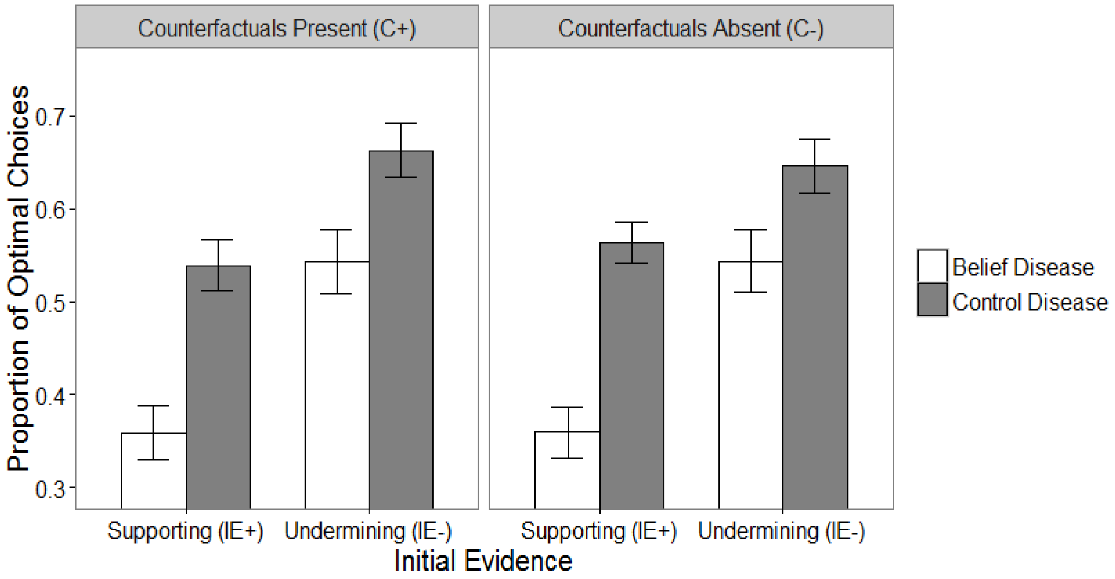


Figure 4.2.1. Experiment 5: Proportion of Optimal Choices. White bars present the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

These main effects showed significantly more choices were made for the sub-optimal medicine when favoured by initial evidence, or having received a belief favouring that option. Furthermore, there was a significant *Belief* x *Initial Evidence* interaction, $F(1,240) = 4.104$, $p = .044$, $\eta^2 = .017$, indicating that the conjunction of both *belief* and *initial evidence* results in significantly greater bias. However, exploring this interaction further by splitting the analysis into supporting and undermining *initial evidence* groups revealed a significant effect of *belief* in *both* supporting, $F(1,125) = 51.776$, $p < .001$, $\eta^2 = .293$, and undermining, $F(1,115) = 13.63$, $p < .001$, $\eta^2 = .106$, groups. Given the significance of *belief* in both of the supporting and undermining *initial evidence* conditions, revisions regarding the power of initial evidence in refutation are elaborated upon in the discussion, although there are large difference in *F* values.

Block Data: Assessing Learning. To assess the impact of the within and between-subject factors on learning, an additional choice data analysis was conducted using 25 trial blocks for each disease. Accordingly, an additional within-subjects factor (hereafter termed *block*) was added to the mixed ANOVA used for total choices, replacing the total choices for each disease with the 1st and 2nd block of the belief disease, and the 1st and 2nd block of the control disease. This allowed for the assessment of *block* and its interaction with the factors of interest – all other effects (those not including *block*) reflect those found in the total choice analysis above.

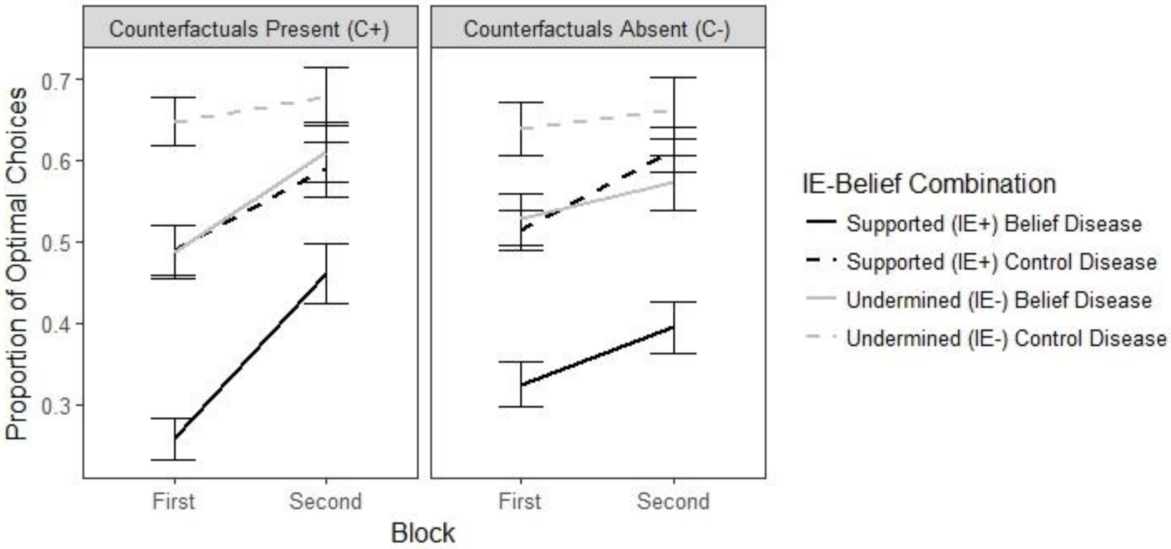


Figure 4.2.2. Experiment 5: Proportion of Optimal Choices, split by 25 trial block. Line colour reflects initial evidence condition (black = *supporting* initial evidence conditions, grey = *undermining* initial evidence conditions), line type reflects within-subject disease (solid = *belief* disease, *dashed* = control disease). Error bars reflect Between-subject Standard Errors.

There was a main effect of *block*, $F(1,240) = 79.807$, $p < .001$, $\eta^2 = .25$, wherein there was an overall learning effect towards the optimal option within a pair. *Block* also interacted with the three main factors; *belief*, $F(1,240) = 7.783$, $p = .006$, $\eta^2 = .031$, as belief diseases (solid lines in Figure 4.2.2 above) generally start from a lower (i.e. sub-optimal) prior starting point; *initial evidence*, $F(1,240) = 9.252$, $p = .003$, $\eta^2 = .037$, as initial evidence dictates different starting points in block 1; and with *counterfactuals*, $F(1,240) = 7.658$, $p = .006$, $\eta^2 = .031$. The latter of these interactions demonstrates the impact counterfactuals have on facilitating learning (i.e. learning is faster when counterfactuals are available – see steeper lines in left-hand facet of Figure 4.2.2). There was no significant four-way interaction between the *Belief* x *Initial Evidence* interaction, and the *Block* x *Counterfactual* (learning) interaction.

Posteriors. Where appropriate, the same format of mixed ANOVA as that conducted in the total choices analysis was used to assess the impact of the independent variables on posterior judgements.

Probability Estimate. This analysis protocol was used to assess the effect of the independent variables on posterior probability estimates in the belief and control diseases (as the within-subject *belief* factor). Although there was once again a main effect of *initial evidence*, $F(1,240) = 25.387$, $p < .001$, $\eta^2 = .096$, the main effect of *belief* did not quite reach significance ($p = .051$).

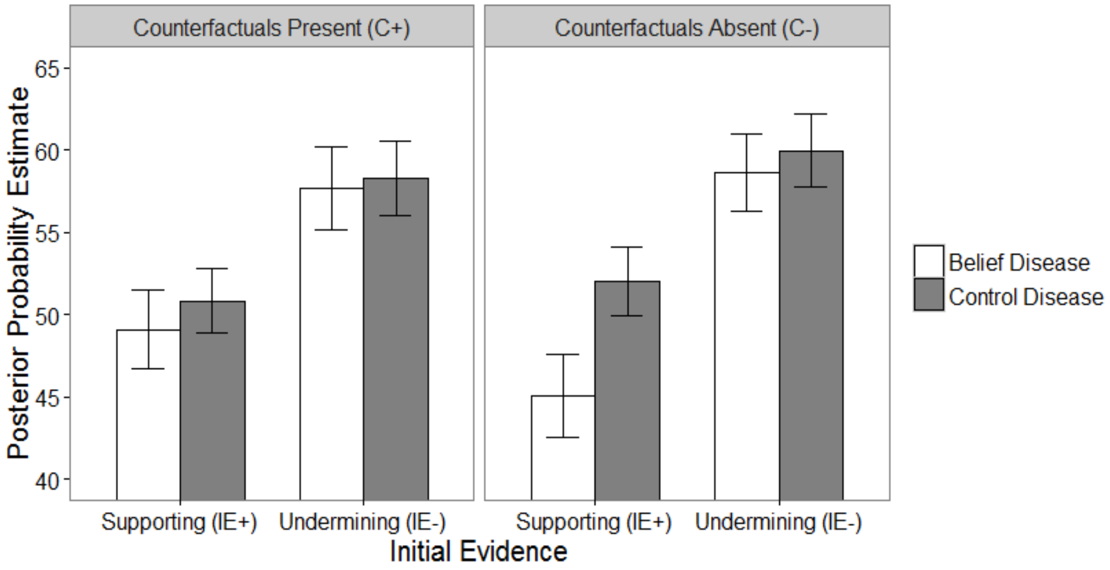


Figure 4.2.3. Experiment 5: Posterior Probability Estimates (estimate of percentage of optimal outcomes in favour of initially dominant medicine), split by group. Error bars reflect Between-subject Standard Errors.

Furthermore, although there was no main effect of *counterfactuals*, as can be seen from the pattern in Figure 4.2.3, when initial evidence is coupled with the absence of counterfactuals, there is a significant effect of *belief*, $F(1,63) = 6.633$, $p = .012$, $\eta^2 = .095$.

Posteriors reflect the impact of counterfactuals in reducing bias effects that were evident from block data, however overall trends suggest a *Belief* x *Initial Evidence* interaction (although this did not reach significance, $p = .21$)independent of *counterfactuals*. Importantly, significant correlations were found between the degree of bias (taken as difference between belief pair and non-belief pair) in choice data and the degree of bias in posteriors, $r = .494$, $N = 244$, $p < .001$.

Binary Preferences and Confidence. Binary preferences required a different approach due to the binary nature of the dependent variable. As such, a mixed-effects logistic regression was run in R to test for each main effect, and the interaction between the between-subject (fixed) and within-subject factors.

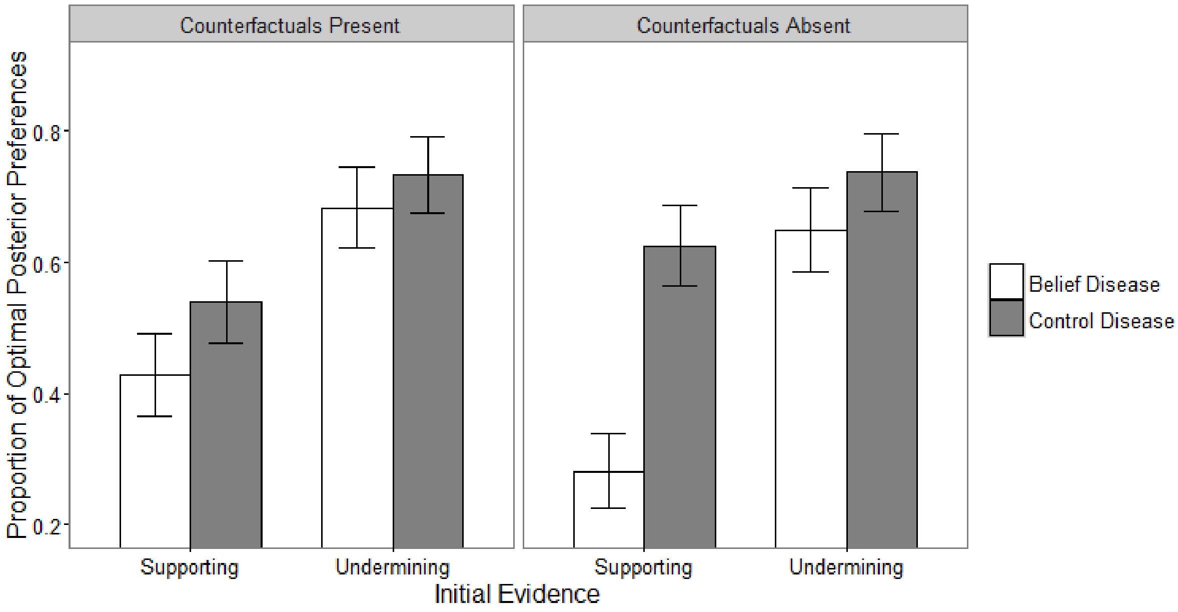


Figure 4.2.4. Experiment 5: Binary Preferences (proportion of participants indicating a preference for the optimal option, split by group). White bars represent the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

Such an analysis compared sequentially more complex models, first finding a significant improvement over the basic model (intercept + subject random-effect) through the inclusion of *belief* as a factor (belief (white) bars lower than control (grey) bars, Figure 4.2.4 above), χ^2 (1) = 12.1322, p < .001. In the same way, a main effect of *initial evidence* was also found (left-hand pairs of bars within each facet are lower than right-hand in Figure 4.2.4), χ^2 (1) = 26.79, p < .001, whilst there was no main effect of *counterfactuals*, χ^2 (1) = 0.3642, p = .5462. Finally, a model containing both significant main effects and the interaction (*Belief* x *Initial Evidence*) term was compared to the null model of solely the main effects. This model comparison revealed no significant improvement due to the inclusion of the interaction term, χ^2 (1) = 2.5535, p = .11.

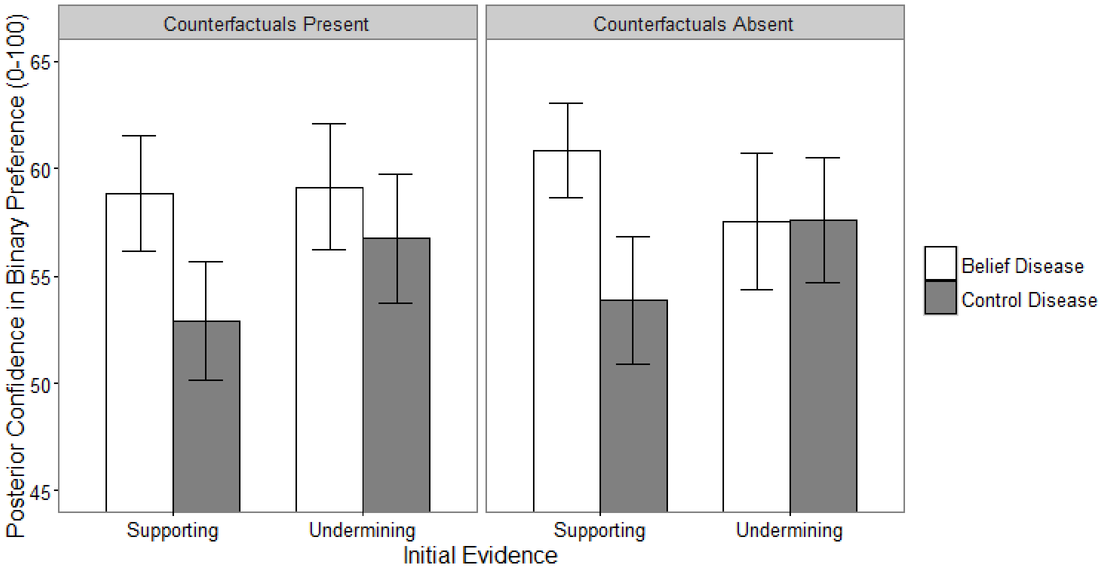


Figure 4.2.5. Experiment 5: Confidence in Binary Preference (0-100%), split by group. Error bars reflect Between-subject Standard Errors.

Regarding confidence in this binary preference (and returning to the mixed ANOVA format), the belief disease posterior confidence and control disease posterior confidence were used as the within-subject (*belief*) 2-level factor. There was significantly greater confidence in belief disease preferences (white bars in Figure 4.2.5 above), F(1,240) = 9.912, p = .002, η^2 = .04. Despite there being no significant main effects of *initial evidence* or *counterfactuals*, there was a significant *Belief* x *Initial Evidence* interaction, $F(1,240) = 4.749$, $p = .03$, $\eta^2 = .019$, wherein the difference in confidence between belief and control disease confidence was exacerbated by initially supporting evidence (left-hand pairs of graphs within each facet of Figure 4.2.5 show larger belief-control differences).

4.2.3 Discussion

The purpose of Experiment 5 was to extend the findings of the previous work (Chapter 3) regarding the role of *belief* and *initial evidence* into a new paradigm, and explore the role counterfactuals play through explicit manipulation of their presence or absence. Results showed that in choice data, both the within-subject *belief* and the between-subject *initial evidence* (the first couple of trials of evidence as either supporting or refuting the sub-optimal – and thus, the belief in the case of the belief disease) factors had strong main effects on choices. However, along with the significant decrease in the overall amount of optimal choices due to either sub-optimally supporting initial evidence or a communicated (misleading) belief, the interaction of the two compounded such biasing effects. This finding supports our previous work indicating the impactful role initial evidence plays in consolidating and exacerbating the biasing effects of communicated beliefs (Chapter 3) and replicates in a more applied setting the work of Staudinger & Büschel (2013). Furthermore, the finding of main effects of *belief* (although weaker) in initially undermining evidence conditions does corroborate the trends found as initial evidence phases were reduced in latter experiments of Chapter 3 (i.e. the subtler the phase of refutation, which in this Experiment is the first 2 trials, the lower the level of refutation is likely to be).

The second principal finding demonstrated by the choice data is the limited role counterfactuals play. Although block data revealed that counterfactuals play a significant role in facilitating learning (i.e. those in conditions without counterfactuals present learned *less* than those with counterfactuals present), this was *independent* of the *Belief* x *Initial Evidence* effect. Accordingly, the presence or absence of counterfactuals does have an impact on the way in which people learn within the task, and further explains the reduced effects found in the posterior measures. In particular, most groups have learnt by the time posteriors are sampled, that the belief is no longer valid. The exception to this is those who both consolidated the belief (i.e. those received initially supporting evidence) *and* were not exposed to counterfactual information (see Figure 4.2.3). Although tentative, this finding is in line with research on selective search strategies in confirmation bias (Jonas et al., 2001), which find stronger confirmation effects of previously held opinions having had the opportunity to select further evidence (in theory to either refute *or* confirm their beliefs, but opting asymmetrically for the latter).

Interestingly, we find a continuation of the posterior confidence effects found in Chapter 3, Experiment 4; participants in groups with the initially supporting evidence (i.e. those most likely to have consolidated the belief) show the greatest disparity in confidence in their posterior binary preference for the belief disease over the control disease (Figure 4.2.4). In other words, those who are in conditions which provoke the strongest degree of bias are also the most confident in the consequent (mistaken) preference.

4.3 Experiment 6: Replication and Intervention

The primary aim of Experiment 6 was to replicate the main findings of Experiment 5; namely the main effects of *belief* and *initial evidence*, as well as their interaction in choice data, the role counterfactuals play in facilitating learning, and exploring these same effects further in posterior judgements. The secondary aim of Experiment 6 was to investigate possible intervention effects on the initial evidence and belief effects founds in Experiment 5. In other words, the aim was to assess whether the impact of initial evidence as consolidating or refuting a communicated belief could be intervened upon explicitly through a directed statement.

In essence, by directing participants towards the potentially misleading or biasing nature of initial evidence, the question of whether such an intervention can lead to a reduction in bias has both interesting theoretical and practical applications. From a theoretical standpoint, such an intervention (i.e. formal instruction / advice) can address the degree to which the evaluation of the communicated belief, or propositional statement, is *explicitly* evaluated based on initial evidence. Methodologically, such a theoretical implication can be broken down further into three possible categories of effect:

Firstly, such an intervention reduces the impact of initial evidence *irrespective* of disease category (i.e. whether there is a communicated belief regarding the options or not), which would suggest the evidence based primacy effect responds to explicit intervention. Previous work on clinical decision making has successfully used interventions based on redirecting attention to the latter evidence in a sequence to reduce primacy effects (Curley et al., 1988), suggesting a top-down impact of interventions through attention mechanisms.

Secondly , the intervention, by calling attention to the potentially misleading nature of early impressions, instead reduces the impact of the communicated belief *irrespective* of *initial evidence* condition. Such an effect would suggest the intervention was interpreted as directly relevant to the participant in assessing the validity of using the communicated belief as a prior.

Finally, which is of most relevance to the effects previously discussed in prior work, the intervention may selectively the consolidation of beliefs, where such beliefs would otherwise have been supported by initially supporting evidence. Although the last of these possibilities bears the most interest for the processes discussed in this work, all three potential effects have practical implications for reducing the impact of erroneous information transfer or evidence sampling (Anderson, 1965; Bar-Hillel & Wagenaar, 1991; Curley et al., 1988; Dennis & Ahn, 2001; Fiedler & Kareev, 2006; Nurek et al., 2014; Peterson & DuCharme, 1967; Staudinger & Büchel, 2013).

4.3.1 Method

Accordingly, Experiment 6 followed the methodology of Experiment 5, with a single exception:

An additional factor, *intervention*, was added. This factor manipulated either the presence of a warning statement regarding primacy (in layman's terms) to the participant, or a control warning statement about avoiding exiting the task early. These statements were pre-tested to ensure participants in either condition read, understood, and could remember which statement they had seen at the end of the experiment. Both intervention statements consisted of two parts, the first was placed between the comment manipulation page and the first trial, which required participants to read a warning statement and then press "Start" to continue on to the trials. The warning statements were as follows:

For the primacy intervention condition:

"You are about to start the task. Don't make up your mind too quickly about which medicine is best. Research has shown that people who take more time to decide perform better."

Whilst those in the control intervention condition saw:

"You are about to start the task. Please be aware that exiting out of the task before you have finished and received the completion code will result in no possible payment."

The second part of the intervention manipulations were a repeat of the condition's warning statement on the feedback screen of the first 6 trials, breaking down into 3 trials for both the belief and control disease. This warning appeared after initial feedback for the given trial had been shown. In this way, the second part of the intervention statements coincided with the duration of the initial evidence manipulation. These secondary intervention statements were as follows:

For the primacy intervention condition:

"Remember not to make up your mind too quickly."

Whilst those in the control intervention condition saw:

"Remember not to exit the task without completing."

As mentioned previously, aside from the addition of an intervention statement manipulation, the remainder of the methodology for Experiment 6 followed that of Experiment 5. This led to the use of the same within-subject factor (the difference between belief disease choices / judgements control disease choices / judgements; *belief*), between subject factors; *initial evidence* (supporting, or undermining), and *counterfactuals* (present, or absent). The intervention statement manipulation added one further between-subject factor: *intervention* (primacy, or control). In summary, this resulted in a *Belief* (2) x *Initial Evidence* (2) x *Counterfactuals* (2) x *Intervention* (2) design.

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over

500 prior HITs. Participants completed the experiment under the assumption that the purpose of the investigation was to improve medical decision making. Participants were English speakers between the ages of 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

4.3.2 Results

Descriptives and Processing. The addition of *intervention* (primacy or control) as a factor resulted in 8 groups for analysis (see Table 4.3.1 below). Two pilots of 20 participants were used to check that people understood the new intervention statements. Using the effect size from the *Belief* x *Initial Evidence* interaction in the choice data of Experiment 5, a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 6. This resulted in an estimate of 50 participants per group, which was conservatively increased to 60 per group to encompass the likely forgetting rate. Given the increase from 4 to 8 groups, the resulting total sample size was 480.

Table 4.3.1: Experiment 6: Participant Breakdown by Group, along with the number of participants passing the belief manipulation check.

Intervention					***N***	**Passed Manipulation Check**
Primacy	**Counterfactuals**	Present	**Initial Evidence**	Supporting	57	51
				Undermining	60	51
		Absent	**Initial Evidence**	Supporting	57	47
				Undermining	58	51
Control	**Counterfactuals**	Present	**Initial Evidence**	Supporting	60	55
				Undermining	54	47
		Absent	**Initial Evidence**	Supporting	64	50
				Undermining	59	49

Participants were recruited online using MTurk. Those who had taken part in previous experiments were ruled out from participating. The participants recruited were US based, randomised into one of the eight possible conditions (see the first column of

Table 4.3.1). The average age was 35.57 years (*SD* = 11.385) and the sample was 55% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comments, they were removed from subsequent analysis, the remaining group numbers are shown in the second column of Table 4.3.1. The decision to remove those who failed to remember the comment manipulations was taken following the same protocol and reasoning as Experiment 5. The following analyses were conducted using the remaining 401 participants, with an average age of 35.51 years (*SD* = 11.135) and 55.1% female. The results below will not be presented exhaustively; for the sake of brevity, only those of central interest to the thesis questions are included.

Correlates. A correlation matrix was run to check for possible relationships between how variables such as Need for Closure, age, gender, and counterbalancing loaded onto the independent variables (i.e. whether cells significantly differed in any of these variables). Further, the correlation matrix also checked for possible relationships between the potentially confounding variables above and the dependent variables. Gender and age did correlate with Need for Closure once again (higher in females; $p < .001$, and higher in older people; $p = .002$).

Choice Data. To assess the impact of the *belief*, *initial evidence*, *counterfactuals* and *intervention* on choices, a mixed ANOVA was run using the total number of optimal choices for the belief disease and the total number of optimal choices for the control disease as the two-level within-subjects factor (*belief*). The between-subject factors included in the analysis were *initial evidence*, *counterfactuals* (presence or absence), and *intervention*. As can be seen in Figure 4.3.1, there were significant main effects of *belief*, $F(1,393) = 49.48$, $p < .001$, $\eta^2 = .112$, and *initial evidence,* $F(1,393) =$

47.745, p < .001, η^2 = .108, whilst *counterfactuals* and *intervention* showed no main effect (p = .108 and p = .102, respectively).

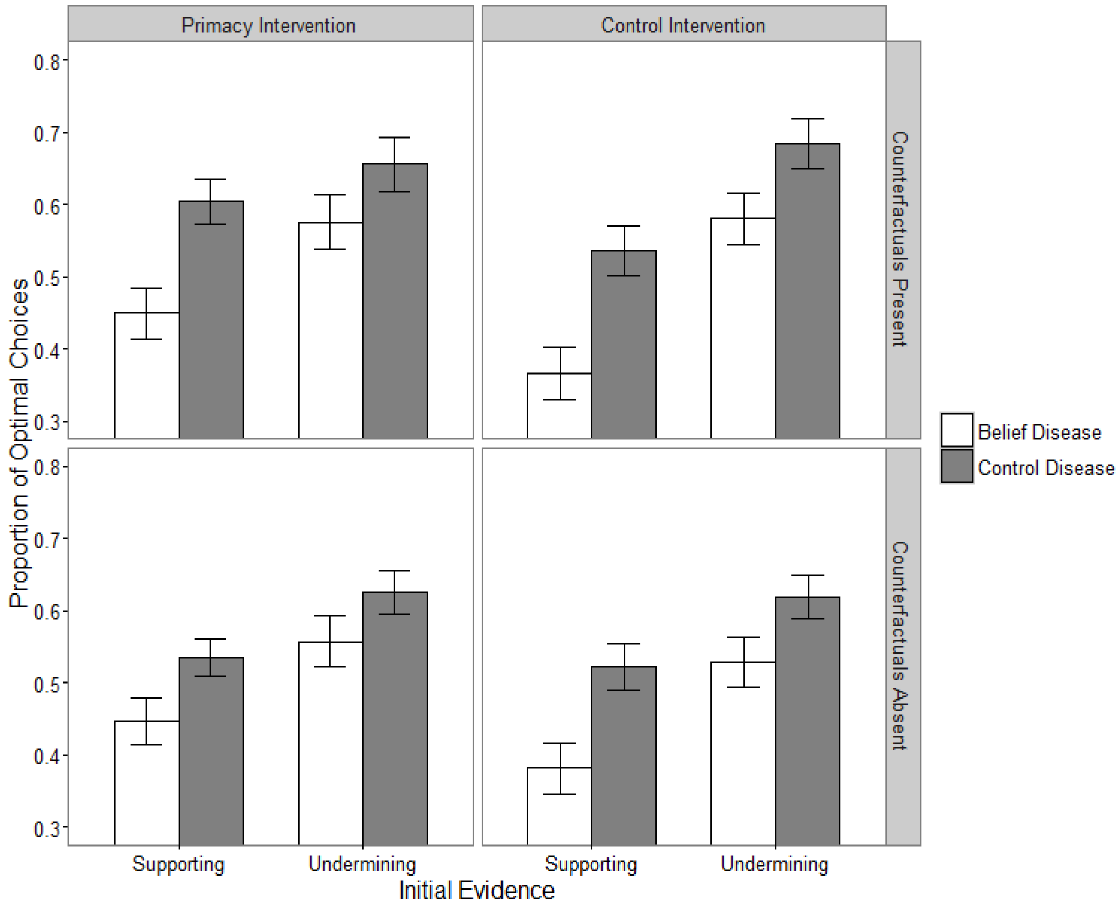


Figure 4.3.1. Experiment 6: Proportion of Optimal Choices. White bars present the disease that received a belief indicating the sub-optimal option. Error bars reflect between-subject standard error.

These main effects showed that significantly more choices were made for the sub-optimal medicine when favoured by initial evidence, or having received a belief favouring that option. There were, however, no significant interactions between factors, although the *Belief* x *Initial Evidence*, and *Initial Evidence* x *Intervention* interactions approached significance (p = .101, and p = .111, respectively). Given the proximity of the latter interaction to significance, and the a priori assertions that the intervention manipulation is *designed* to intervene on the consolidation effect, it is logical to assess the effect of *intervention* on the subgroup that are predicted to have consolidation effects (i.e. the subgroups that have the combination of both a belief and initially supporting evidence for that belief). Thus, a secondary analysis was performed on the subgroup of the initially supporting evidence conditions[23].

Restricting the analysis to those who also remembered the intervention manipulation statement, and focusing on the subgroup that has shown consolidation (supporting initial evidence groups), the total N = 187. A repeated-measures ANOVA was conducted, again using total choices for belief and control diseases as the within-subject factor (*belief*), and intervention as the between-subjects factor. Main effects were found for *belief*, $F(1,185) = 40.192$, $p < .001$, $\eta^2 = .178$, and for *intervention*, $F(1,185) = 6.091$, $p = .014$, $\eta^2 = .032$, although the interaction between the two was not significant. Investigating this further using univariate ANOVA on each of the two DVs independently (belief disease choices and control disease choices), *intervention* had no effect on control choices, $p = .186$, whilst the belief choices showed a significant effect of *intervention*, $F(1,186) = 5.236$, $p = .023$, $\eta^2 = .028$. This suggests the intervention selectively impacts belief disease choices.

Block Data: Assessing Learning. To assess the impact of the within and between-subject factors on learning, an additional choice data analysis was conducted using 25 trial blocks for each disease. Accordingly, an additional within-subjects factor (*block*) was added to the mixed ANOVA used for total choices, replacing the total choices DVs for the 1st and 2nd block of the belief disease, and the 1st and 2nd block of the control disease. Such an analysis allowed for the additional assessment of *block* and the way it interacted with the factors of interest – all other effects (those not including *block*) reflect those found in the total choice analysis above.

[23] When focusing on this secondary analysis, an additional manipulation check was performed to confirm that participants recalled the intervention statement. Those who failed to recall the correct statement at the end of the task were removed from this analysis (a total of 25 removals), leaving 376 participants. No individual group lost more than 10% of their total. Furthermore, all effects found using this restricted version of the secondary targeted analysis hold when such a restriction is not in place.

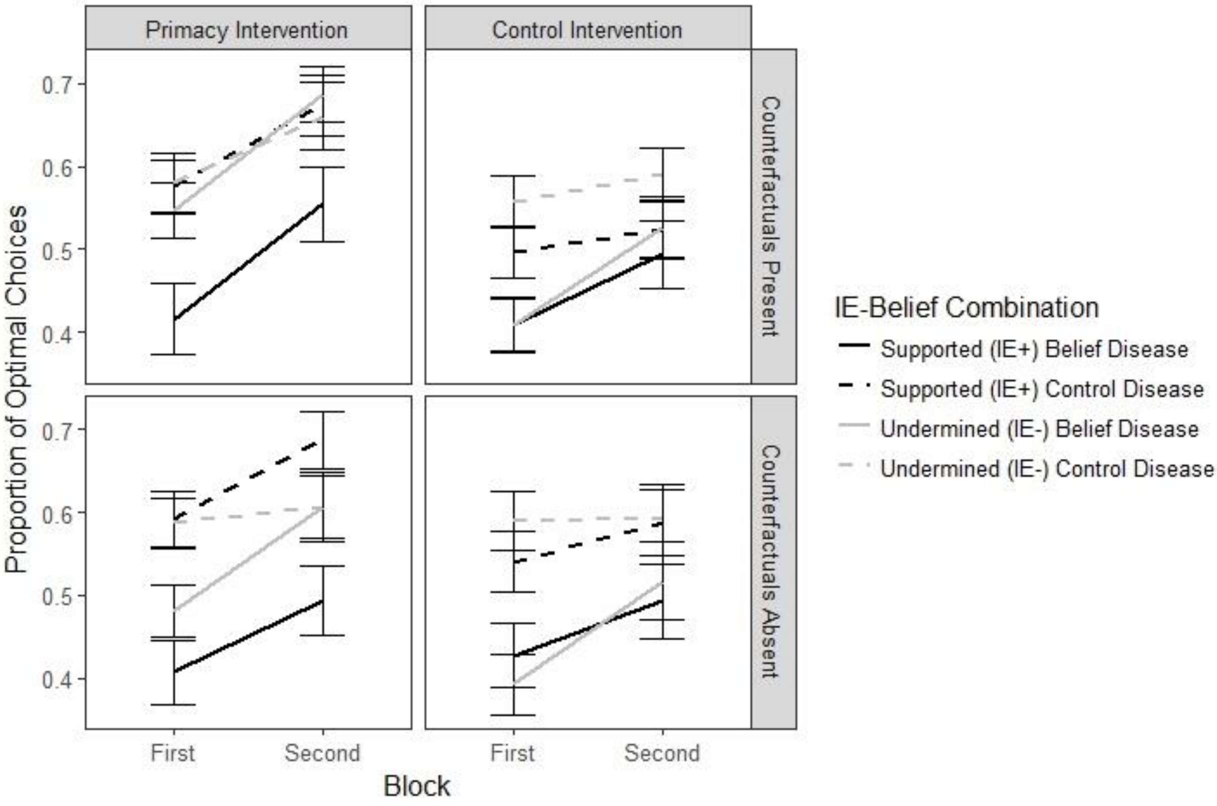


Figure 4.3.2. Experiment 6: Proportion of Optimal Choices, split by 25 trial block. Line colour reflects initial evidence condition (black = *supporting* initial evidence conditions, grey = *undermining* initial evidence conditions), line type reflects within-subject disease (solid = *belief* disease, *dashed* = control disease). Error bars reflect Between-subject Standard Errors.

There was a main effect of *block*, $F(1,393) = 123.589$, $p < .001$, $\eta^2 = .239$, wherein there was an overall learning effect towards the optimal option within a pair. *Block* also interacted with three of the main factors; *belief*, $F(1,393) = 22.288$, $p < .001$, $\eta^2 = .054$, as belief diseases generally start from a lower (i.e. sub-optimal) prior starting point (see solid lines in Figure 4.3.2 above); *initial evidence*, $F(1,393) = 9.124$, $p = .003$, $\eta^2 = .023$, as initial evidence dictates different starting points in block 1; and *counterfactuals*, $F(1,393) = 17.258$, $p < .001$, $\eta^2 = .042$. The latter of these interactions once again demonstrates the impact counterfactuals have on facilitating learning (i.e. learning is faster when counterfactuals are available – see steeper lines in upper versus lower facet rows in Figure 4.3.2). Finally, there was a significant *Belief* x *Block* x

Counterfactuals x *Initial Evidence* x *Intervention* interaction, *F*(1,393) = 4.371, *p* = .037, η^2 = .011. Consequently, and in accordance with the both the choice data secondary analysis, and the a priori assertions regarding the selective impact of interventions on the initially supported (i.e. belief consolidated subgroup), a secondary, sub-group analysis was performed to explore this interaction. A significant *Belief* x *Block* x *Counterfactuals* x *Intervention* interaction was found only in the initially supported evidence groups, *F*(1,183) = 9.428, *p* = .002, η^2 = .049, which fits with the fact that an intervention can only have impact when there is a deviation upon which to intervene.

The significant 4-way interaction showed the primacy intervention facilitated learning (away from the belief-indicated option), in that the primacy intervention reduced the degree of bias (the difference between belief and control diseases) across blocks, but in a different manner depending on the presence or absence of counterfactuals. When counterfactuals were present (top row of Figure 4.3.2), the intervention reduced the impact of the belief-biasing effect gradually across blocks (black solid lines *end* higher due to intervention – left versus right facet of top row, Figure 4.3.2), as evidenced by both the *Belief* x *Block* x *Intervention* interaction, *F*(1,96) = 4.136, *p* = .034, and the further *Belief* x *Block* interaction in the primacy intervention group, *F*(1,46) = 10.277, *p* = .002, a learning effect that did not occur in the control intervention group. Alternatively, when counterfactuals were absent (bottom row of Figure 4.3.2), the bias-reducing impact of the primacy intervention occurred within the first block (black solid lines already are higher in first block in left versus right facet comparison of bottom row in Figure 4.3.2). This is evidenced by the significant *Belief* x *Block* x *Intervention* interaction, *F*(1,96) = 5.292, *p* = .024, breaking down to reveal no significant *Belief* x *Block* interaction in the primacy intervention group, but instead an immediate, significant reduction in the impact of *belief* within the first block, *F*(1,88) =

4.675, p = .033. Inversely, the reduction in the effect of *belief* in the control intervention group occurred gradually across blocks, as evidenced by the significant *Belief* x *Block* interaction, $F(1,43) = 12.423$, $p = .001$. This difference between present and absent counterfactuals can be summarised as a more immediate effect of *intervention* in counterfactual absent conditions, whilst *intervention* has a more gradual effect in counterfactual present conditions.

Finally, the impact of *intervention,* when broken down by disease type (belief or control disease) found the above impact entirely dependent upon changes in the belief disease, whilst the primacy intervention had no impact on the control disease choices.

Posteriors. Where appropriate, the same format of mixed ANOVA as that conducted in the total choices analysis was used to assess the impact of the independent variables on posterior judgements. Given the extended analyses warranted for this experiment, the posteriors have been further split into sections by dependent variable. Importantly, significant correlations were found between the degree of bias (taken as difference between belief pair and non-belief pair) in choice data and the degree of bias in posteriors, $r = .485$, $N = 376$, $p < .001$.

Probability Estimates. This analysis protocol was used to assess the effect of the independent variables on posterior probability estimates in the belief and control diseases (the within-subject, *belief* factor). Although there were significant main effects for both *belief* (belief (white) bars lower than control (grey) in Figure 4.3.3 below), $F(1,393) = 11.582$, $p = .001$, $\eta^2 = .029$, and *initial evidence* (left-hand pairs of bars within each facet lower than right-hand, Figure 4.3.3), $F(1,393) = 15.989$, $p < .001$, $\eta^2 = .039$, there were no main effects for either *counterfactuals* or *intervention*.

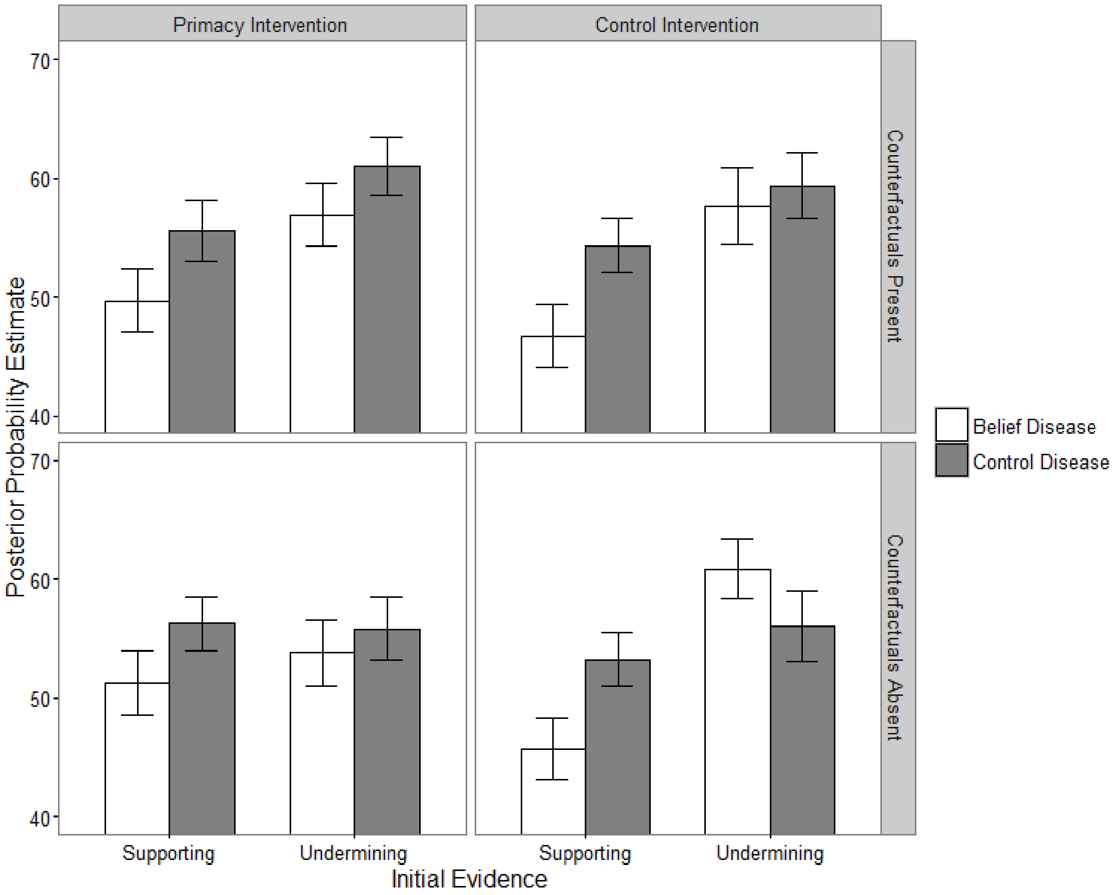


Figure 4.3.3. Experiment 6: Posterior Probability Estimates (estimate of percentage of optimal outcomes in favour of initially dominant medicine), split by group. Error bars reflect Between-subject Standard Errors.

Importantly, the *Belief* x *Initial Evidence* interaction was significant, $F(1,393) = 7.308$, $p = .007$, $\eta^2 = .018$, indicating that those who received a belief that was coupled with initially supportive evidence showed the greatest degree of bias in posterior probability estimates.

Binary Preferences and Confidence. Binary preferences required a different analysis due to the binary nature of the dependent variable. As such, a mixed-effects logistic regression was run in R to test for each main effect, and the interaction between the between-subject (*initial evidence*) and within-subject factors (*belief*).

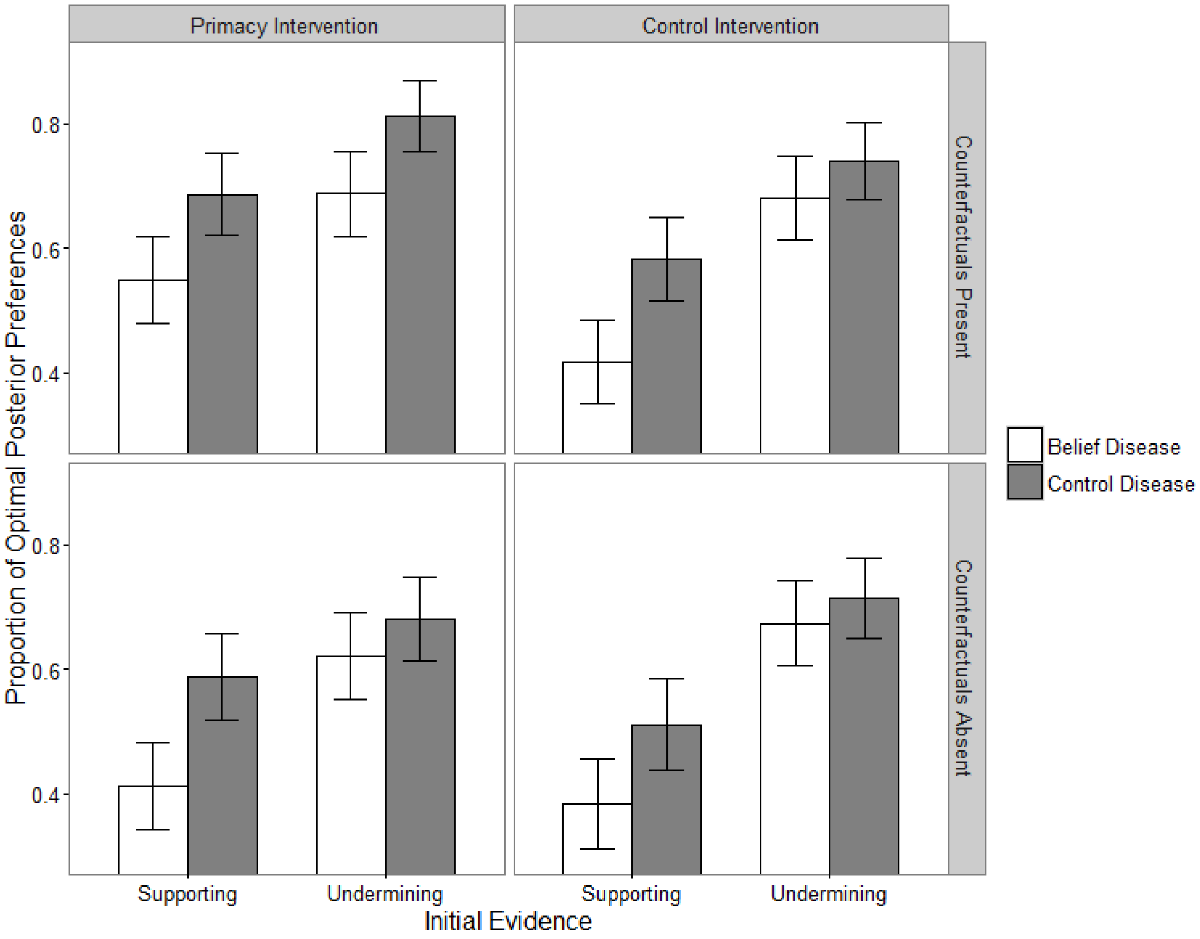


Figure 4.3.4. Experiment 6: Binary Preferences (proportion of participants indicating a preference for the optimal option, split by group). White bars represent the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

This analysis compared sequentially more complex models, first finding a significant improvement over the basic model (intercept + subject random-effect) through the inclusion of *belief* as a factor (belief (white) bars lower than control (grey) bars, Figure 4.3.4 above), χ^2 (1) = 11.42, p < .001. In the same way, a main effect of *initial evidence* was also found (left-hand pairs of bars within each facet are lower than right-hand in Figure 4.3.4), χ^2 (1) = 27.188, p < .001, whilst there was no main effect of *counterfactuals*, χ^2 (1) = 3.4934, p = .062, or *intervention*, χ^2 (1) = 1.3087, p = .253. Finally, a model containing both main effects and the interaction term (*Belief* x *Initial Evidence*) was compared to the null model of solely the main effects. This model comparison revealed no significant improvement due to the inclusion of the interaction term, χ^2 (1) = 0.9261, *p* = .3359.

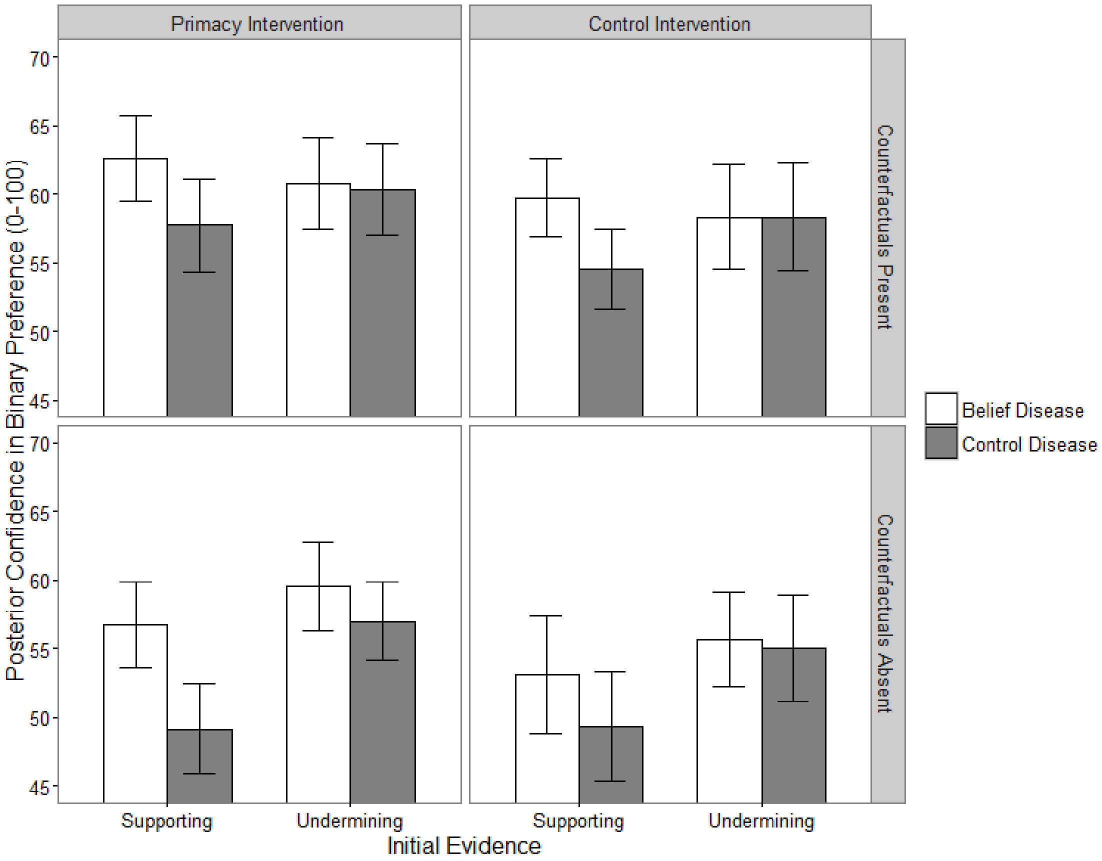


Figure 4.3.5. Experiment 6: Confidence in Binary Preference (0-100%), split by group. Error bars reflect Between-subject Standard Errors.

Regarding confidence in binary preferences (and returning to the mixed analysis of variance format), the belief disease posterior confidence and control disease posterior confidence were used as the within-subject (*belief*) 2-level factor, along with *initial evidence*, *counterfactuals*, and *intervention* as between-subject factors. There was also significantly higher confidence in belief disease preferences (white bars higher than grey in Figure 4.3.5 above), *F*(1,393) = 10.271, *p* = .001, η^2 = .025. Further, although there was no main effect of *initial evidence*, it did interact with *belief*, *F*(1,393) = 5.179, *p* = .023, η^2 = .013, with those who had received both a belief *and* initially supporting evidence showing the highest levels of confidence. Interestingly, although there was no effect of *intervention* on confidence, the presence of counterfactuals did lead to significantly higher confidence, $F(1,393) = 4.28$, $p = .039$, $\eta^2 = .011$.

4.3.3 Discussion

The purpose of Experiment 6 was two-fold; to replicate the general findings of Experiment 5 (control intervention group), and explore the impact of explicit intervention. Briefly, before turning to the impact of *intervention*, it is worth noting that the general findings of Experiment 6 are in line with the effects found in Experiment 5. In particular, these were the main effects of *belief* and *initial evidence* across all measures (both choice data and posterior judgements), and the role of counterfactuals in facilitating learning (although once again independent of any biasing effect).

However, in the choice data, the previously found *Belief* x *Initial Evidence* interaction (although in the correct direction) did not quite reach significance. Breaking choices down by block to assess the role of learning over the course of the task revealed the impact of counterfactuals in facilitating learning, replicating the findings of Experiment 5, along with the independence of this effect from any biasing effect. Within posterior measures, across all three measures (binary preferences, preference confidence, and probability estimates), results again revealed significant main effects of *belief* and *initial evidence*, as well as their interaction in confidence and (unlike Experiment 5) probability estimates.

The analysis of the effect of *intervention* on the choice data focused on the initially supporting evidence conditions (i.e. those who received initial evidence that deviated from the overall probabilities, and thus supported the belief manipulation in the case of the belief disease) as it is these groups who possess a deviation to intervene upon. Choice data revealed an effect of *intervention* in reducing the proportion of sub-optimal choices (i.e. those in supporting initial evidence conditions, who usually perform sub-optimally via primacy and or belief effects, found these effects reduced if provided with an explicit intervention statement warning of the dangers of forming pre-mature impressions).

Furthermore, and of most relevance to the hypothesised impact of such an intervention, this reduction was driven by the belief disease choices. This was further supported by the analysis of learning over time (block data); wherein the intervention was found to significantly impact the effect of *belief* across blocks, dependent on *counterfactuals*. Specifically, when counterfactuals were present, the intervention increased the learning effect (the reduction in the belief-control difference across blocks) evenly across both blocks. Conversely, when counterfactuals were absent, the intervention had a more immediate effect, significantly reducing the impact of *belief* within the first block, such that the belief had been fully abandoned by the second. Such a difference is attributed to exploration requirements that the intervention provokes in counterfactual absent conditions, whilst participants in counterfactual present conditions may consider the intervention whilst continuing to sample one option. In this way, the consequence of the increased exploration in counterfactual absent intervention recipients is a significantly reduced (or removed) consolidation of belief, whereas the counterfactual present alternative incorporates the intervention more gradually (as its impact does not rely on exploration).

This finding supports the proposed impact of the intervention in disrupting the belief -biasing effect, albeit at different moments in the learning process, dependent on counterfactual availability. Although trends indicated a similar impact of the intervention in the posterior measures, it did not reach significance by this stage in the task. Accordingly, although this initial attempt at interrupting the potential negative consequences of communicated beliefs and initial evidence shows promise, further work is required before robust conclusions can be drawn. From a theoretical standpoint, the effects found do support a propositional account of belief adoption and consolidation (De Houwer, 2009; Mitchell et al., 2009; Chapter 3) given the capacity of explicit verbal statements to impact truth value assessments.

4.4 Joint Analysis

Given the methodological similarity between Experiments 5 and 6, and the likely low power issues with the latter, it was decided that – to more robustly assess the validity of the new methodology – both experiments should be combined into a joint Bayesian analysis of the main analyses used in Experiments 5 and 6. Specifically, the effects of interest were the impact (and potential interaction) of *initial evidence* and *belief* on choice data and posterior judgements, and *counterfactuals* in facilitating learning. However, before doing so it was necessary to ratify the assumption of there being no difference between the two experiments.

4.4.1 Bayesian Analysis

To assess the assumption of similarity between Experiments 5 and 6, both sets of data were combined and a new variable was created for the experiment number. Using this factor, a series of Bayesian T-tests were conducted on all dependent variables of interest using the JASP statistical programme (JASP Team, 2016), using the default uniform prior and MCMC methods (as in all subsequent analyses unless specified otherwise). Strong support was found for the null across all dependent variables of interest, in accordance with the <1/3rd. cut off recommendation for Bayes Factors (Dienes, 2014).

Descriptives and Processing. Combining Experiments 5 and 6 resulted in a total sample size of 806. The average age was 35.9 years (*SD* = 11.87) and the sample was 56.3% female. However, following the same protocol as in Experiments 5 and 6, if participants had no recollection of a manipulation comments, they were removed from subsequent analysis. The decision to remove those who failed to remember the manipulation comments was taken following the same protocol and reasoning as Experiment 5[24]. The following analyses were conducted using the remaining 646 participants, with an average age of 35.46 years (*SD* = 11.492) and 55% female.

Correlates. There were no unexpected correlations regarding counterbalancing factors such as disease or medicine outcome assignments. Importantly, significant correlations were found using a Bayesian Pearson Correlation between the choice data and posterior probability estimates in the belief disease, $r = .525$, $N = 646$, BF = 1.27 * 10^{54}, and the choice data and posterior probability estimates in the control disease, $r = .464$, $N = 646$, BF = 1.602 * 10^{38}.

Choice Data. The key question of interest for the overall choice data is to assess the role of *belief* and *initial evidence* factors, and whether they interact. To answer this question, a Bayesian mixed ANOVA was conducted, using the total number of optimal choices in the belief disease and in the control disease as the within-subject *belief* factor. The between-subject factor included in the analysis was *initial evidence*.

[24] The key analyses of section 4.4 were also run with all participants included ($N = 806$). All dependent variables still showed a strong null difference for the effect of Experiment number (ratifying their inclusion). All joint analysis findings were also found when including all participants. Critically, the Bayes Factor for the interaction term (*Belief* x *Initial Evidence*) was significant for both choice data, BF = 26.23, and posterior probability estimates, BF = 8.936. Similarly, the expected strong support for the null of the *Counterfactuals* x *Block* x *Belief* x *Initial Evidence* interaction was found, BF = 7.642 * 10^{-6}.

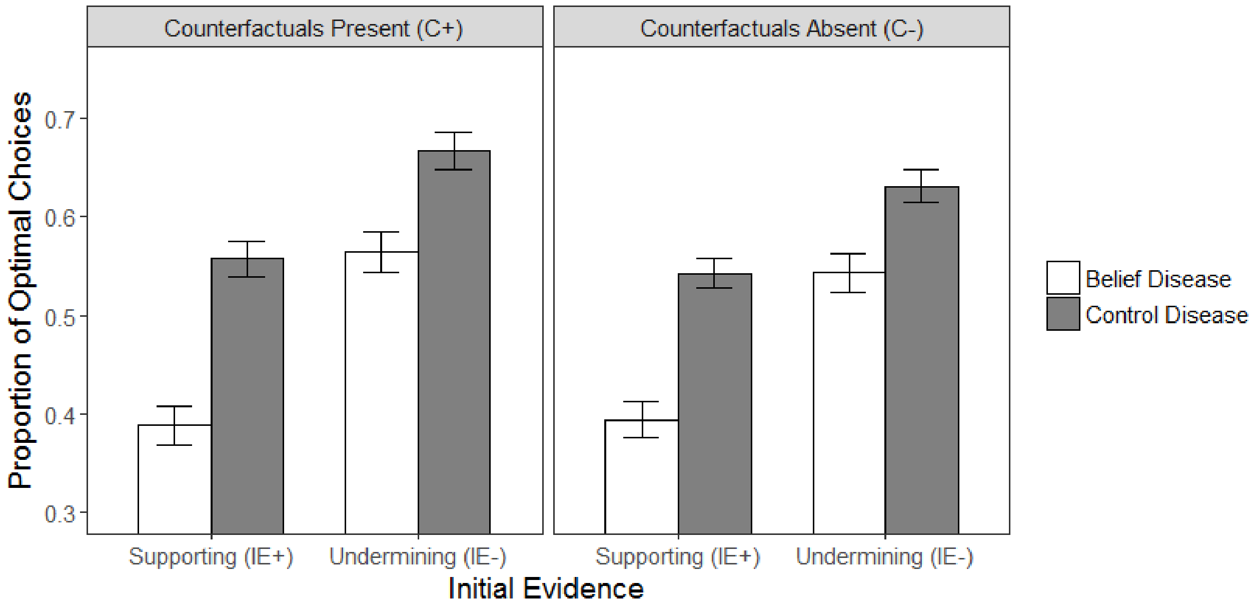


Figure 4.4.1. Joint Analysis: Proportion of Optimal Choices. White bars present the disease that received a belief indicating the sub-optimal option. Error bars reflect between-subject standard error.

This analysis found that a model including *belief*, *initial evidence*, and their interaction term yielded a highly significant Bayes Factor, $BF_{10} = 8.974 * 10^{37}$, a significant improvement upon the model including only main effects of *belief* and *initial evidence*, $BF_M = 10.993$. This was further ratified in an analysis of the effects themselves, with highly significant Bayes Factors for *belief* (belief (white) bars lower than control (grey) bars, Figure 4.4.1 above), $BF = 6.005 * 10^{15}$, *initial evidence* (left-hand pairs of bars within each facet are lower than right-hand in Figure 4.4.1), $BF = 6.005 * 10^{15}$, and the interaction term (difference between belief (white) and control (grey) bars larger in left-hand versus right-hand pairs of bars, within facets, Figure 4.4.1), BF = 10.99.

Block Data: Assessing Learning. For the assessment of learning within the choice data, the main question of interest is whether counterfactuals do improve learning, and whether they impact the degree of bias. To answer this question, a second Bayesian mixed ANOVA was conducted, this time including an additional within-subject 2 level factor (*block*), and an additional between-subject factor (*counterfactuals*). The 1st and 2nd block of the belief disease, and the 1st and 2nd block of the control disease were the resultant dependent variables.

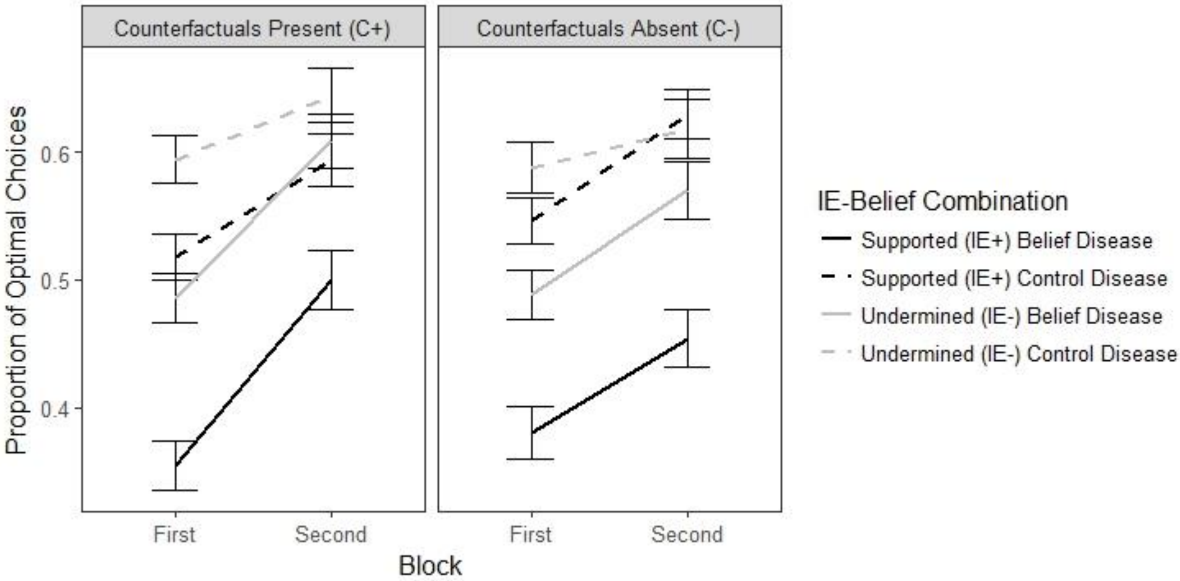


Figure 4.4.2. Joint Analysis: Proportion of Optimal Choices, split by 25 trial block. Line colour reflects initial evidence condition (black = *supporting* initial evidence conditions, grey = *undermining* initial evidence conditions), line type reflects within-subject disease (solid = *belief* disease, *dashed* = control disease). Error bars reflect Between-subject Standard Errors.

By looking at the analysis of effects, it is clear that *counterfactuals* as a main effect does not have any significant impact on choices, BF = 1.046. However, along with a further demonstration of the *Belief* x *Initial Evidence* interaction, BF = 74.911, there was a significant *Block* x *Counterfactuals* interaction (lines are steeper in left-hand (counterfactuals present) than in right-hand facet, Figure 4.4.2), BF = 6.536. This latter interaction indicates that the presence of counterfactuals improves learning. Importantly, the *Belief* x *Initial Evidence* x *Counterfactuals* interaction revealed strong support for the null, BF = .008, which indicates that although counterfactuals *generally* improve learning, their inclusion does not affect the *Belief* x *Initial Evidence* biasing effect. This is further evidenced by the *Belief* x *Block* x *Initial Evidence* x *Counterfactuals* interaction also revealing a strong null effect, BF = 2.275 * 10^{-5}.

Posteriors. Once more for the posterior measures, the key effects of interest are the impact of *belief*, *initial evidence* and their interaction. Given that the block analysis had indicated a lack of interaction between *counterfactuals* and biasing effects, *counterfactuals* were not included in final posterior analyses.

Probability Estimates. The same Bayesian mixed ANOVA was conducted for probability estimates, replacing the within-subject 2-level factor (*belief*) with probability estimates for the belief and control diseases.

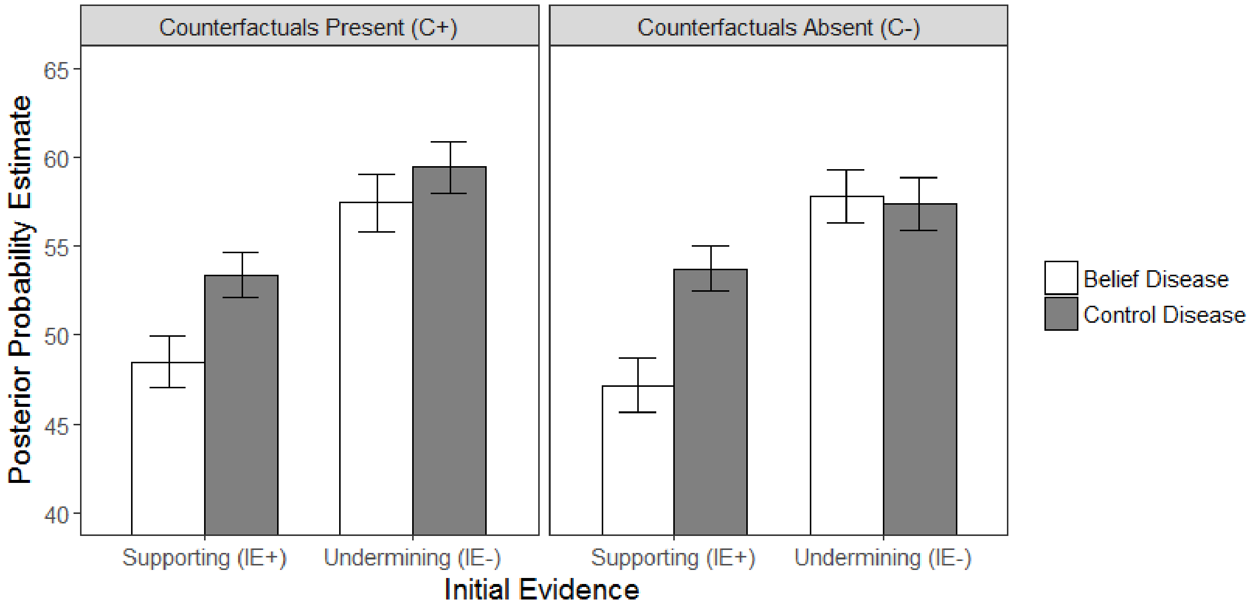


Figure 4.4.3. Joint Analysis: Posterior Probability Estimates (estimate of percentage of optimal outcomes in favour of initially dominant medicine), split by group. Error bars reflect Between-subject Standard Errors.

Replicating the effects found in the choice data, the model including *belief*, *initial evidence*, and their interaction term yielded a highly significant Bayes Factor, BF_{10} = 1.354 * 10^{10}, a significant improvement upon the model only including main effects of *belief* and *initial evidence*, BF_M = 33.393. This was further ratified in an analysis of the effects themselves, with highly significant Bayes Factors for *belief* (belief (white) bars lower than control (grey) bars, Figure 4.4.3 above),, BF = 959.79, *initial evidence* (left-hand pairs of bars within each facet are lower than right-hand in Figure 4.4.3), BF = 7.231 * 10^7, and the interaction term (difference between belief (white) and control (grey) bars larger in left-hand versus right-hand pairs of bars, within facets, Figure 4.4.3), BF = 33.39.

Binary Preferences. To determine whether *belief* and *initial evidence* affect binary preference proportions both independently, and in interaction, a mixed-effects logistic regression was run in R to test for each main effect, and the interaction between the between-subject (*initial evidence*) and within-subject factors (*belief*). Such an analysis is based on the difference in log-likelihoods when comparing sequentially more complex models, and though not a Bayesian analysis in itself, the method approximates a Bayesian method using an informed prior.

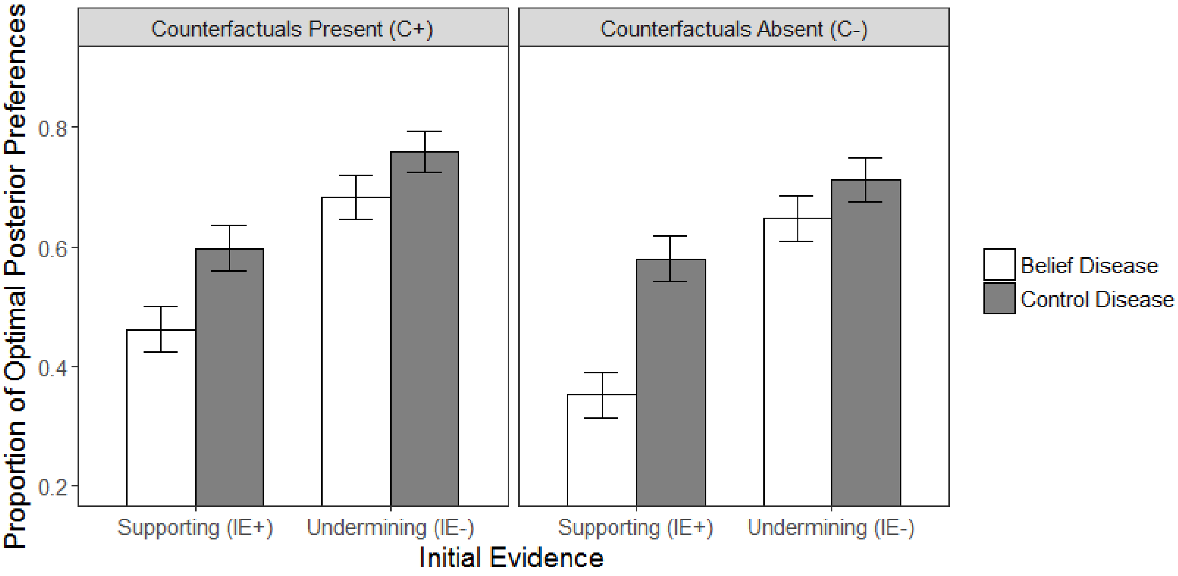


Figure 4.4.4. Joint Analysis: Binary Preferences (proportion of participants indicating a preference for the optimal option, split by group). White bars represent the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

Accordingly, a significant improvement over the basic model (fixed intercept + subject random-effect) was found when including *belief* (belief (white) bars lower than control (grey) bars, Figure 4.4.4 above), χ^2 (1) = 23.155, p < .001. In the same way, a main effect of *initial evidence* was also found (left-hand pairs of bars within each facet are lower than right-hand in Figure 4.4.4), χ^2 (1) = 53.287, *p* < .001. Finally, a model containing both main effects and the interaction term was compared to a null model of solely the main effects. This model comparison revealed that although the inclusion of an interaction term improved the model fit over the null model, it did not quite reach significance, χ^2 (1) = 3.0474, *p* = .081. This is further ratified by the difference in Bayes Factors when running a Bayesian contingency table on the impact of *initial evidence* on control, BF_{10} = 357.9, and belief, BF_{10} = 4.866 * 10^8, preferences. Although both indicate strong primacy effects, the Bayes Factor for belief preferences (tentatively) indicates a substantially larger impact of *initial evidence* than control preferences (the difference between white bars within-facet is generally greater than the difference between grey bars within-facet, Figure 4.4.4).

Confidence. Turning to the posterior measure of confidence, a Bayesian mixed ANOVA was conducted, using the confidence in the binary preference for the belief disease and for the control disease as the within-subject *belief* factor. The between-subject factor was *initial evidence* condition.

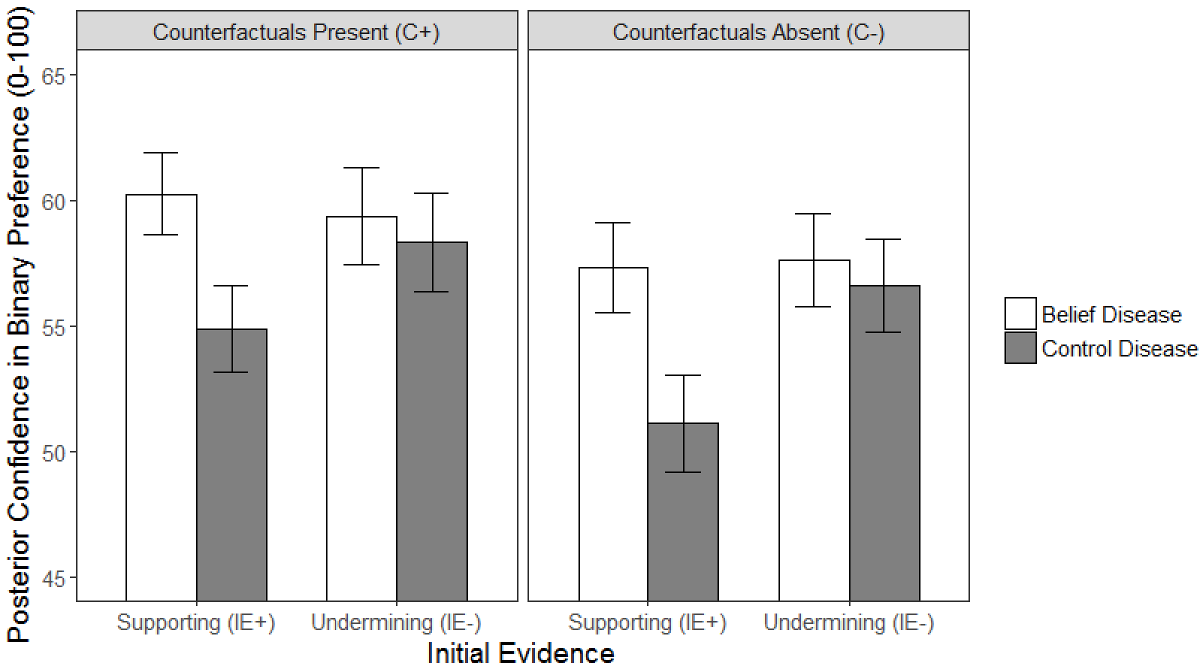


Figure 4.4.5. Experiment 6: Confidence in Binary Preference (0-100%), split by group. Error bars reflect Between-subject Standard Errors.

This analysis found that a model including *belief*, *initial evidence*, and their interaction term yielded a highly significant Bayes Factor, BF_{10} = 4769.923, a significant improvement upon the model including only main effects of *belief* and *initial evidence*, BF_M = 11.075. This was further ratified in an analysis of the effects themselves, with highly significant Bayes Factors for *belief* (belief (white) bars higher than control (grey) bars, Figure 4.4.5 above), BF = 3394.305, and the *Belief* x *Initial Evidence* interaction (difference between belief (white) and control (grey) bars larger in left-hand versus right-hand pairs of bars, within facets, Figure 4.4.5), BF = 11.075, whilst there was no main effect of *initial evidence* on confidence, BF = 2.543.

4.4.2 Discussion

The purpose of the joint analysis above was to use Bayesian statistics to more robustly assess the key findings of the adapted methodology employed in the present chapter. To validate the inclusion of measures in this joint analysis, a series of Bayesian T-tests were conducted on each dependent variable, which all found strong support for the null difference between experiments (Dienes, 2014). Consequently, the principal effect of interest – the *Belief* x *Initial Evidence* interaction – was found in both choice data and posterior judgements including probability estimates and confidence, providing strong evidence for the consolidation / refutation effect.

Furthermore, a Bayesian mixed ANOVA of the block data once again revealed a strong interaction between counterfactuals and blocks, but counterfactual presence or absence did not interact with the *Belief* x *Initial Evidence* biasing effect – in fact showing strong support for the null in both 3 and 4-way (*block* included in the model) interactions. This finding supports the previous assertions of counterfactuals facilitating learning, but that such effects are *independent* of the biasing effects of communicated beliefs. This, together with the evidence for the consolidation effects outlined above, is discussed along with their implications and limitations in the general discussion below.

4.5 General Discussion

The present work looked to extend the findings of Chapter 3 regarding the biasing effects of communicated beliefs as dependent upon initial evidence. By adapting a within-subject belief manipulation, into a probabilistic design used in reinforcement learning (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013), it was possible to assess the impact of immediate, short-term fluctuations in the evidence on belief consolidation and consequent biased choices and judgements. Furthermore, the manipulation of the presence or absence of counterfactual feedback allowed for more direct assessment of integration based bias accounts forwarded in previous work (Chapter 3). Finally, Experiment 6 sought to extend this work through an intervention manipulation aimed at interrupting the previously found consolidation effects. Such an extension allowed for both the testing of theoretical assumptions regarding the propositional nature of communicated beliefs, and future applications aimed at reducing erroneous belief transfer.

The principal finding of the experiments included in the current work, most neatly demonstrated in the joint Bayesian analysis, is the interaction between communicated beliefs and initial evidence in both choice and posterior judgement data. The significantly higher number of suboptimal choices and judgements as a consequence of a communicated belief supported by the first few trials of experienced evidence not only supports previous work on the interaction between these two elements (Chapter 3), but further extends the potency of such effects to immediate, short-term fluctuations of evidence. It should be noted that large main effects were found for both the *belief* and *initial evidence* manipulations, irrespective of an interaction. This indicates that both these elements, whether the first couple of trials of evidence, or a (fallacious) communicated belief indicating a superior option, have powerful effects on the overall interpretation of evidence.

However, the finding of a significant interaction, both in choice and posterior judgement data, does lend support to the forwarded mechanism of initial evidence acting as a “gatekeeper” to confirmation bias effects of communicated beliefs. Specifically, when presented with a belief from a source that is stripped of cues that impact belief adherence, such as credibility (Bovens & Hartmann, 2003; Briñol & Petty, 2009; Hahn et al., 2009), attractiveness (Kelman, 1958), or similarity to recipient (Petty & Cacioppo, 1979), initial evidence is required to validate both the belief and (by proxy) the source, increasing the confidence in the two elements being true (i.e. the belief is valid, and therefore the source is reliable). When such cues exist, such as experimenter authority / expertise cues (Goodwin, 2011; Sniezek & Van Swol, 2001), or affiliation with the source (Frost et al., 2015), alternative cognitive and motivational explanations are added, such as increased confidence (or scepticism) provoked by the former (Hahn et al., 2009; Harris et al., 2015; Schul & Peri, 2015; Sniezek & Van Swol, 2001), and directional motivations added by the latter (Klein & Kunda, 1989; Kunda, 1990). Thus, when such cues are stripped away, we are left with an account of belief biasing effects that relies on evidence-based consolidation (and thus depend on evidence order effects), followed by integration. Such an account fits with the notion of a communicated belief acting as a proposition possessing a truth value that is evaluated by the recipient (De Houwer, 2009; Mitchell et al., 2009). Although this approach allows for a singular mechanistic explanation, and has ecological validity to the on-line context of belief transmission (i.e. highly variable or non-existent source cues), further work could benefit from integrating source elements into the existing, initial evidence-based framework.

The second finding, most ably demonstrated in the joint Bayesian analysis, of counterfactuals impacting the degree of general learning across trials (wherein the presence of counterfactual feedback facilitates learning of optimal choices), demonstrates that counterfactual information is attended to, when available, within the task. However, the strong support for the *null* interaction of counterfactuals with the aforementioned biasing effect (i.e. the interaction between communicated beliefs and initial evidence) suggests that such biasing effects occur independently of counterfactuals. Such a finding may be attributed to the removal of directional motivations for confirmation, something that is present in selective search studies that find strong biasing impacts of the absence of counterfactual information (Jonas et al., 2001; Lord et al., 1979). In other words, the *belief* effect does not work through *directed* search of information. Consequently, further support is found for the forwarded integrative or second-order confirmation bias mechanism forwarded in Chapter 3, whereby the consolidated communicated belief acts as a pattern against which subsequent evidence is interpreted as either confirmatory or *not*. Such an interpretation results in an asymmetry in updating as pattern-evidence matches receive stronger updating signals than evidence that does not fit a pattern (Whitman et al., 2015), as there is no equivalent *null* pattern to confirm.

The final, albeit more tentative finding from Experiment 6 further speaks to the propositional account of communicated beliefs. Specifically, the selective impact of a verbal intervention statement (designed to warn participants against forming pre-mature impressions) on the proportion of sub-optimal choices that followed the conjunction of beliefs and supporting initial evidence, indicates an interruption of consolidation via prolonged deliberation. These results require further replication, most notably considering their restricted impact on choice data alone. The findings do, however, suggest that along with interesting theoretical implications, there are fruitful avenues of further research for investigating such interruption effects of interventions on sub-optimal reasoning outcomes in other domains, including consumer research (Ha & Hoch, 1989; Hoch & Ha, 1986), impression formation (Anderson, 1965; Mann & Ferguson, 2015; Smith & Collins, 2009), advice taking (Biele, Rieskamp, & Gonzalez, 2009; Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000), gambling research (Ayton & Fischer, 2004; Balodis, MacDonald, & Olmstead, 2006; Gilovich, 1983; Joukhador, Blaszczynski, & Maccallum, 2004), illusion of control (Langer, 1975; Yarritu et al., 2013) and superstition / pseudoscience research (Matute et al., 2011; Ono, 1987; Rudski et al., 1999).

The results discussed above, as has been mentioned in previous discussions, are occasionally limited in parts (for example the interaction of *belief* and *initial evidence* is found in choice data, but does not reach significance in posterior probability estimates in Experiment 5, whilst the reverse is found in the targeted replication analysis of Experiment 6). Such inconsistencies have been attributed to lower than anticipated power given the subtlety of the manipulations in the newly developed methods. To resolve such a limitation, we feel the joint Bayesian analyses of the two experiments to assess principal effects helps (given its valid application) inform firmer overall conclusions of the effects at hand. Further, it is anticipated that although the size of the *Belief* x *Initial Evidence* interactions relative to the main effects of belief and initial evidence, are modest, we feel such interactions are likely to be exacerbated in less artificial environments in which directional motivations and selective search factors are present.

The second limitation pertains to the use of MTurk within this work. As has been mentioned in previous work (Chapter 3), the lack of experimental control associated with online studies (Goodman, Cryder, & Cheema, 2013) is in part compensated for by larger and more representative sampling (Henrich, Heine, & Norenzayan, 2010). Further, the online context for the study allows for the excision of demand characteristics as unwanted motivational explanations for bias (i.e. a participant is more motivated to adhere to a belief if it comes from an experimenter, rather than an anonymous source), separating this work from more artificial, lab based studies (Staudinger & Büchel, 2013). Lastly, the use of an online context (and therein, an online communicated belief manipulation), has a growing real-world validity in light of the increasingly dominant role of the internet in communication networks in the modern world.

Such real-world applicability of the methods used in the current work touches on a potent implication of this work in the susceptibility of internet users (and humans as reasoners in general) to the combination of communicated beliefs and immediate, short-term fluctuations in evidence. The coupling of a communicated directional belief (i.e. hypothesis) and the sensitivity to (or over-reliance on) initial evidence demonstrated in this work, *irrespective* of counterfactual evidence exposure and alternative motivations for biased reasoning (such as demand characteristics and self-concept preservation) point towards an integrative account of confirmation bias (Klayman, 1995; MacDougall, 1906; Nickerson, 1998). Furthermore, the demonstration of the critical impact of initial evidence in conjunction with a communicated belief (or hypothesis) has ramifications for areas of research that involve these two elements, including impression formation (Anderson, 1965), causal learning (Yarritu, Matute, & Luque, 2015; Yu & Lagnado, 2012), consumer research (Ha & Hoch, 1989; Hoch & Ha, 1986), and decision making (Nurek et al., 2014), the latter of which has started to explore the interplay between mismatched communicated and experienced evidence (see Weiss-Cohen et al., 2016).

From an applied perspective, the impact of interventions shown in Experiment 6 follows previous suggestions (Chapter 3) and demonstrates not only the notion of critical periods of evidence evaluation, but a potential avenue for reducing erroneous belief acquisition. Although interventions aimed at reducing erroneous belief formation have found success with re-categorisation and positive exposure strategies (Gaertner & Dovidio, 2014), as well as exposure to and directed attention towards null evidence (Blanco, Barberia, & Matute, 2014), the current research suggests evidence order as a fruitful addition.

Finally, the effects discussed in the current work bear particular relevance to other areas of psychological research involving misinterpretations of first-hand evidence, including pseudo-contingencies (Fiedler & Freytag, 2004; Yarritu et al., 2013; Yarritu, Matute, & Vadillo, 2014), superstitious responding (Ono, 1987; Rudski, Lischner, & Albert, 2012), and placebo effects (Wickramasekera, 1980; Wager & Atlas, 2015). In particular, the relationship between reliance on prior information and the degree of ambiguity in experienced evidence (Abbott & Sherratt, 2011) involved in these literatures in various forms is a likely entry point for the application of false-belief uptake effects. For example, to increase the uptake of placebo responses (a placebo effect being the consequence of the false belief that one has received the medically potent treatment), investigating possible interactions between what treatment beliefs are communicated to patients, and initially supportive evidence for such beliefs, may yield advantageous uptake and adherence rates. As such, the current work seeks to provide further insight into the way in which unsupported beliefs may be taken up and adhered to, laying the groundwork for an ecological model of the core mechanisms involved.

CHAPTER 5: FALSE PROPHETS AND CASSANDRA'S CURSE: A LITTLE TRUST GOES A LONG WAY

Along with the human capacity to communicate information to others, the capacity to convey misinformation has expanded along with it. Whether such misinformation is the result of ill-intent, or first-hand misperceptions and misinterpretations, the consequences can be severe. Whether it is the mistaken belief that vaccines cause autism, or that the more diluted a substance is, the more potent it becomes (the principle behind homeopathy), the fact such beliefs are capable of being *communicated to others* – the reach of which has exponentially increased with the advent of mass-communication and the internet – means that instead of such erroneous beliefs dying with the originator, they spread to new hosts, carrying the costs with them. Of particular relevance when investigating communicated beliefs, is from whom the belief originates. The source of a belief undeniably matters greatly; people are inclined to dismiss a statement from a drunkard on the street, such as global warming being a hoax created by the Chinese, but when the same statement is made by the Republican nominee for the US Presidency, the results are alarmingly different[25]. How the perceived credibility of a source influences not only the uptake of a belief, but how recipients then interpret subsequent evidence in light of this belief, is the central question of this chapter. In answering it, the predictions of the present work are informed by current models of source credibility, most notably the Bayesian model of source credibility (Bovens & Hartmann, 2003; Hahn et al., 2009; Hahn, Oaksford, & Harris, 2012).

Although the general question of how people update their beliefs given expected testimony (or evidence) is of interest, the present work focuses on two situations:

[25] Donald Trump went on to win the 2016 United States Presidential election, despite such pronouncements.

Firstly, what happens when a credible source tells a lie, and secondly, what happens when a dis-credited source tells the truth? In both instances, it is of interest how these factors (the belief and its source) influence the subsequent interpretation of evidence – either to accept and adhere to the belief, or refute it. Previous work has shown that anonymously sourced, erroneous communicated beliefs can lead to confirmation bias effects in the integration of subsequent first-hand, probabilistic evidence (Chapters 3 & 4), which results in adherence to sub-optimal action-outcome hypotheses (or beliefs). Importantly, such biasing effects are influenced by initial evidence exposure, wherein beliefs that are undermined by initial evidence have significantly less impact on choices and judgements than beliefs which are initially supported. In this way, initial evidence plays a "gatekeeping" role in the validation of the belief (and by proxy, the unknown source of said belief). Before discussing the incorporation of source credibility elements into this work, it is first worth fleshing out the proposed mechanism underpinning the misinterpretation of first-hand evidence that follows consolidation: confirmation bias.

5.1.1 Confirmation Bias

Confirmation bias is generally defined as the overweighting of confirmatory instances, or underweighting of contradictory instances (Klayman, 1995; Nickerson, 1998). Such a definition covers a plethora of associated mechanisms and effects (Hahn & Harris, 2014), linked by the common consequence: the belief or hypothesis is (erroneously) preserved despite evidence indicating the contrary. Research investigating these effects has typically focused on the role of cognitive mechanisms (Dave & Wolfe, 2003; Decker et al., 2015; Doll et al., 2011; Klayman, 1995; Staudinger & Büchel, 2013), which can broadly be placed into two categories (MacDougal, 1906): First-order, or *input* based biases, involve selective exposure to evidence that favours the current hypothesis, such as positive test strategies (Hendrickson et al., 2014; Klayman & Ha, 1987; Navarro & Perfors, 2011) and selective search (Jonas et al., 2001; Lord et al., 1979). Second-order, or *integration* based biases, involve the over or underweighting of evidence as it is integrated to update the currently held belief (Hahn & Harris, 2014; Klayman, 1995; MacDougall, 1906). Support for an integrative account comes from the finding that asymmetrically stronger updating signals occur when evidence is pattern-matching (confirmatory), than when evidence is pattern-mismatching (null, or contradictory; Whitman et al, 2015). In conclusion, there is no singular mechanism ascribed to cognitive explanations behind confirmation bias effects, but instead a collection of different effects and strategies grouped under this same outcome.

Further, beyond purely cognitive mechanisms behind confirmation bias, research in the field of social cognition has provided motivated reasoning explanations (Kunda, 1990). In a similar manner to the cognitive mechanisms described above, motivational explanations of bias can be broadly categorized into two, interrelated camps. The first of these can be summarised under the term the motivation for *accuracy*, or the desire to find the *correct* answer (Chen, Shechter, & Chaiken, 1996; Hahn & Harris, 2014; Kunda, 1990; Tetlock, 1983b). Typically manipulated through accuracy incentives (e.g., judgements being made public and open to scrutiny), associated effects include higher quality inferences (Kunda, 1990) and deeper processing (Tetlock, 1983a, 1985a). However, effective reductions in deviation are argued to be limited to biases that follow from *hasty* reasoning (Kruglanski & Fruend, 1983), and as such, do not prevent the use of other, faulty reasoning strategies, including the availability heuristic (Tversky & Kahneman, 1973) and hindsight bias (Fischhoff, 1977).

Conversely, the second category of motivations, known as *directional*, can be considered a motivation to come to the *right* (i.e. of most benefit to the individual) answer (Hahn & Harris, 2014; Kunda, 1990), rather than the *correct* answer. This has been proposed to result in a consequent increase in the use of sub-optimal cognitive strategies, such as the use of selective search. In doing so, this facilitates the desired outcome, such as preserving a coherent or positive self-concept (Cialdini & Goldstein, 2004; Festinger, 1962; Kusec et al., 2016), whilst the individual remains confident of their objectivity (Kunda, 1990). One pertinent example of the potential influence of directional motivations comes from the notion of social coherence and proofing (Cialdini & Goldstein, 2004). Whether it is the previous (public) subscription to a given belief, or the perceived social acceptability of holding an alternative belief, both, via internal coherence in the former, and perceived social costs in (both the former and) the latter, are likely to motivate reasoners in the direction of a particular (acceptable) outcome (Heine et al., 2006; Proulx & Inzlicht, 2012). For example, if the perceived consensus on capital punishment is against it, then an individual (especially someone who has already declared they are also against it) is motivated to interpret ambiguous evidence in a manner that results in the same (anti-capital punishment) conclusion. Concluding the opposite carries perceived costs, including social exclusion, and the (self-)perception of inconsistency.

In previous work (Chapters 3 & 4) we have sought to disentangle these mechanisms (both cognitive and motivational) behind the confirmation bias effects in evidence integration that follow from communicated beliefs. Through the use of counterfactual feedback (i.e. avoiding selective exposure explanations) and accuracy rather than belief adherence incentives (i.e. ruling out directional motivation explanations), we were able to demonstrate an integrative, a-motivational explanation for the biasing effects found. Furthermore, the inclusion of an accuracy incentive indicated the biasing effect was not due to a hasty reasoning process, as the presence of accuracy motivations did not remove the bias. The paradigm, in which a communicated belief is updated in light of sequentially presented probabilistic evidence, as used in Chapter 4, consequently forms the methodological backbone to the present work and onto which is grafted manipulations of source credibility. This paradigm defined beliefs as generic (or “unquantified”) statements (Cimpian et al., 2010; Leslie, 2008), in line with an account of beliefs as propositional statements about actions and outcomes (Mitchell et al., 2009), wherein the “truth” of a statement may be evaluated as an (in)accurate representation of the world. Further, this characterisation of belief fits with the naturalistic communication of real world (erroneous) beliefs (Gilovich, 1993).

Importantly, to focus solely on the interplay between evidence factors (such as ambiguity, quantity and order) and communicated beliefs, cues regarding the source of the belief were purposefully removed. Factors including the perceived expertise (Goodwin, 2011; Sniezek & Van Swol, 2001), trustworthiness (Siegrist et al., 2005; Twyman et al., 2008), attractiveness (Kelman, 1958) and similarity to recipient (Petty & Cacioppo, 1979) have been shown to impact argument strength (Briñol & Petty, 2009; Hahn et al., 2009) in persuasion, advice taking, and argumentation literature. Correspondingly, work in social psychology has explored the role of warmth and competence in sources as traits (Cuddy, Glick, & Beninger, 2011; Kenworthy & Tausch, 2008), their impact on positive perceptions and behaviour (Fiske, Cuddy, & Glick, 2007), and their relationship to the motivational goals of the recipient (Tausch, Kenworthy, & Hewstone, 2007). Similarly, the literature on management has demonstrated the impact of trust between employees and management (Mayer, Davis, & David, 1995) on various outcomes, including risk taking (Colquitt, Scott, & LePine, 2007) and job performance (Mayer & Davis, 1999; Mayer & Gavin, 2005). These source factors, given such an influence, add alternative motivational and cognitive explanations for belief adherence, and thus obfuscate more universal evidence factors. Models such as the Elaboration-Likelihood Model (Petty & Cacioppo, 1984) and the Heuristic-Systematic Model (Chaiken & Maheswaran, 1994) have been developed in response to the question of how such source cues are used. These models predict that a belief recipient will use source information as cues or shallow heuristics when evaluating the likely truth of an argument (although such models indicate this information should be overruled under conditions of effortful engagement). Pertinently, recent evidence has shown source information can directly impact evidence integration. Source factors such as affiliation to recipient (i.e. friend vs stranger) have contributed to confirmation bias effects in the perception and recollection of confirmatory versus contradictory declarations (Frost et al., 2015). This leads to the necessary discussion of literature that has involved the *evaluation* and use of sources in light of more objective evidence.

Previous work in the decision making literature has investigated the roles trust and expertise play in the uptake of advice when making decisions or judgements. Typically, this research has focused on the adjustments an individual (or judge) will make to a judgement in light of supplemental information from an advisor (Harvey & Fischer, 1997; Bonaccio & Dalal, 2006). This has explored the interaction between the difference in judgements between the advisor and the judge (how different is the advice from my own assessment), and previous success of the advisor in question (Schöbel et al., 2016; Twyman et al., 2008). Research has not only shown that individuals pay greater attention to advisors that have demonstrated expertise and trustworthiness (Twyman et al., 2008), ignoring less reliable advisors, but that individuals are also willing to adjust their own judgements further as a consequence (Sniezek & Van Swol, 2001). This research has typically used quantified, statistical "advice", that typically pertains to a one-off judgement or decision, often focused on "risky" versus "safe" decisions (Bonaccio & Dalal, 2006) or forecasting (Harvey & Fischer, 1997), which separates it from the context of the present work via the focus on generic beliefs (Cimpian et al., 2010) rather than quantified or probabilistic statements.

In summary, there are a wide range of influencing factors that may contribute to belief-maintenance biases. A large body of work on confirmation biases has investigated various mechanisms behind evidence-based belief updating, typically through cognitive mechanisms such as selective exposure or search (Doherty et al., 1979; Jonas et al., 2001; Lord et al., 1979; Wason, 1960), and biased *integration* (Gilovich, 1983; Klayman, 1995; MacDougall, 1906; Nickerson, 1998; Whitman et al., 2015). Further, directional motivation accounts (Cialdini & Goldstein, 2004; Klein & Kunda, 1989; Kunda, 1987) have been reconciled with these cognitive processes as increasing the likelihood of sub-optimal strategy deployment (Hart et al., 2009; Kunda, 1990). Pertinently, work on second-hand influences on belief-updating, including argumentation (Goodwin, 2011; Hahn et al., 2009; Harris et al., 2015; Madsen, 2016), risk communication (Ludvig & Spetch, 2011; Siegrist et al., 2000; Siegrist et al., 2005), marketing (Ha & Hoch, 1989; Hoch & Ha, 1986; Zhou & Guo, 2016) and advice taking (Harvey & Fischer, 1997; Twyman et al., 2008; Weiss-Cohen et al., 2016; Yaniv & Kleinberger, 2000; Yaniv, 2004) has explored a similar semantic concept to the present work regarding the influence of information sources. These source factors have been proposed to influence recipients through both directional motivation (Cialdini, 2003; Cialdini & Goldstein, 2004) and cognitive processes (Chaiken & Maheswaran, 1994).

The present work, however, seeks to shed new light on communicated belief uptake and maintenance processes. By building off previous work (Chapters 3 & 4), we seek to place such effects within a motivational (and social) context. Further, using generic belief statements (Cimpian et al., 2010; Leslie, 2008) and probabilistic evidence-exposure, this work seeks to explore the interplay between source credibility elements, and the previously found order effects in determining subsequent belief uptake and maintenance. However, before synthesising these elements into the current research paradigm, it is worth providing a formalisation of the relationship between source credibility, (generic) belief content, and evidence, as it pertains to the maintenance of a second-hand belief in light of first-hand evidence (rather than a first-hand belief updated by second-hand information).

5.1.2 A Bayesian Model of Source Credibility

How a human reasoner deals with new information from a given source is an intriguing and important question. Indeed, given that a large portion of the information we receive and use on a daily basis is not generated from first-hand experience, but originates instead from various sources, such as the newsreader or meteorologist on TV, colleagues, friends and family, or complete strangers, among many others. Although the evidence may vary in assessing a belief, the source itself carries cues that may influence such assessment. For example, although both a drunkard and a nurse may provide the same information regarding getting a mole on one’s arm checked, the persuasiveness of the claim differs based on the source.

However, as the truth value of a proposition has been argued to be independent of the source, appeals to authority have traditionally been considered an argument fallacy (*ad verecundiam*). Perhaps given this reasoning, dual-process models such as the Elaboration Likelihood Model and the Heuristic-Systematic Model (Briñol & Petty, 2009; Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984) have considered attention to source cues as a shallow heuristic. To explain further, such accounts consider source information as capable of providing directional predictions, but given the opportunity for increased consideration, a belief recipient should instead defer to the evidence, minimising the influence of source characteristics. It is worth noting, that unlike in the present work (in which the truth value of evidence is always assumed to be 100%), these models are based on work in which “evidence” or arguments are either weak or strong (and accordingly have a variable truth value).

However, developments in coherence-based models have suggested a more gradated conceptualisation, involving domain expertise, trustworthiness, coherence, and back-up evidence in qualitatively formalising appeals to authority (Walton 1997, table 1, p. 102). Subsequently, Bovens and Hartmann (2003) developed this concept further in their proposal of a formal Bayesian model of reliability and coherence (see also Schum, 1981). This work has since provided the formal foundation for a Bayesian model of source credibility (Hahn et al., 2009, 2012).

The model (see equation 1, below) integrates expertise and trustworthiness in an independent and orthogonal manner, providing predictions for the posterior degrees of belief in the hypothesis (H) given the representation (Rep) by a source:

$$\mathrm{P(H|Rep)} = \frac{\mathrm{P(H)x\,P(Rep|H)}}{\mathrm{P(H)x\,P(Rep|H)} + \mathrm{P(\neg H)x\,P(Rep|\neg H)}}\,{}^{26} \qquad (1)$$

The model conceptualises expertise as referring to the degree of access to accurate information the source is believed to have for the domain in question. Trustworthiness, however, refers to the degree of belief in the source being willing to communicate information as faithfully as possible, to the best of the source's ability. In short, expertise refers to access whilst trustworthiness refers to intentions.

The model has found empirical support in argumentation (Harris et al., 2015) and political endorsements (Madsen, 2016), indicating that the model captures essential and predictive characteristics of appeals to authority, given both the fit with observed posterior degrees of belief, and statements from uncertain sources. Further, despite the model's development in cognitive psychology, the demonstrated impact of expertise and trustworthiness bear a close parallel to findings in social psychology.

[26] To integrate expertise, trustworthiness and conditional probabilities in model predictions, P(Rep|H) = P(Rep|H, Exp, T) * P(Exp) * P(T) + P(Rep|H, ¬Exp, T) * P(¬Exp) * P(T) + P(Rep|H, ¬Exp, ¬T) * P(¬Exp) * P(¬T) + P(Rep|H, Exp, ¬T) * P(Exp) * P(¬T); mutatis mutandis for P(Rep|¬H).

As mentioned previously, source reliability is conceptualised as an amalgamation of warmth (trust) and competence (expertise) traits (Cuddy et al., 2011; Fiske et al., 2007). Branching across findings in argumentation and social psychology studies, trustworthiness has generally been indicated as more influential than expertise regarding persuasion and influence. Aside from its effect, evidence also suggests that trustworthiness is easy to lose and hard to (re)gain whilst expertise is relatively more difficult to lose, but easier to (re)gain. As a consequence, when inferring new predictions from a Bayesian source credibility model to a decision making style paradigm, one initial prediction is that trustworthiness will impact decisions more than expertise. Given the previous empirical support for the model in belief revision tasks, as well as the consistency of social psychological findings, it is plausible that sensitivity to sources should extend behaviourally as well. Consequently, when integrating source credibility into the present paradigm, predictions are derived by taking the forwarded mechanisms of previous work (Chapters 3 & 4), and integrating directionality inspired by the Bayesian source credibility model.

5.1.3 Present Research

In the synthesis of manipulations of source credibility cues and initial evidence in the paradigm used in Chapter 4, there are several predictions regarding the interplay of these factors. It should be noted, however, that the present work does not seek to provide specific, concrete Bayesian predictions, but rather take inspiration from the Bayesian source credibility model in deriving quantified, behavioural predictions.

Previous work has found that given an unknown source, initial evidence plays a critical role in the consolidation of a belief (see Chapter 3 & 4). Initial evidence has thus been argued as playing a pivotal role in shaping not only the perceived validity of the communicated information, but by proxy the reliability of the source as well. Given this, we argue that much like an unknown source, when source credibility cues indicate that a source may be unreliable (and thus confidence in the truth of the belief is low), initial evidence will continue to play a role in consolidation and refutation (or "gatekeeping").

Conversely, if source credibility cues indicate the source of the belief as being reliable, in line with expectations of the Bayesian source credibility model, we predict that initial evidence will no longer play a role in consolidating or refuting the belief. Such a prediction is based on the notion of credibility cues acting as evidence to inform confidence in the belief being true – *prior to evidence exposure*. Furthermore, given this supplanting role of source credibility cues, and in line with Bayesian source credibility predictions, it can thus be predicted that more credible cues will (irrespective of initial evidence) elicit higher degrees of belief compliance.

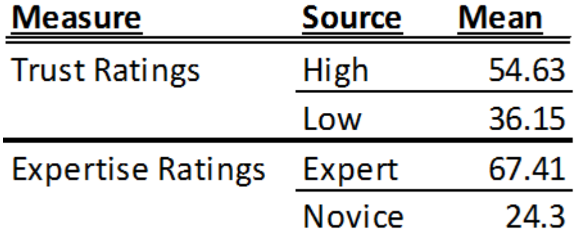
It is further worth noting that such predictions are distinct from those of alternative models of the influence of sources, such as the Elaboration Likelihood Model and the Heuristic-Systematic Model (Briñol & Petty, 2009; Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984). These models, given the present paradigm's combination of accuracy incentives, evidence-exposure, and the erroneous nature of the belief (see 4.2.1), predict abandonment of source cues in favour of evidence, and thus factors like source (un)trustworthiness should be overruled.

5.2 Pre-testing: Credibility Statements

Accordingly, to incorporate elements of source credibility into the current paradigm, their introduction is inherently interlinked to both the content, and communication medium, of the belief itself. In this way, to infer the behavioural consequence of, for example, a communicated belief from a high trust, expert source, it was necessary to determine how the manipulations of trustworthiness and expertise are interpreted *in light of* the content of the belief itself. However, it is stressed here that expertise and trustworthiness are considered independent of *one another* within the Bayesian source credibility model (Bovens & Hartmann, 2003; Hahn et al., 2012; Harris et al., 2015). Thus, to include source credibility elements into the previous work investigating beliefs, initial evidence, and consequent confirmation bias effects (Chapters 3 & 4), it is necessary to first understand how the combination of statements indicating components of source credibility, in conjunction with the belief itself, are interpreted prior to evidence exposure. Outlined below are the key outcomes from the pre-testing process required to understand the current experiments. Full details of the pre-testing process can be found in Appendix B.

Briefly, two questions were designed to assess the ratings of trustworthiness and expertise; "*The participant is* ***completely trustworthy*** *(i.e. the participant will be truthful* ***to the best of their abilities)****?*" (from 0; certainly false, to 100; certainly true) to assess trustworthiness, and "*The participant has* ***accurate knowledge*** *about the* ***effectiveness*** *of the medicines?*" (from 0; certainly false, to 100; certainly true) to assess expertise. These questions were similarly implemented in the full experiment versions. Further, this allowed for the validation of a high and low trust statement, and a high and low expertise statement. The statement validated for high trustworthiness was "*The participant below was told they would be* ***paid double*** *if the next participant group performed* ***better*** *than them.*", and for low trustworthiness was "*The participant below was told they would be* ***paid double*** *if the next participant group performed* ***worse*** *than them.*". Whilst the validated statement for high expertise was "*This participant was asked to make a comment after completing* ***all*** *of the 1000 trials.*", and for low expertise was "*This participant was asked to make a comment after completing* ***only 1*** *of the 1000 trials.*". These two sets of statements were thus used in conjunction with the belief statement to create the *Trust* (high or low) x *Expertise* (high or low) source credibility independent variables.

Table 5.2.1: Pre-test: Expertise and Trust Ratings for Source Credibility Statements



Measure	Source	Mean
Trust Ratings	High	54.63
	Low	36.15
Expertise Ratings	Expert	67.41
	Novice	24.3

As can be seen in Table 5.2.1, the ratings of the corresponding high and low trust and expertise statements obtained from the pre-testing demonstrates that such manipulations do provide the desired directional impact on credibility variables of interest. As mentioned previously, full details of how such results were obtained can be found in Appendix B.

5.3 Experiment 7: False Prophets

The purpose of Experiment 7 is to assess the role source credibility factors play in the uptake and maintenance of a communicated belief, when the belief recipient is subsequently exposed to a prolonged period of first-hand evidence. Moreover, following previous work in Chapters 3 & 4, Experiment 7 focused on *erroneous* beliefs (i.e. all belief manipulations are factually incorrect, by suggesting the probabilistically inferior option is in fact superior). The questions that can then be asked using this paradigm have immediate, real world relevance.

Firstly, if initial evidence, previously found to be a gatekeeper to consolidating erroneous beliefs (see Chapter 4, section 4.4), is found to be overlooked in cases where the belief comes from a trustworthy source, a neat parallel can be drawn to the potential dangers of abuse of authority in the spread of misinformation. It should be noted, however, that appeals to authority are not therefore implied to be irrational (Goodwin, 2011; Harris et al., 2015; Walton, 1997). For example, it is advantageous to take advice from a village elder about the potential dangers of the nearby forest, rather than solely relying on first-hand evidence. The experiment is designed to focus specifically on the instances of falsehoods, it cannot pass comment on the general proportion of valid to invalid belief dissemination. Such a finding, wherein source credibility trumps initial evidence, would suggest that initial evidence plays a secondary role. However, conditions in which low credibility beliefs are supported by initial evidence will help answer such a prioritization (i.e. if a belief is already deemed suspicious, can initial evidence redeem it?).

Secondly, it is possible to assess how low credibility sources are interpreted. More precisely, is low trustworthiness taken directly as a sign of ulterior motives (Twyman et al., 2008), and consequent choices then reflect this assumption, whereby recipients choose the opposite option from that indicated by the belief? Importantly, not only does this neatly demonstrate belief processing within a social context, but moreover, if biasing effects are found in support of the *assumption*, then a more robust account of beliefs-as-hypotheses has been demonstrated. In other words, one can use source trustworthiness to "flip" the biasing effect of the belief in either direction.

Finally, not only can the impact of beliefs and source credibility factors be seen on subsequent choices, but the impact of the probabilistic (refuting) evidence experienced over time can be seen in both the maintenance of beliefs *and* the assessments of credibility. Although such a comparison is made across experiments (pre-test to Experiment 7), how ratings of trust and expertise of the original source of the belief are affected by the combination of a communicated belief and sustained, probabilistic evidence (that fails to support it) has important ramifications for the cyclical, long-term impact of sources. For example, assuming high trust sources provoke lower amounts of scepticism and consequently lead to higher proportions of "believers" (despite the truly erroneous nature of the belief), if trust is then updated even higher (in accordance with the supposition that the recipient *believes* they have been told the truth), then no correction occurs and instead the source becomes more dangerous / unassailable. In the same way, the reverse of this may occur when confronted with beliefs from a low credibility source (higher scepticism, less bias, more refuters, lower credibility ratings) in line with work on attitudes (Priester & Petty, 1995). Importantly, unlike the preceding predictions (also see 5.1.3); the analysis of the impact of belief-evaluation on the ratings of trust and expertise is considered tentative, as the current method does not allow for direct within-subject comparison of credibility ratings before and after evidence exposure.

5.3.1 Method

Following the outline set out above, Experiment 7 incorporated the health context evidence integration task used in the previous Chapter, along with the new comment section that allowed for the manipulation of source credibility factors.

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 500 prior HITs. Participants completed the experiment under the assumption that the purpose of the investigation was to improve medical decision making. Participants were English speakers between the ages of 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

Design. For each of the two diseases, their two medicine options (“Mox” and “Nep” for the “Lannixis” disease, and “Byt” and “Zol” for the “Deswir” disease), all generated either “Cured” of “No Effect” outcomes, with one medicine for each disease curing at a 60% rate (optimal option), and one medicine curing at a 40% rate (sub-optimal option). The instructions given to participants explained that each trial is a new patient presenting with one of the two diseases, and it is their job to prescribe one of the two medicine options for that disease for the patient, and assess the outcome. Successful cures earned them points (+3), whilst failures to cure cost them 1 point. Participants were further instructed that each patient may react differently to the medicines, so it is their job to discern the *overall* efficacy of the medicine options. Participants were further incentivized by a performance bonus on top of the standard payment, based on the amount of points earned.

Before starting the trials, all participants were exposed to a comment from a previous participant regarding **one** of the two diseases (counterbalanced), the comment regarded one of the two medicines (always the **sub-optimal** option) being better. This comment was accompanied by the manipulation of the two source credibility factors: a statement regarding the trustworthiness of the source and a statement regarding the expertise of the source. Both of these between-subject factors were randomized independently to be either low or high.

Which medicines were optimal and sub-optimal were also counterbalanced. In this way, the first factor of interest in the design is the within-subject difference in the choices and judgements for the "control" disease (the disease that did not have a comment section), and the "belief" disease. As the sole difference between the two diseases was the presence or absence of communicated beliefs, any subsequent difference could be discerned as due to this within-subject manipulation.

After source credibility and belief, the next factor of interest follows from the previous Chapter. An initial evidence manipulation was applied to the current methodology, the result being a further between-subject (2-level) factor. Having received the belief, the first few trials of evidence either support it (initially supportive condition; IE+), or undermine it (initially undermining condition; IE-). To explain further, given that the belief (falsely) indicates the sub-optimal option as superior, in the supporting initial evidence (IE+) condition, the sub-optimal options for *both* diseases receive two positive trials (cures), followed by one negative (no effect), whilst the optimal options for *both* diseases receive the opposite pattern; two negative trials (no effect), followed by one positive (cure). Conversely, the undermining initial evidence (IE-) condition follows the opposite pattern, with the *optimal* options now receiving two positive trials followed by one negative, and the *sub-optimal* options now receiving two negative trials followed by one positive. In this way, for the belief disease, the belief receives initially undermining evidence (much in the same way as the Belief Initially Undermined (BIU) groups of Chapter 3 experiments; see 3.2 for full description). All trials following this three-trial manipulation followed the 60/40 probability distribution outlined above.

In summary, there were, along with the within-subject belief-control disease difference, three between subject factors under investigation: initial evidence (supportive or undermining), source trustworthiness (high / trustworthy, or low / untrustworthy), and source expertise (high / expert, or low / novice). Whereas in previous experiments within this thesis, the counterfactual outcomes were either constantly present (see sections 3.2, 3.3, 3.4, & 3.5), or had their presence / absence manipulated as a factor (see section 4.2 & 4.3), in this paradigm their *absence* was held constant. This decision was made for three reasons. Firstly, previous work has demonstrated counterfactual presence or absence plays a minimal role in the belief consolidation process (see section 4.4.2). Secondly, the goal of the present chapter was not to explicitly test integrative versus selective bias explanations (see Chapters 3 & 4 for such a focus), but rather the interplay of initial evidence and source credibility. Thirdly, given the preceding reasons indicating manipulation being unnecessary, to provide a more realistic task context, the decision was taken to hold counterfactuals absent. The main dependent variables under investigation were the proportion of choices made in favour of the sub-optimal medicine, posterior measures of binary preferences, confidence in that preference, and probability estimates. Lastly, ratings of the trustworthiness and expertise of the source were also added, following evidence exposure.

Procedure. Before starting the trials, participants were shown the "comment section" in which previous participants had written their thoughts regarding one of the diseases, and the task in general. The comment was rigged to appear to be from a previous MTurk participant (complete with fake MTurk ID number), and indicated a directional hypothesis regarding one of the medicines for the disease ("I think the Zol medicine was the most effective"). The medicine indicated as superior, was in fact always (unknown to the participant) the sub-optimal option. Along with this comment were the two statements regarding the trustworthiness and expertise of the previous participant (each either low or high, and randomly assigned between subjects).

On each trial participants selected which of the two medicines to prescribe to the patient, with “Mox” and “Nep” the two options for when the patient presented with “Lannixis” disease, and “Byt” and “Zol” for presentations of the “Deswir” disease. Which disease was the control condition and which was the “belief” condition was counterbalanced between participants, along with which medicine in each pair was set to be optimal or sub-optimal. For each trial, the side of the screen on which each medicine was shown was randomized. Participants were invited to earn as many points as possible, based on the number of cures. Each trial cost participants one point, so a failure cure resulted in a net loss of −1 point, whilst a cure earned 3 points. Participants were aware of their current total points earned during the trials, and instructed that their total amount of points directly corresponded to an increasing bonus payment in dollars, on top of the standard payment.

For each trial, participants selected one of the two medicine options to prescribe to the patient, generating the outcomes. The outcomes were “Cured” written in green if the medicine led to a cure, and “No Effect” in red if the medicine was unsuccessful. The outcome text was surrounded by a highlight box of the same colour (see Appendix A.3.2 for an example feedback screen).

Once all 100 trials (50 per disease, alternating each trial) were completed, participants then completed posterior measures for each of the diseases. This consisted of a binary preference for that diseases pair of medicines, the confidence in that preference, and a probability estimate for the distribution of cures between those medicines. Following this, participants were asked if they could recall the previous participants comment (manipulation check), after which they could then post an open text response “comment” in the comment section they had seen before the task. Participants were instructed that when posting their comment, they would receive an additional bonus if they either successfully deceived or aided the participant that followed them. The instruction to deceive or help was randomized between-subjects, and further assisted the validity of the context for the trust statement they had seen regarding the previous participant. This manipulation was added to ensure the participants bought into the realism of the comment context (i.e. their commenting procedure fitted with the “previous participant” comment they had seen). Upon posting their comment, participants then completed the trust and expertise ratings for the previous participant (in the same format as that used in the pre-test methodology, see 5.2). Finally, this was followed by a demographics questionnaire and the Need for Closure measure (Roets & Van Hiel, 2011; Webster & Kruglanski, 1994). Following completion of the task, participants were debriefed and given an email to contact if they had any further questions.

Table 5.3.1 below summarises the key information from the above methodological description. This includes the phrasing of the task instructions, incentives scheme, and belief manipulation, as well as the measures taken (including posteriors and manipulation check question phrasing).

Table 5.3.1: Experiment 7: Summary table of task setup and measures.

Setup / Manipulations	**Description**	**Details**
Task Instructions	Formal instructions given to participants from the experimenter on how to perform the task.	"In this task, you will be attempting to cure cases of two different diseases. Each disease has a pair of medicines designed to cure it... ...Your objective is to try and cure as many patients as you can by learning the **effectiveness** of the medicines. Each successful cure will earn you points. The medicines both cost 1 point to take. At the end **your score will determine your bonus in dollars as follows:**"
Incentive Scheme	Bonus scheme outlined to participants based on performance. With each increasing points boundary, the change in bonus also increases. [Full scheme not shown to participants, only first three levels to indicate increasing performance bonuses.]	**Total Payment, based on points:** <50 points = standard $.50, >50 points = $.60, >100 points = $.75, >150 points = $.90, >200 points = $1.05, >250 points = $1.20, >300 points = $1.40
Belief Manipulation	Single isolated comment. Presented to participants with accompanying trust and expertise statements (see section 5.2).	**Manipulation Comment**: "I think the Byt medicine is the most effective"(see Appendix A.3.1 for Sample comment screen with trust and expertise manipulations included)

Measures	**Description**	**Wording**	**Values**
Choice Data	Binary forced choice between two medicines for each trial. 50 trials for each disease (alternating), to make 100 trials total. Each disease had an optimal (60%) and suboptimal (40%) option.	n/a	Byt, Zol / Mox, Nep
Binary Preference	Posterior Measure: Following trials, participants were asked which of the two machines they preferred.	"Which medicine do you think is better?"	Byt, Zol / Mox, Nep
Confidence in Binary Preference	Posterior Measure: Having given their binary preference, participants were asked how confident they were in their preference.	"How confident are you in this preference?"	0-100 slider (default value of 0)
Probability Estimate	Posterior Measure: Participants were asked to give a probability estimate of the distribution of cures between the two medicines. **Note: All posterior measures were completed sequentially on two screens, split by disease.**	"What is the distribution of cures between the two medicines?"	100% Byt / Mox, through 50/50, to 100% Zol / Nep (slider, default value 50/50)
Trust Rating	Participants were asked (following all other measures) to rate a statement regarding the trustworthiness of the previous participant.	"The participant is **completely trustworthy** (i.e. the participant will be truthful **to the best of their abilities**)?"	0 (certainly false) to 100 (certainly true)
Expertise Rating	Participants were asked (following all other measures) to rate a statement regarding the expertise of the previous participant.	"The participant has **accurate knowledge** about the **effectiveness** of the medicines?"	0 (certainly false) to 100 (certainly true)
Manipulation Check	Questions asked of participants to determine if participants still recalled manipulation comment by the end of the trial procedure.	"Do you think comments were biased towards one medicine?"	Byt, Zol, Mox, Nep, No

5.3.2 Results

Descriptives and Processing. In addition to the 2 between subject factors from the Pre-tests, source trustworthiness (high or low; hereafter termed *trust*) and source expertise (expert or novice; hereafter termed *expertise*), the additional factor of *initial evidence* (supporting or undermining) was added. This resulted in 8 groups for analysis (see Table 5.3.1 below). From the effect size of the *Belief* x *Initial Evidence* interaction in choice data of previous experiments (i.e. Chapter 4: Experiment 5) and the effects found in pilot-testing source credibility effects, the most conservative was selected for a power analysis, run using G*power (Faul et al., 2009, 2007), to estimate sample sizes required for Experiment 7. This resulted in an estimate of 50 participants per group, which when multiplied by a conservative expected forgetting rate, and the 8 groups, resulted in a total sample size of 520.

Table 5.3.2: Experiment 7: Participant Breakdown by Group, along with the number of participants passing the belief manipulation check.

Trust					*N*	Passed Manipulation Check
High	**Expertise**	High	**Initial Evidence**	Supporting	63	52
				Undermining	60	50
		Low	**Initial Evidence**	Supporting	66	49
				Undermining	67	50
Low	**Expertise**	High	**Initial Evidence**	Supporting	64	52
				Undermining	68	49
		Low	**Initial Evidence**	Supporting	65	48
				Undermining	72	51

Participants were randomized into one of the eight possible conditions (see the first column of Table 5.3.2). The average age was 36.85 years (*SD* = 12.424) and the sample was 55.2% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of the manipulation comment they were removed from subsequent analysis, the remaining group numbers are shown in the second column of Table 5.3.2. The decision to remove those who failed was taken following the same protocol and reasoning as in previous Chapters (although this is additionally critical given the now explicit nature of the belief manipulation, in conjunction with trust and expertise statements). The following analyses were conducted using the remaining 401 participants, with an average age of 36.54 years (*SD* = 12.253) and 56.6% female. The results below will not be presented exhaustively; for the sake of brevity, only those of central interest to the thesis questions are included.

Correlates. There were no unexpected correlations regarding counterbalancing factors such as disease or medicine outcome assignments. Importantly, a significant correlation was found between the degree of bias (taken as difference between belief pair and non-belief pair) in choice data and the degree of bias in posteriors, $r = .537$, $N = 401$, $p < .001$. This demonstrated corroboration between the impact of *belief* in trial-by-trial choices and in participants' end-of-sequence judgements.

Choice Data. To assess the impact of the *belief, initial evidence* and source credibility manipulations on choices, a mixed ANOVA was run using the total number of optimal choices for the belief disease and the total number of optimal choices for the control disease as the two-level within-subjects factor (*belief*). The between-subject factors included in the analysis were *initial evidence, trust*, and *expertise*. As can be seen in Figure 5.3.1, there were significant main effects of *belief*, $F(1,393) = 6.272$, $p = .013$, $\eta^2 = .016$, *initial evidence*, $F(1,393) = 24.803$, $p < .001$, $\eta^2 = .059$, and *trust*, $F(1,393) = 6.462$, $p = .011$, $\eta^2 = .016$, whilst *expertise*, although trending towards significance, showed no main effect ($p = .096$).

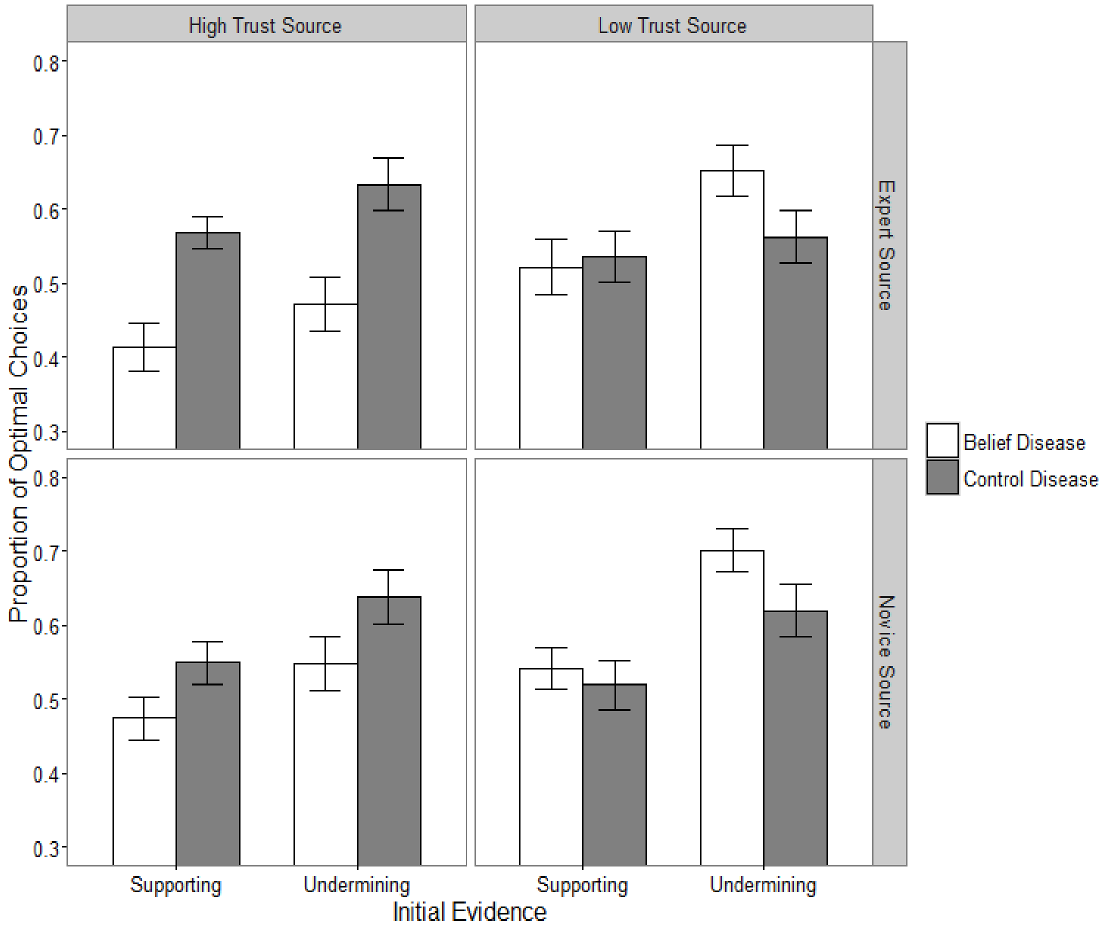


Figure 5.3.1. Experiment 7: Proportion of Optimal Choices. White bars present the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

These main effects showed that significantly more choices were made for the sub-optimal medicine when favoured by initial evidence, having received a belief favouring that option, or receiving a belief from a high trust source. Furthermore, there was a significant interaction between *belief* and *trust*, $F(1,393) = 29.926$, $p < .001$, $\eta^2 = .071$, indicating that the conjunction of both a belief *and* a high trust source results in significantly more suboptimal choices. Visual inspection of Figure 5.3.1 indicates a pattern of coherence / incoherence between *initial evidence* and *trust*, which is further explored in the sub-group analyses below. The decision to split by *initial evidence* is further supported by the expected moderating effects of initial evidence on belief uptake, and its predicted selective impact depending on source credibility (see 5.3).

Splitting by Initial Evidence. By splitting the sample into supporting and undermining initial evidence conditions and running a mixed ANOVA for *belief* (within-subject factor), *trust*, and *expertise*, it was possible to start determining the strength of possible coherence effects. Unsurprisingly, those in the supporting initial evidence condition showed a significant main effect of *belief*, $F(1,197) = 7.834$, $p = .006$, $\eta^2 = .038$, supporting effects found in previous experiments (see 4.4). Although there was no significant main effect of *trust*, it did interact significantly with *belief*, $F(1,197) = 9.105$, $p = .003$, $\eta^2 = .044$. This latter effect showed that the coherence of both high trust and supporting initial evidence resulted in biased choices, whilst a belief from an untrustworthy source which receives supporting evidence does not lead to such a bias.

Conversely, those in undermining initial evidence conditions do not show a main effect of *belief*, but do show a main effect of *trust*, $F(1,196) = 5.5479$, $p = .02$, $\eta^2 = .027$. However, the reason for the lack of a main effect of *belief* is likely due to the strong interaction between *belief* and *trust*, $F(1,196) = 21.33$, $p < .001$, $\eta^2 = .098$, in this case showing that the coherence of low trust and undermining initial evidence resulted in biased choices in the *opposite* direction to those in the incoherent, high trust, undermining initial evidence condition. There are two effects of consequence here, the first of which is that whilst low trust and supporting evidence failed to produce a bias, the reverse (high trust, but undermining initial evidence) procured a bias despite the proposed refuting effect of undermining initial evidence. This finding corroborates the second prediction regarding the impact of source credibility in subsuming the role of initial evidence.

The second effect of interest is that source trustworthiness is predictive of choice directionality[27]. The interaction between *belief* and *trust* in undermining initial evidence groups demonstrates not only that beliefs are processed *in the context of source cues*, but that the resultant hypothesis ("I do not trust the person telling me "A is better", so I believe B is in fact better") results in the same form of biasing effect as when initial evidence supports a believed hypothesis (i.e. initial evidence is undermining a belief they already believe to be untrue, *confirming the lie*), just in the opposite direction. Taken in conjunction with the finding that incoherence between supporting initial evidence and an untrustworthy source (i.e. the evidence does not fit the suspicion of the belief being false) results in no impact of belief in either direction, such a finding corroborates the first prediction regarding the preservation of the role of initial evidence when confidence in the reliability of the source is low.

Block Data: Assessing Learning. To assess the impact of the within (*belief*) and between-subject (*initial evidence*, *trust*, and *expertise*) factors on learning, so as to determine potential limitations in the impact of these factors across choices, an additional choice data analysis was conducted using 25 trial blocks for each disease. In this way, it was possible to assess whether the impact of these factors diminishes with experience in the task. Accordingly, an additional within-subjects factor (hereafter termed *block*) was added to the mixed ANOVA used for total choices, replacing the total choices DVs for the 1st and 2nd block of the belief disease, and the 1st and 2nd block of the control disease. Such an analysis allowed for the assessment of *block* and its interactions with the factors of interest – all other effects (those not including *block*) reflect those found in the total choice analysis above.

[27] This claim is supported by a Chi squared analysis on first choice data, finding a highly significant effect of *trust*, χ^2 (1, N = 401) = 95.373, $p < .001$, with those in low trust groups more often choosing the *non-specified* option, rather than the option specified by the belief, whilst the reverse is true of high trust groups.

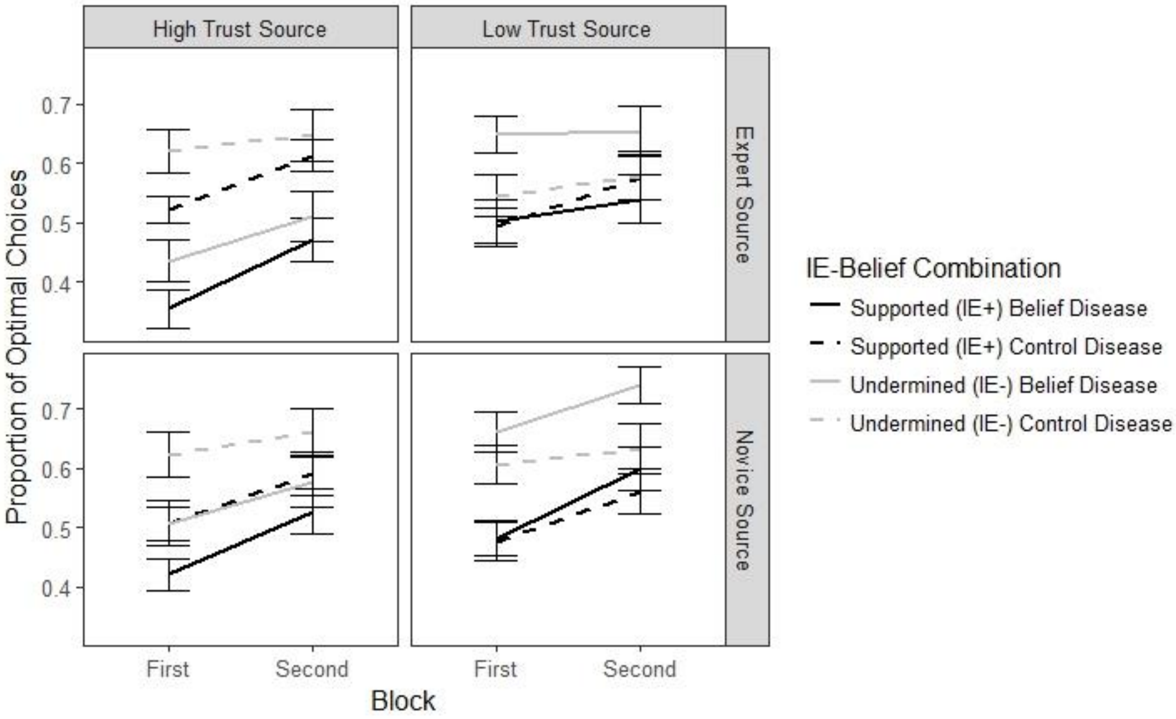


Figure 5.3.2. Experiment 7: Proportion of Optimal Choices, split by 25 trial block. Line colour reflects initial evidence condition (black = *supporting* initial evidence conditions, grey = *undermining* initial evidence conditions), line type reflects within-subject disease (solid = *belief* disease, *dashed* = control disease). Error bars reflect Between-subject Standard Errors.

Accordingly, there was a main effect of *block*, $F(1,393) = 75.43$, $p < .001$, $\eta^2 = .161$, wherein participants chose the optimal option within a pair more often as trials progressed. *Block* also interacted with *initial evidence* (black lines show greater change across blocks than grey lines in Figure 5.3.2 above), $F(1,393) = 8.3$, $p = .004$, $\eta^2 = .021$. Taken together, this shows participants chose the optimal option more often as they moved through the task (i.e. a learning effect), which is exaggerated by the more (sub-)optimal starting points based on *initial evidence* (black lines start lower, Figure 5.3.2). Further, it is worth noting the main effects of *trust*, and its interaction with *belief*, is not found to be reduced across blocks, suggesting the impact of *trust* is not reduced over time.

Posteriors. Where appropriate, the same format of mixed ANOVA as that conducted in the total choices analysis was used to assess the impact of the independent variables on posterior judgements. Given the extended analyses warranted for this experiment, the posteriors were further split into sections by dependent variable.

Probability Estimates. Using this protocol, the effect of the independent variables on posterior probability estimates in the belief and control diseases (as the within-subject, *belief*, factor) was assessed. Although there were significant main effects of both *belief*, $F(1,393) = 4.432$, $p = .036$, $\eta^2 = .011$, and *initial evidence*, $F(1,393) = 16.656$, $p < .001$, $\eta^2 = .041$, there were no main effects for either *trust* or *expertise*. Interestingly, the interaction between *belief* and *trust* was significant, $F(1,393) = 3.888$, $p = .049$, $\eta^2 = .01$, indicating that those who received a belief that was coupled with a high trust source showed a greater degree of bias in posterior probability estimates (left hand column of Figure 5.3.3).

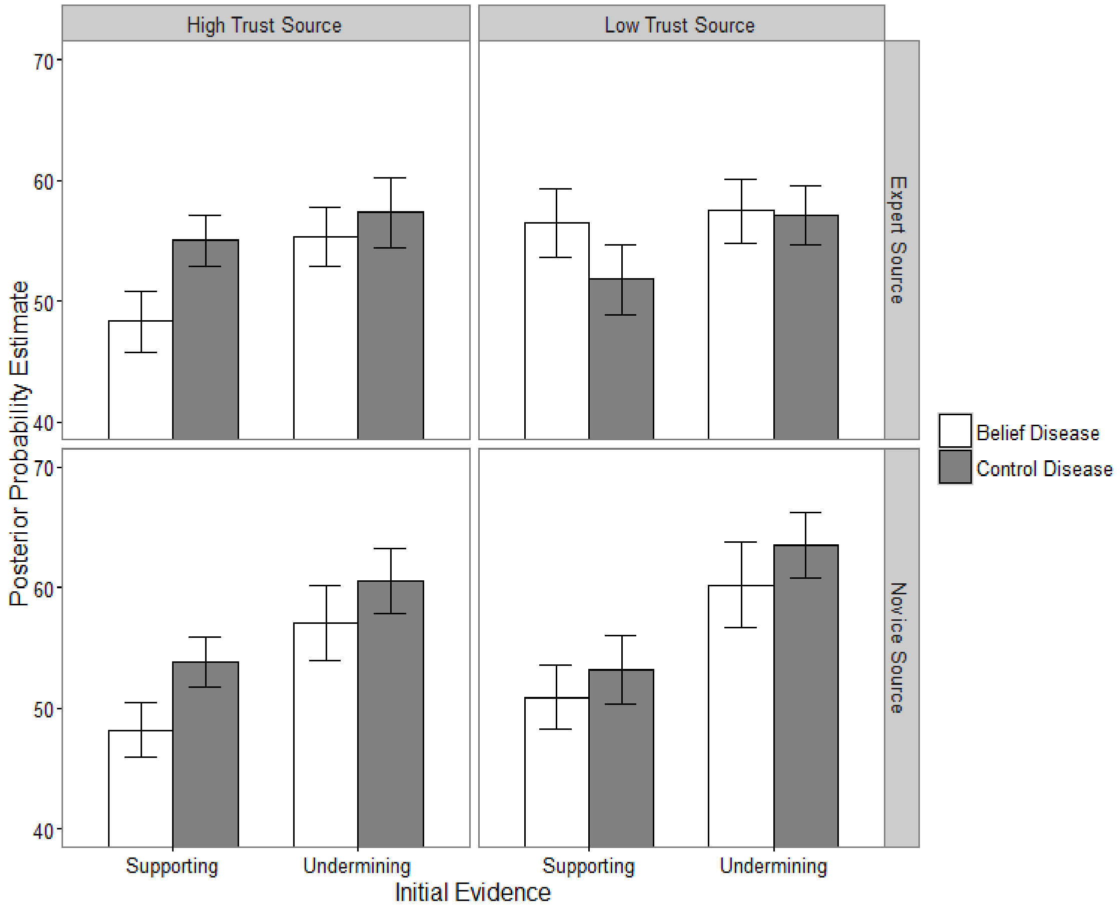


Figure 5.3.3. Experiment 7: Posterior Probability Estimates (estimate of percentage of optimal outcomes in favour of initially dominant medicine), split by group. Error bars reflect Between-subject Standard Errors.

In general, trends indicate the effects found in probability estimates are subsumed by the impact of *initial evidence*. One notable trend is that when moving from left to right across the columns of Figure 5.3.3, as the number of factors that should influence poor choices decreases, the number of optimal choices increases (note: both belief and control diseases have their *initial evidence* manipulated in the same direction, whilst *trust* and *expertise* manipulations are localized to the belief disease).

Binary Preferences and Confidence. The exception to the above mixed ANOVA protocol is for the binary preference, which required a different analysis due to the binary nature of the dependent variable. As such, a mixed-effects logistic regression was run in R to test for each main effect, and if appropriate, possible interactions between the between-subject (fixed) and within-subject factors.

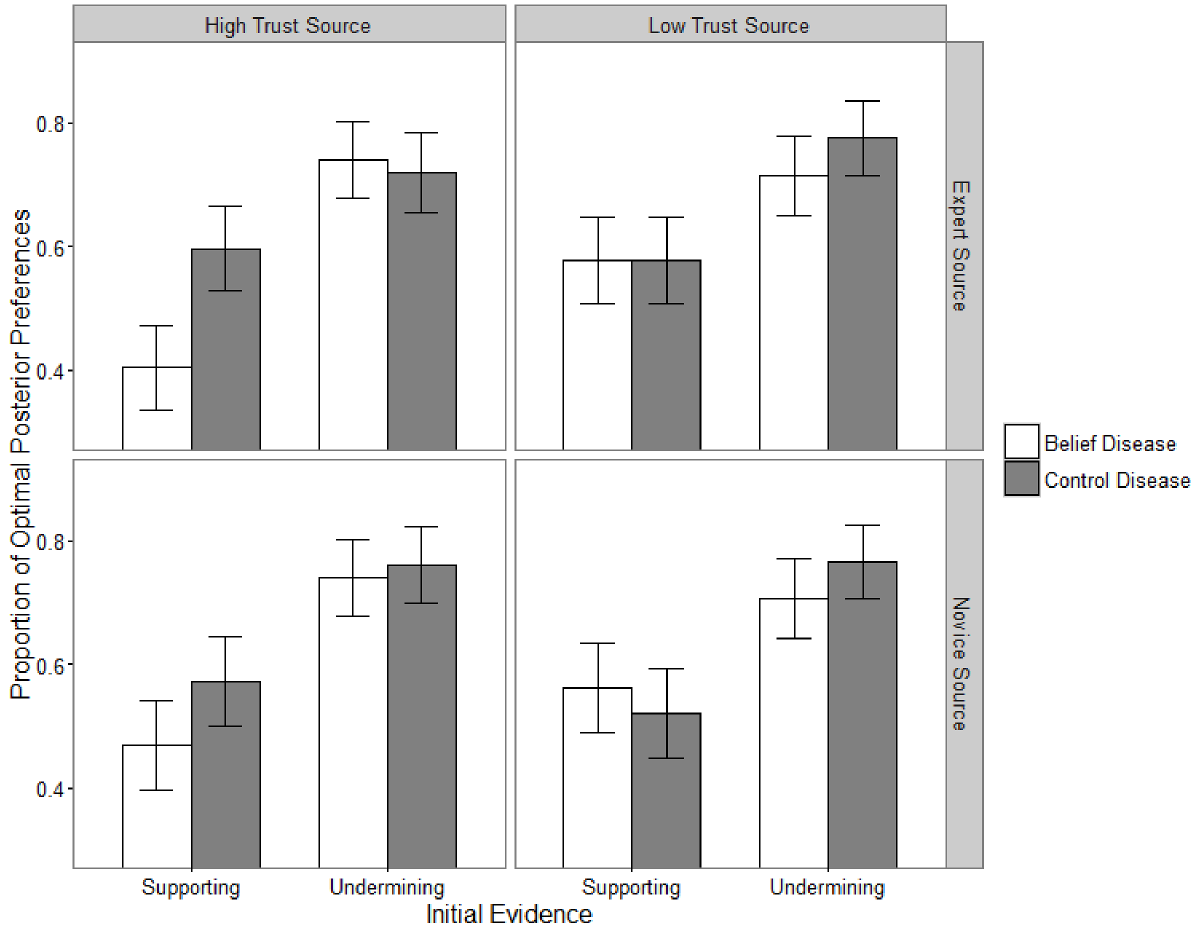


Figure 5.3.4. Experiment 7: Binary Preferences (proportion of participants indicating a preference for the optimal option, split by group). White bars represent the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

Such an analysis compared sequentially more complex models, first finding a significant improvement over the basic model (fixed intercept + subject random-effect) through the inclusion of *initial evidence* as a factor (left-hand pairs of bars within each facet are lower than right-hand in Figure 5.3.4), χ^2 (1) = 35.174, p < .001. However, as no main effects were found for *belief*, χ^2 (1) = 2.0943, p = .148, *trust*, χ^2 (1) = 0.5406, p = .4622, or *expertise*, χ^2 (1) = .0102, p = .9195, no further models including interaction terms were assessed. Such a finding suggests, in line with posterior probability estimates, that after such prolonged evidence exposure, the impact of source cues accompanying the communicated belief have been reduced.

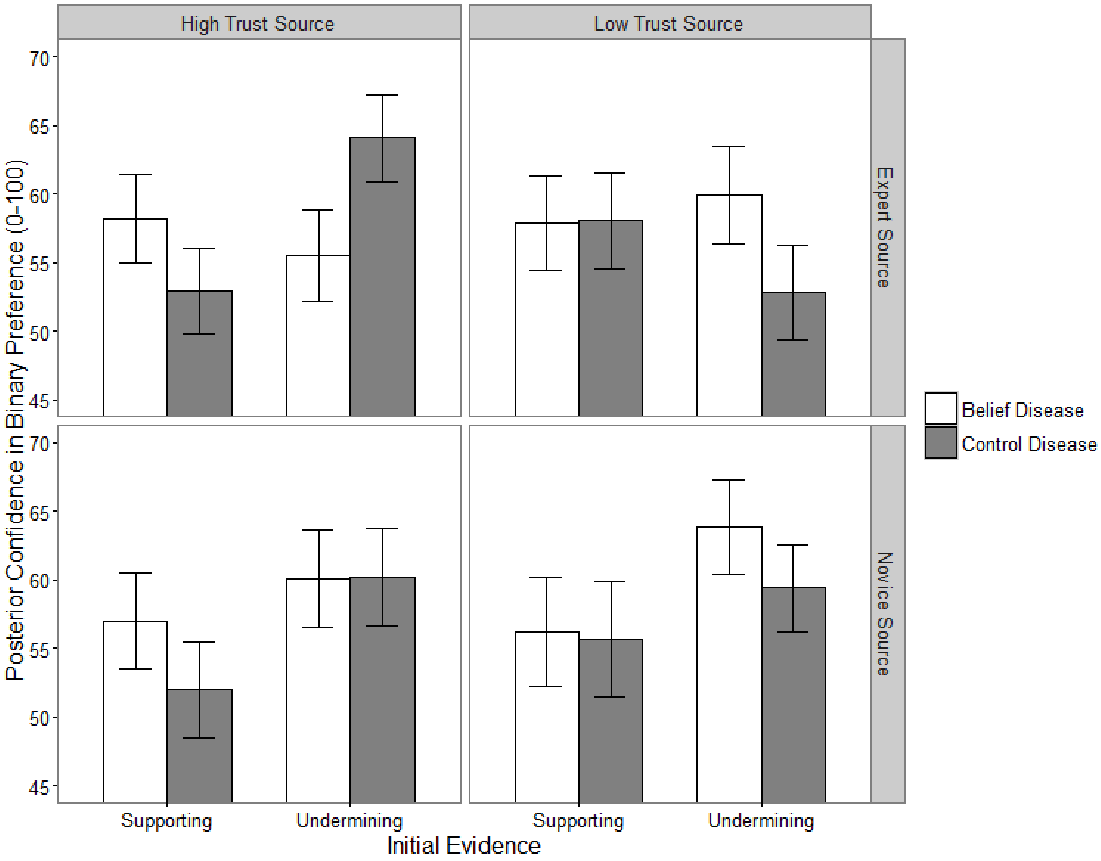


Figure 5.3.5. Experiment 7: Confidence in Binary Preference (0-100%), split by group. Error bars reflect Between-subject Standard Errors.

Regarding confidence in binary preferences (and returning to the mixed ANOVA), the belief disease posterior confidence and control disease posterior confidence were used as the within-subject (*belief*) 2-level factor, along with *initial evidence*, *trust*, and *expertise* as between-subject factors. Although there was no main effect of *initial evidence*, *belief*, *trust* or *expertise*, there was however a *Belief* x *Trust* x *Initial Evidence* interaction, $F(1,393) = 15.994$, $p < .001$, $\eta^2 = .039$. Investigating this interaction further, in the same manner as the analysis of the overall choice data, a secondary mixed ANOVA was conducted by splitting participants by *initial evidence* condition (left-hand versus right-hand pairs of columns in Figure 5.3.5 above).

Splitting by Initial Evidence. In line with previous work (Chapters 3 & 4), those in supporting initial evidence conditions showed a significant main effect of *belief* (white bars higher than grey in left-hand columns of Figure 5.3.5), $F(1,197) = 4.945$, $p = .027$, $\eta^2 = .024$, on confidence. Further, although *trust* did not have a significant main effect on confidence, it did interact significantly with *belief* (left-hand versus right-hand facets, left-hand columns, difference between white and grey bars, Figure 5.3.5), $F(1,197) = 4.284$, $p = .04$, $\eta^2 = .021$. This latter effect shows that the coherence of a belief supported by both high trust and initial evidence results in greater confidence, whilst a belief from an untrustworthy source which still receives supporting evidence leads to no such difference. Such a finding fits with predictions regarding the preservation of a gatekeeping initial evidence role, given untrustworthy sources. Conversely, those in undermining initial evidence conditions do not show a main effect of *belief* or *trust*, but do show a strong interaction *between belief* and *trust* (left-hand versus right-hand facets, right-hand columns, difference between white and grey bars, Figure 5.3.5, $F(1,196) = 12.093$, $p = .001$, $\eta^2 = .058$, in this case showing the coherence of low trust and undermining initial evidence results in greater confidence that the *opposite* option to that indicated by the belief is true.

Ratings of Trust. An analysis of variance was conducted to assess the effect of the *trust* manipulation as a factor on posterior ratings of trust. This analysis demonstrated that irrespective of group and experienced evidence, those in high trust groups rated the source as significantly more trustworthy ($M = 64.84$, $SD = 27.321$) than those in low trust groups ($M = 26.92$, $SD = 24.062$), $F(1,400) = 217.518$, $p < .001$, $\eta^2 = .353$. Following this initial manipulation efficacy assessment, there were two questions remaining regarding the ratings of trust; firstly, how does first-hand evidence exposure affect ratings of trust, and secondly, if such evidence factors affect trust ratings, how do such ratings compare to pre-test values for the source, prior to evidence exposure? It should be noted that the following analyses are considered tentative.

Evidence on Trust. To assess the role that experienced evidence played on trust ratings, the analysis was initially restricted to look at the impact of having *refuted* the belief put forward by the source. To do this, two posterior measures were used, the binary preferences and probability estimates. Using the former of these, those who had refuted the belief (assessed using the posterior binary preference for the belief disease – with refutation as selection of the non-belief medicine option) rated the trust of the source significantly lower (M = 38.81, SD = 31.268) than those who adhered to the belief (M = 56.69, SD = 29.785), F(1,400) = 15.529, p < .001, η^2 = .037. More interestingly still, this reduction in trust is primarily driven by the high expertise conditions. When splitting by *expertise* factor and re-running the analyses, high expertise conditions show this significant reduction in trust as a consequence of belief refutation, F(1,202) = 16.439, p < .001, η^2 = .076, whilst low expertise conditions show no such significant reduction, *ns* (p = .121). This suggests that those in low expertise conditions are in effect "excused" for getting the belief wrong (or rather, are not necessarily expected to get it right), whilst experts suffer the reduction in trust as they should have known better. This raises questions regarding the independence of expertise and trust within the Bayesian source credibility model (Bovens & Hartmann, 2003; Hahn et al., 2012; Harris et al., 2015), however, such an inference is albeit constrained so far by situational specificity.

Using the posterior probability estimates, a similar overall effect is found, whereby trust ratings are negatively correlated with posterior probability estimate judgements, r = -.193, N = 401, p < .001, indicating those who are more inclined towards the belief indicated medicine also rate the trust in the source as higher.

Pre-test Comparison. By comparing Experiment 7's posterior ratings of trust to the pre-test ratings of trust, it was possible to determine the impact of the various conditions (and first-hand evidence) on the trustworthiness of a source that has communicated an erroneous belief.

In theory, given that all groups are receiving erroneous beliefs, all trust values should decrease having experienced evidence. However, there is instead a polarizing effect of evidence: Those receiving information from high trust sources rate the trustworthiness of the source as even higher having experienced evidence (pre-evidence, *M* = 54.35, *SD* = 33.169; post-evidence, *M* = 64.84, *SD* = 27.321), *F*(1,288) = 7.887, *p* = .005, η^2 = .027, (evidence that should undermine the belief, and trust). Conversely, those receiving beliefs from low trust sources rate the trustworthiness of the source as even lower (pre-evidence, *M* = 35.94, *SD* = 28.113; post-evidence, *M* = 26.92, *SD* = 24.062), *F*(1,299) = 8.364, *p* = .004, η^2 = .027, having confirmed their suspicions. These effects may be attributed to the ambiguity of evidence, allowing for multiple interpretations. By splitting this analysis into supporting and undermining initial evidence groups, it is clear such effects are driven by congruence between *trust* and *initial evidence* (as mentioned in choice data and posterior measures). To explain further, the increase in trust ratings as a consequence of high trust sources is driven by the supporting initial evidence group, *F*(1,188) = 10.278, *p* = .002, η^2 = .052, whilst the decrease in trust ratings as a consequence of low trust sources is driven by the undermining initial evidence group, *F*(1,199) = 9.713, *p* = .002, η^2 = .047. In contrast to these congruent group effects, groups in which *initial evidence* and *trust* are incongruent (e.g., high trust but undermining initial evidence) show no significant effects. Taken together, the analysis of trust ratings indicates (albeit tentatively) that highly trusted sources not only retain, but increase their perception of trust, via higher levels of belief uptake and maintenance, despite communicating a falsehood. Conversely, those initially perceived to be untrustworthy, via correspondingly low levels of belief uptake and maintenance (and increased scepticism), have their ratings of trust lowered further by communicating a falsehood.

Ratings of Expertise. An analysis of variance was conducted to assess the effect of the *expertise* manipulation as a factor on posterior ratings of expertise. This analysis demonstrated that irrespective of group and experienced evidence, those in high expertise groups rated the source as significantly more expert (*M* = 57.37, *SD* = 27.27) than those in low expertise groups (*M* = 20.75, *SD* = 26.867), *F*(1,400) = 183.432, *p* < .001, η^2 = .315. Following this initial manipulation efficacy assessment, there were again two questions remaining regarding the ratings of expertise; firstly, much like the assessment of trust, how does first-hand evidence exposure affect ratings of expertise, and secondly, if such evidence factors affect expertise ratings, how do such ratings compare to pre-test values for the source, prior to evidence exposure? As with ratings of trust, it should be noted that the following analyses are considered exploratory, and findings are consequently considered tentative.

Evidence on Expertise. To assess the role that experienced evidence played on expertise ratings, the question was initially restricted to look at the impact of having *refuted* the belief put forward by the source, much in the same way as trust ratings were assessed. To do this, the two posterior measures of binary preferences and probability estimates were used. Interestingly, both the former of these – the ANOVA using binary preference of either accepting or refuting the belief as the independent variable for expertise ratings – and the latter probability estimate correlation with expertise ratings, showed no effect of belief assessment outcomes on ratings of expertise (*p* = .395, and *p* = .679, respectively).

Pre-test Comparison. By comparing Experiment 7's posterior ratings of expertise to the pre-test ratings of expertise, it was possible to determine the impact of the various conditions (and first-hand evidence) on the expertise of a source that has communicated an erroneous belief.

All expertise ratings, as a consequence of communicating an erroneous belief, decreased having had the opportunity to evaluate the belief against first-hand evidence. However, such a decrease is localised to those in expert conditions (pre-evidence, M = 68.44, SD = 27.785; post-evidence, M = 57.37, SD = 27.27), $F(1,290) = 10.008$, $p = .002$, $\eta^2 = .033$, whilst sources already considered low in expertise do not see a significant decrease (pre-evidence, M = 23.51, SD = 29.562; post-evidence, M = 20.75, SD = 26.867, $p = .419$). Further, the significant decrease in the expert condition from pre to post-test appears to be primarily driven by high trust conditions: When trust is high there is a significant drop in the rating of expert's expertise as a consequence of first-hand experience (pre-evidence, M = 73.23, SD = 26.645; post-evidence, M = 54.41, SD = 27.312), $F(1,145) = 14.803$, $p < .001$, $\eta^2 = .093$. Whereas when trust is low, the level of expertise assigned to experts remains constant from pre- to post-test, $F(1,144) = .444$, $p = .506$, $\eta^2 = .003$. Possible ramifications of high trust experts having more to lose are discussed below, and related to associated findings in loss of warmth and competence in social psychology findings (Cuddy et al., 2011; Fiske et al., 2007; Kenworthy & Tausch, 2008; Tausch et al., 2007).

5.3.3 Discussion

The purpose of Experiment 7 was to explore the role that source credibility factors play in the uptake and maintenance of an erroneous belief, when exposed to prolonged first hand evidence. Turning first to the choice data, there are three key effects of interest pertaining to current predictions.

Firstly, given a belief from a high trust source, participants are willing to overlook undermining initial evidence. This resulted in continued sub-optimal choices in favour of the belief and overruling the previously found ‘gatekeeping’ effect of initial evidence found with anonymous sources in previous research (Chapters 3 & 4). This effect is in line with the second prediction; when credibility cues indicate a source as being reliable, confidence in the belief being true is already in place prior to initial evidence exposure, supplanting it’s consolidating (or refuting) role. Secondly, when presented with a belief from an untrustworthy source, participants act as if the opposite of the belief is true (i.e. if told that “A is better than B”, they act as if “B is better than A”, given the potential ulterior motive of the source). This demonstrates that beliefs are processed in light of the credibility cues, and participants act accordingly.

Furthermore, when such suspicions (“the source is lying to me”) are confirmed by initial evidence (i.e. the belief, which is believed to be a lie, is undermined by initial evidence, thus confirming the suspicion), we find a mirror image of the biasing effect in the other congruent condition (high trust and supporting initial evidence). Interestingly, when such an assumption does not receive initially supporting evidence, we find the inverse of the gatekeeping effect found in previous work (Chapter 4). In other words, when dealing with low trust sources, there is once again a reliance on initial evidence, in this case to confirm the anterior of the belief (i.e. the suspicion that the source is lying), in line with predictions regarding the gatekeeping effect of initial evidence.

In this way, not only do these effects fit with work on source factors in attitude persuasion, which has indicated low trustworthiness promotes scepticism (Priester & Petty, 1995), but we have further extended this to demonstrate the evidence dependent consequences of this scepticism. If initial evidence instead confirms the belief (and thus undermines the *suspicion*), then we once again find the refutation / gatekeeper effect (i.e. no consolidation), which leads to no impact of belief (in either direction).

Conversely, source expertise did not play a significant role influencing choice data (although it did trend close to significance). This is broadly in line with expectations of the Bayesian source credibility model (see 5.1.2), and findings in advice taking of trust's importance relative to expertise (Sniezek & Van Swol, 2001), as well as mirroring findings in social psychology that warmth (trust) is more important and immediately accessed than competence (Cuddy et al., 2011; Fiske et al., 2007; Kenworthy & Tausch, 2008; Tausch et al., 2007). Further, this suggests a reliance on trust for initial directionality, whilst the role of expertise is somewhat relegated by the dominance of first-hand evidence, in line with previous findings in egocentric discounting (Yaniv & Kleinberger, 2000).

Regarding posterior measures, as with previous research, those showing the greatest degrees of bias in choice data (for example, high trust, supporting initial evidence groups) also retain that bias in posterior judgements and show significantly higher confidence in those judgements. Critically, it should be noted that the subtler interactions between *belief* and *trust* (such as the coherence effect of low trust, undermining initial evidence groups), has dissipated somewhat in posterior measures. However, the central, main effects of *belief* and *initial evidence* persist across all measures, along with the interaction between *belief* and *trust* as found in choice data. Interestingly, as found in pre-testing (see Appendix B.3), there is a main effect of *expertise* on confidence in posterior judgements (those receiving beliefs from experts are more confident in their judgements), which is further supported by the significant positive correlation between expertise ratings (irrespective of condition) and confidence. Such a finding fits with work on the impact of source factors on confidence (Earle, Siegrist, & Gutscher, 2010; Twyman et al., 2008) and convincingness (Harris et al., 2015).

Finally, when comparing trust and expertise ratings of the belief's source in Experiment 7 to the ratings elicited in the pre-tests (taken prior to any evidence based evaluation of the belief could take place), there are several, tentative findings of interest. Primary among them is that the level of source trustworthiness (high or low), leads to polarized outcomes. Specifically, if a belief comes from a source that is believed to be highly trustworthy, then not only does this result in greater belief preservation (despite its incorrectness), but having believed the source (i.e. believe they have been told the truth), the source is rated as even more trustworthy. In principle, this finding corroborates work on trust in advice taking literatures, whereby an advisor believed to be trustworthy is favoured more (Schöbel et al., 2016; Twyman et al., 2008). However, the demonstration of trustworthy sources not only getting away with, but profiting from, the communication of a falsehood, is unique, and hereafter termed a false-prophet cycle.

Conversely, ratings of expertise go down after experiencing first-hand evidence; an effect that is driven by expert sources (novices are likely already close to floor). Such an effect is likely attributable to experienced ambiguity of the evidence at hand. In other words, a source that has experienced "1000 trials" is rated as more credible before one realises how ambiguous the evidence is.

The second effect of interest mirrors the effect described above, wherein untrustworthy sources that communicate an erroneous belief are subject to more scepticism (and thus lower levels of consolidation on the belief and subsequent maintenance, or outright refutation and belief in the anterior). This scepticism leads to lower levels of belief uptake (and bias), and as a result, their already low ratings of trust are rated *even lower* (i.e. they have been confirmed as untrustworthy given the correct assessment of the belief's falsity, and update the source appropriately).

This polarization of trust ratings based on starting point (high gets higher, low gets lower) is compounded by *initial evidence* through congruency effects (i.e. when the initial evidence matches the proposition: “this trustworthy source is correct”, “this untrustworthy source is lying”, consolidation occurs). This effect (albeit tentative given the between-experiment comparison), has interesting implications for the impact of initial assessments of source trustworthiness in environments where statements are generic and first-hand evidence is either scarce or ambiguous (i.e. susceptible to confirmation bias effects), such as politics, health, and marketing. For example, within placebo effect research, cues to authority have been identified as impactful on the degree of placebo uptake (Wager & Atlas, 2015). Combining such cues with initial evidence effects (e.g. by using an initially potent, rather than medically inert, pill), may not only lead to greater ratings of trust in the prescribing doctor, but higher subsequent levels of placebo uptake given this update. Furthermore, although coherence-based models of source credibility have proposed the potential impact of testimony on reliability updating (Bovens & Hartmann, 2003; Harris & Hahn, 2009), empirically, such updating has received little attention in cognitive psychology (but see Harris et al., 2015; Jarvstad & Hahn, 2011). Accordingly, such findings, despite their tentative nature, indicate a fruitful potential avenue of further research.

5.4 Experiment 8: Cassandra’s Curse

So far, we have investigated the roles source credibility and initial evidence play in the adoption and maintenance of a fallacious belief. In doing so, we have found evidence of cyclical effects in the propagation of fallacious beliefs from high trust sources. Specifically, high trust sources lead to greater levels of belief uptake and maintenance, leading agents to conclude that the belief is true, which in turn leads agents to update the source’s trust rating as even higher (and thus allowing the cycle to continue). The effect of high trust on belief uptake has been shown to even override the previously found gatekeeping role that initial evidence has played in consolidating or refuting such communicated beliefs when source cues are absent (Chapters 3 & 4). Conversely, low trust sources result in individuals picking the opposite choice than that indicated within the belief, demonstrating that beliefs are processed in light of source credibility cues. Further, when this supposition is confirmed by initial evidence (i.e. "I suspect this source's comment that 'A is optimal' is fallacious, therefore I suspect B is actually optimal." – and initial evidence supports B being optimal), belief-uptake effects then occur in confirmation of this suspicion.

Consequently, an interesting question provoked by these effects is how such factors (credibility and initial evidence) play a role in the adoption of valid (i.e. directionally truthful) beliefs? Such a design allows for the testing of questions including whether a low trust source benefits from telling the truth in an uncertain environment?

The character of Cassandra in Greek mythology is the daughter of the King of Troy. The god Apollo falls in love with her unmatched beauty, and in attempting to woo her, gives her the gift of prophecy. However, when Cassandra refuses Apollo's advances, he curses her so that nobody will ever believe her. In this way, Cassandra is doomed to always foresee events, but have her warnings ignored. We draw a parallel here to the methodological set up of a low trust source attempting to convey a truthful (and thus, beneficial) belief to others.

Given the methodological similarity to Experiment 7, we predict the same general pattern of results (and in line with predictions of 5.1.3): High trust sources result in higher proportions of belief congruent (and in this case, optimal) choices and judgements, with initial evidence being overruled if contradicting the high trust source. Conversely, low trust sources result in initial choices that reflect the anterior of the belief (which may well, when consolidated by initial evidence, result in the aforementioned “Cassandra’s curse” outcome), as initial evidence that contradicts the suspicion (“the source is lying and the opposite is true”) resulting in no effect of belief in either direction. Finally, tentative predictions are made regarding reliability updating: Source expertise will have a role limited to affecting confidence levels (in line with findings discussed in 5.3.3), and high trust sources will benefit the most from the belief integration process (i.e. receive the greatest increase in trust ratings), in line with previous findings (see 5.3.3 similarly for a discussion).

5.4.1 Method

Accordingly, Experiment 8 followed the methodology of Experiment 7, with a single exception:

Whereas in Experiment 7, the belief manipulation always indicated the sub-optimal medicine was optimal (i.e. the belief was fallacious), in Experiment 8 this was reversed so that the optimal medicine is correctly indicated by the belief (i.e. the belief is truthful). In accordance with this change, the *initial evidence* manipulation is reverse coded in all Experiment 8 analyses, as what has previously been undermining evidence of fallacious beliefs are now supportive of the truthful belief, and vice-versa. Aside from this change in design, all remaining features of both design and procedure are carried over from Experiment 7.

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 500 prior HITs. Participants completed the experiment under the assumption that the purpose of the investigation was to improve medical decision making. Participants were English speakers between the ages of 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

5.4.2 Results

Descriptives and Processing. Given the overall similarity between Experiment 7 and Experiment 8, projected sample size was kept the same, with a minor downwards adjustment of the projected forgetting rate based on the rate found in Experiment 7. This resulted in a total sample size of 500. It is important to note, that the coding of *initial evidence* in all subsequent analyses reflects the change from a belief indicating the sub-optimal option (Experiment 7), to a belief indicating the optimal option (Experiment 8). As such, supporting groups still support the belief, and undermining still undermine the belief, but do so by providing evidence for the optimal (in the former case) and sub-optimal (in the latter).

Table 5.4.1: Experiment 8: Participant Breakdown by Group, along with the number of participants passing the belief manipulation check.

Trust					*N*	Passed Manipulation Check
High	**Expertise**	High	**Initial Evidence**	Supporting	67	51
				Undermining	63	52
		Low	**Initial Evidence**	Supporting	60	51
				Undermining	63	49
Low	**Expertise**	High	**Initial Evidence**	Supporting	61	49
				Undermining	65	51
		Low	**Initial Evidence**	Supporting	62	50
				Undermining	61	49

Participants were randomized into one of the eight possible conditions (see the first column of Table 5.4.1). The average age was 36.09 years (*SD* = 11.396) and the sample was 59% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of the manipulation comment, they were removed from subsequent analysis, the remaining group numbers are shown in the second column of Table 5.4.1. The decision to remove those who failed to remember the comment manipulation was taken following the same protocol and reasoning as in

Experiment 7. The following analyses were conducted using the remaining 401 participants, with an average age of 35.65 years (*SD* = 11.423) and 58.5% female. The results below will not be presented exhaustively; for the sake of brevity, only those of central interest to the thesis questions are included.

Correlates. There were no unexpected correlations regarding counterbalancing factors such as disease or medicine outcome assignments. As has been found in previous experiments within this body of work, gender correlated with Need for Closure, r =.239, N = 402, p < .001, with females showing higher levels, but neither variable correlated with any of the dependent or independent variables. Importantly, significant correlations were found between the degree of bias (taken as difference between belief pair and non-belief pair) in choice data and the degree of bias in posteriors, r = .482, N = 402, p < .001, demonstrating participant-level corroboration between the impact of *belief* in trial-by-trial choices and end-of-sequence judgements.

Choice Data. To assess the impact of the *belief, initial evidence* and source credibility manipulations on choices, a mixed ANOVA was run using the total number of optimal choices for the belief disease and the total number of optimal choices for the control disease, with the difference between the two as the within-subjects factor (*belief*). The between-subject factors included in the analysis were *initial evidence, trust*, and *expertise*. As can be seen in Figure 5.4.1, there were significant main effects of *belief*, F(1,394) = 11.593, p = .001, η^2 = .029, *initial evidence*, F(1,394) = 67.208, p < .001, η^2 = .146, and *trust*, F(1,394) = 12.808, p < .001, η^2 = .031, whilst *expertise* showed no main effect (p = .210). Such a pattern corroborates the general findings of Experiment 7.

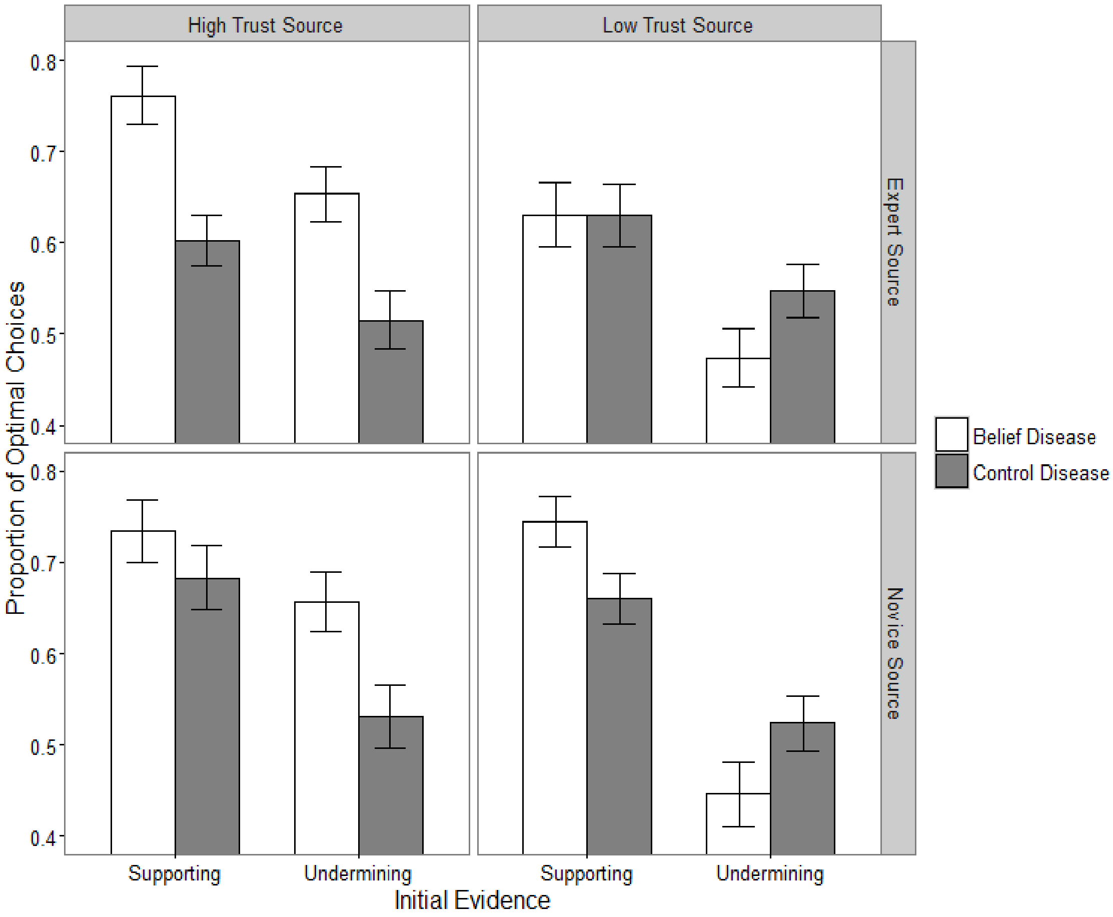


Figure 5.4.1. Experiment 8: Proportion of Optimal Choices. White bars present the disease that received a belief indicating the optimal option. Error bars reflect between-subject standard errors.

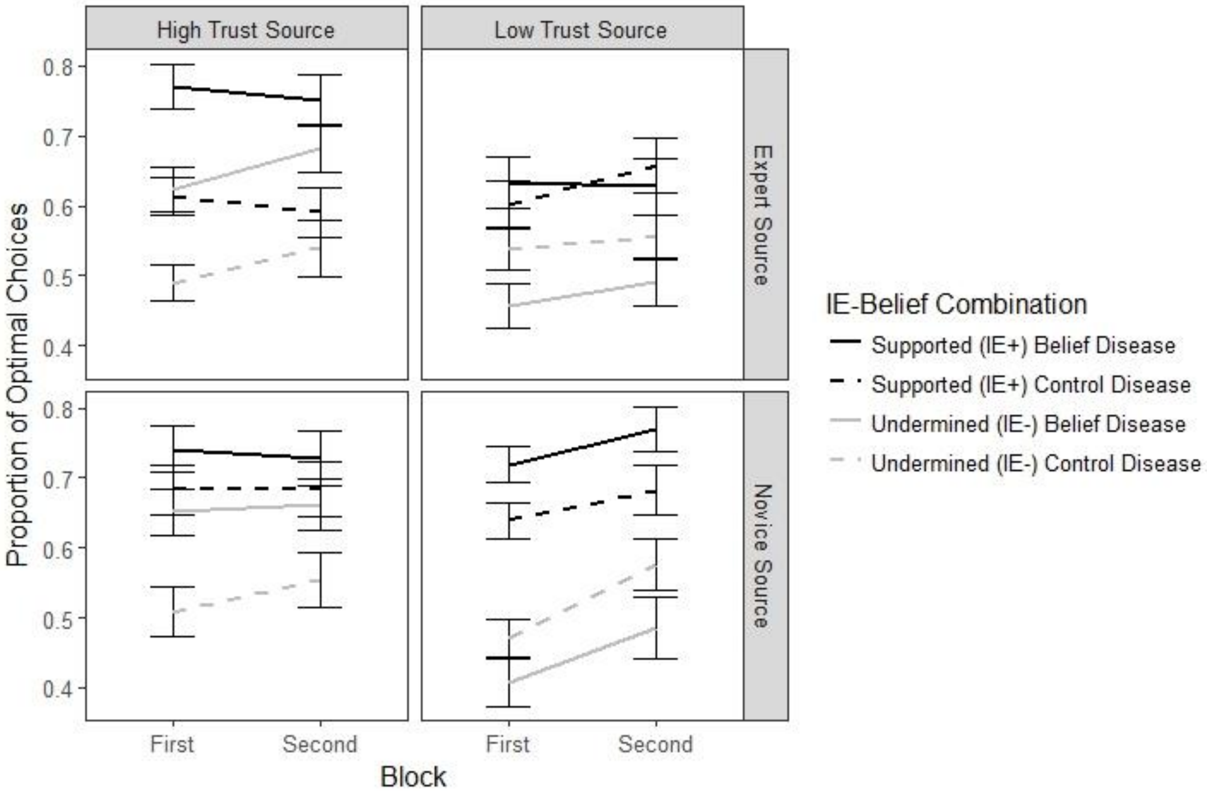
These main effects showed that significantly more choices were made for the optimal medicine when favoured by *initial evidence* (difference between white and grey bars in left-hand columns of Figure 5.4.1 above), having received a belief favouring that option, or receiving a belief from a high trust source. Furthermore, there was a significant interaction between *belief* and *trust*, $F(1,394) = 20.269$, $p < .001$, $\eta^2 = .049$, indicating that the conjunction of both a belief *and* a high trust source results in significantly more optimal choices. Visual inspection of Figure 5.4.1 showed a pattern of coherence / incoherence between *initial evidence* and *trust*, mirroring the effects found in Experiment 7. As such, the same protocol of further sub-group analyses are included below, and is further supported by the significant three-way interaction of *belief*, *trust*, and *initial evidence*, $F(1,394) = 5.825$, $p = .016$, $\eta^2 = .015$.

Splitting by Initial Evidence. By splitting participants into supporting and undermining initial evidence conditions and running a mixed ANOVA for *belief* (within-subject factor), *trust*, and *expertise*, it was possible to start determining the significance of coherence effects. As expected, those in supporting initial evidence conditions showed a significant main effect of *belief*, $F(1,197) = 13.366$, $p < .001$, $\eta^2 = .064$, supporting effects found in previous experiments, along with a novel main effect of *expertise*, $F(1,197) = 4.101$, $p = .044$, $\eta^2 = .02$. However, although *trust* did not have a significant main effect on choices, or a significant interaction with *belief* ($p = .121$) it did interact significantly with *belief* and *expertise*, $F(1,197) = 5.603$, $p = .019$, $\eta^2 = .028$. This latter effect suggests that the coherence pattern was dependent on expertise level within this subgroup.

Conversely, those in undermining initial evidence conditions do not show a main effect of *belief*, but do show a main effect of *trust* once again, $F(1,197) = 16.068$, $p < .001$, $\eta^2 = .075$, corroborating the pattern found in Experiment 7. The reason for the lack of a main effect of *belief* is again likely due to the strong *Belief* x *Trust* interaction, $F(1,197) = 21.771$, $p < .001$, $\eta^2 = .1$, in this case showing that the coherence of low trust and undermining initial evidence results in biased choices in the *opposite* direction to those in the incoherent, high trust and undermining initial evidence.

As in Experiment 7, there are two effects of consequence here; the first is that whilst low trust and supporting initial evidence together led to no impact of belief, the reverse (high trust, but undermining initial evidence) still procured an effect despite the previously found refuting effect of undermining initial evidence (Chapter 4). The second is that source trustworthiness was once again predictive of choice directionality[28]. The *Belief* x *Trust* interaction in undermining initial evidence groups demonstrates once again that not only is the belief processed *in the context of source cues*, but that the resultant hypothesis (“I do not trust the person telling me ‘A is better’, so I believe B is in fact better”) results in the same “suspicion-uptake” effect as when initial evidence supported the suspicion of the belief being false in Experiment 7, just in the opposite direction. Such an effect is expected, given the difference between the two experiments regarding the “correctness” of the belief, and given that participants are making use of the same cognitive processes.

Block Data: Assessing Learning. To assess the impact of the within (*belief*) and between-subject (*initial evidence*, *trust*, and *expertise*) factors on learning, and to determine whether the impact of such factors diminishes with experience in the task, an additional analysis was conducted using 25 trial blocks for each disease.



[28] This claim is supported by a Chi squared analysis on first choice data, finding a highly significant effect of *trust*, χ^2 (1, N = 402) = 66.586, *p* < .001, with those in low trust groups more often choosing the *non-specified* option, rather than the option specified by the belief, whilst the reverse is true of high trust groups.

Figure 5.4.2. Experiment 8: Proportion of Optimal Choices, split by 25 trial block. Line colour reflects initial evidence condition (black = *supporting* initial evidence conditions, grey = *undermining* initial evidence conditions), line type reflects within-subject disease (solid = *belief* disease, *dashed* = control disease). Error bars reflect Between-subject Standard Errors.

Accordingly, the analysis protocol followed that of Experiment 7, revealed a main effect of *block*, $F(1,394) = 19.952$, $p < .001$, $\eta^2 = .048$, wherein there was an overall learning effect towards the optimal option within a pair, corroborating Experiment 7. *Block* also interacted with *initial evidence*, $F(1,394) = 8.066$, $p = .005$, $\eta^2 = .02$, as *initial evidence* dictates a different starting point in block 1 (black lines lower than grey lines in First blocks of Figure 5.4.2), and therefore decreased learning in consolidated individuals (i.e. those receiving initial evidence indicating the optimal option do not have a deviation that experience will correct). The significant interaction between *block* and *trust*, $F(1,394) = 6.001$, $p = .015$, $\eta^2 = .015$, is also indicative of the starting point differences dictated by trust directionality, along with the reduced learning in consolidated individuals. This latter finding is distinct from the pattern found in Experiment 7, and is attributed to the higher levels of belief uptake given the validity of the belief results in participants receiving beliefs from high trust sources already being at ceiling, whilst the misled participants in low trust groups show a learning effect, given their initial deviation.

Posteriors. The same analysis protocol was used to assess posterior judgements as in Experiment 7. Given the extended analyses warranted for this experiment, the posteriors have been further split into sections by dependent variable.

Probability Estimates. The mixed ANOVA protocol was used to assess the effect of the independent variables on posterior probability estimates in the belief and control diseases (as the within-subject, *belief*, factor; white versus grey bars in Figure 5.4.3). There were significant main effects for *belief*, $F(1,394) = 5.184$, $p = .023$, $\eta^2 =$

.013, *initial evidence*, $F(1,394) = 22.893$, $p < .001$, $\eta^2 = .055$, and *trust*, $F(1,394) = 4.924$, $p = .027$, $\eta^2 = .012$, however, there was no main effect of *expertise*. Further, the interaction between *belief* and *trust*, unlike Experiment 7, did not reach significance ($p = .163$).

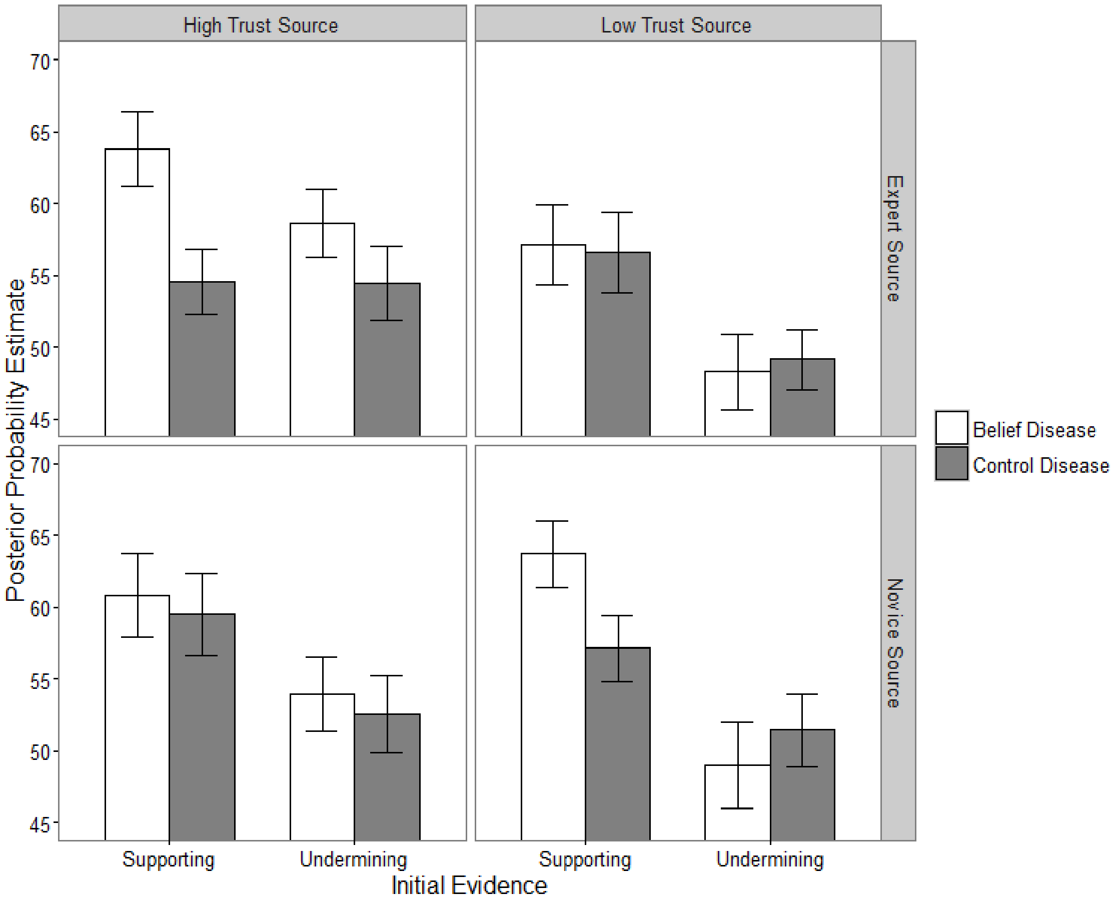


Figure 5.4.3. Experiment 8: Posterior Probability Estimates (estimate of percentage of optimal outcomes in favour of dominant medicine), split by group. Error bars reflect between-subject standard errors.

In general, the effects found in probability estimates are somewhat diminished by the impact of *initial evidence*. Interestingly, when splitting the analysis into subgroups of supporting or undermining initial evidence (in the same manner as choice data), there is a main effect of *belief* in the former, $F(1,197) = 9.17$, $p = .003$, $\eta^2 = .044$, along with a three-way interaction of *belief*, *trust*, and *expertise* (similar to choice data), $F(1,197) = 5.614$, $p = .019$, $\eta^2 = .028$. Meanwhile, the undermining initial evidence group revealed a dominating effect of *trust*, $F(1,197) = 7.648$, $p = .006$, $\eta^2 = .037$. In summary, the posterior probability estimates show an impact of *belief* in increasing optimal probability estimates, but such an effect is driven by coherence with supporting initial evidence. Conversely, when initial evidence undermines the belief, *trust* plays a dominating role, in that high trust sources continue to show a strong influence on probability estimates, despite undermining initial evidence. Conversely, the effect of low trust sources is nullified by the suspicion-refuting initial evidence. Taken together, these effects are in line with predictions regarding the dominance of source cues which indicate reliability, and the remaining gatekeeping effect of initial evidence when such cues do not.

Binary Preferences and Confidence. In line with the analysis protocol of Experiment 7, a mixed-effects logistic regression was run in R to test for each main effect, and if appropriate, possible interactions between the between-subject (fixed) and within-subject factors on binary preferences.

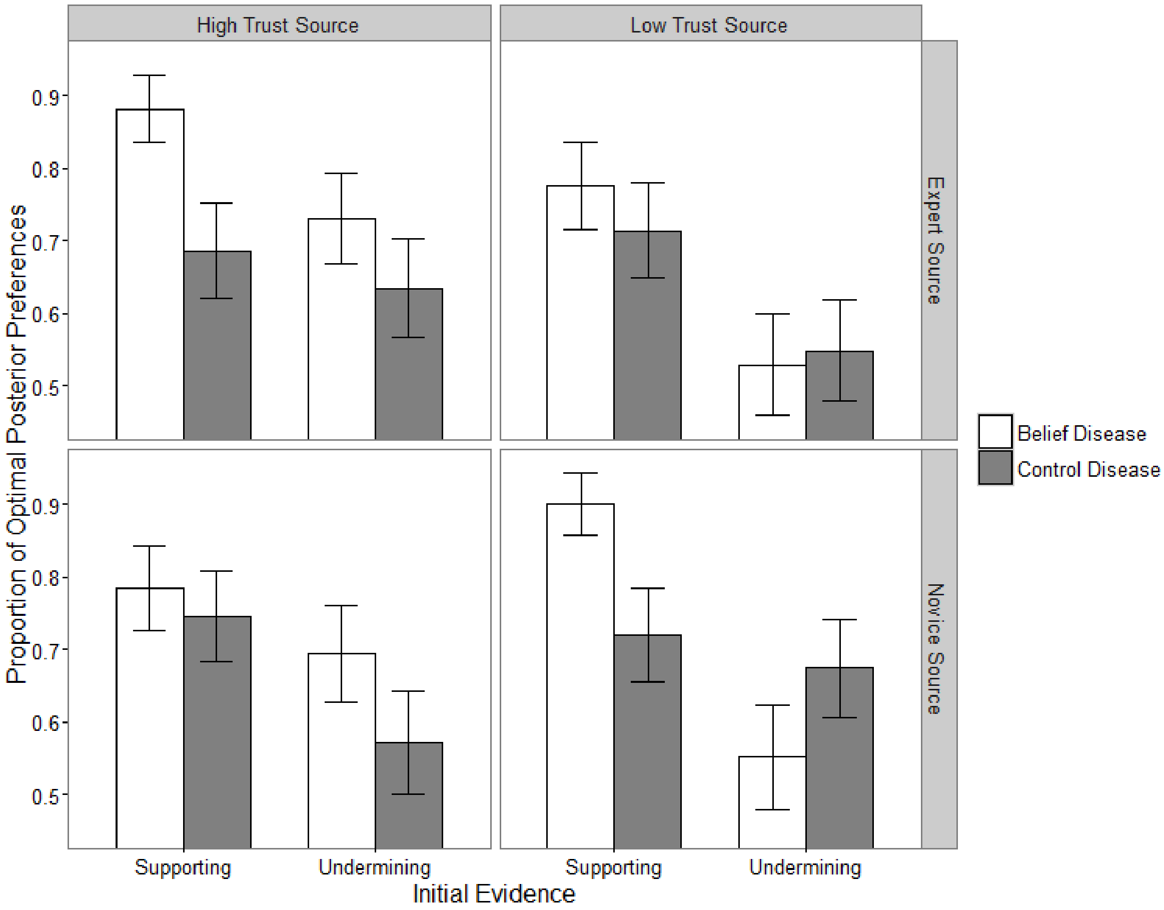


Figure 5.4.4. Experiment 8: Binary Preferences (proportion of participants indicating a preference for the optimal option, split by group). White bars represent the disease that received a belief indicating the sub-optimal option. Error bars reflect Between-subject Standard Errors.

Such an analysis found a significant improvement over the basic model (fixed intercept + subject random-effect) through the inclusion of *initial evidence* as a factor (left-hand pairs of bars within each facet are higher than right-hand in Figure 5.4.4), χ^2 (1) = 24.182, p < .001, and *belief* as a factor (belief (white) bars higher than control (grey) bars, Figure 5.4.4 above), χ^2 (1) = 4.6976, p = .03. However, no main effects were found for *trust*, χ^2 (1) = 1.5671, p = .2106, or *expertise*, χ^2 (1) = .3328, p = .564. These latter effects corroborate the (lack of) findings in Experiment 7, whilst the significance of *belief* is novel, but unsurprising given the validity of the belief communication. Consequently, given the significant main effects of *belief* and *initial evidence*, a model including the interaction of these two factors was compared to a model containing just the main effects. This model comparison yielded an improvement through the inclusion of an interaction term, however it did not quite reach significance, χ^2 (1) = 3.7313, p = .053. Such a model comparison was not possible in Experiment 7, but is again attributable to the novel validity (rather than falsehood) of the belief manipulation in Experiment 8.

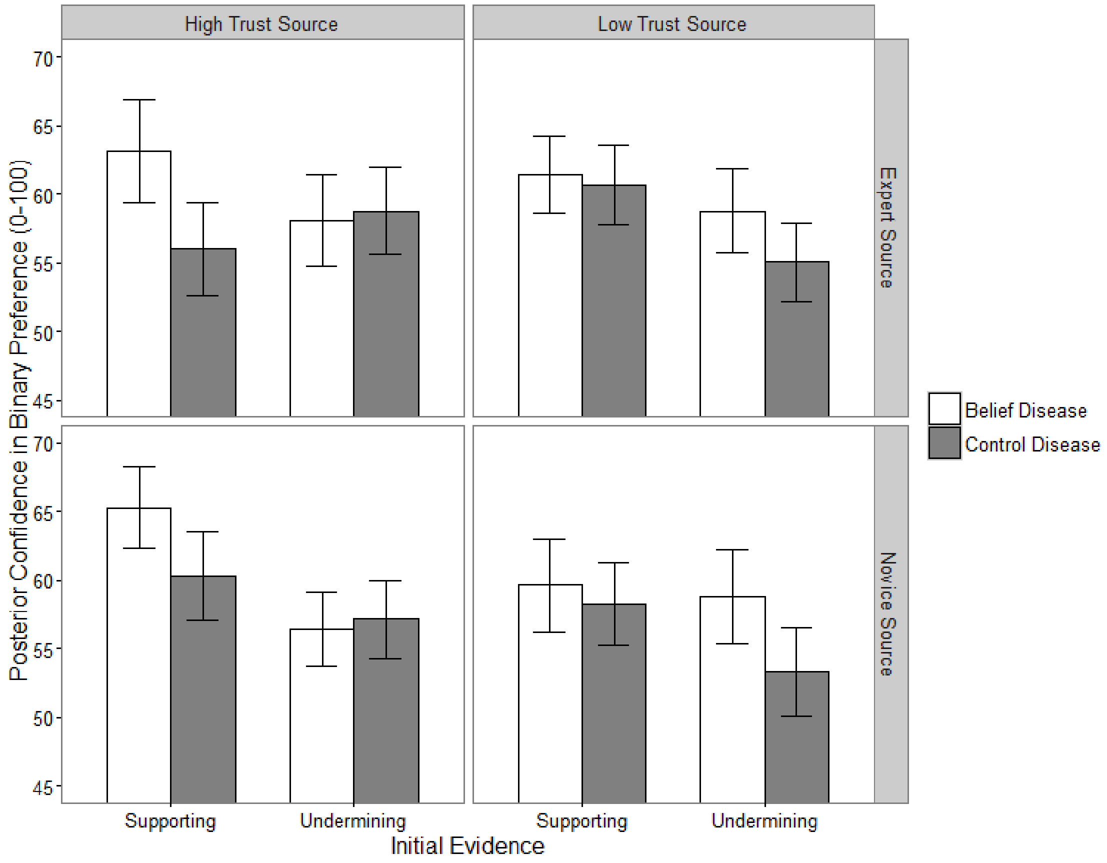


Figure 5.4.5. Experiment 8: Confidence in Binary Preference (0-100%), split by group. Error bars reflect between-subject standard errors.

Regarding confidence in binary preferences (and returning to the mixed analysis of variance format), the belief disease posterior confidence and control disease posterior confidence were used as the within-subject (*belief*) 2-level factor, along with *initial evidence*, *trust*, and *expertise* as between-subject factors. Although the only main effect was of *belief* (belief (white) bars higher than control (grey) bars, Figure 5.4.5 above), F(1,394) = 9.273, p = .002, η^2 = .023, there was, as in Experiment 7, a three-way interaction between *belief*, *trust*, and *initial evidence*, $F(1,393) = 8.058$, $p = .005$, $\eta^2 = .02$. This once again suggested the selective impact of initial evidence when trust is low. Investigating this interaction further, in the same manner as the analysis of the overall choice data, a secondary mixed ANOVA was conducted on the sub-groups of *initial evidence*.

Splitting by Initial Evidence. As expected, those in the supporting initial evidence conditions (left-hand columns of Figure 5.4.5) showed a significant main effect of *belief*, $F(1,197) = 7.473$, $p = .007$, $\eta^2 = .037$, supporting effects found in overall choices. However, *trust* did not have a significant main effect on confidence, and the interaction between *belief* and *trust* did not quite reach significance ($p = .059$). Conversely, those in undermining initial evidence conditions (right-hand columns of Figure 5.4.4) did not overall show a main effect of *belief* or *trust*, but did show an interaction *between belief* and *trust* (difference between right-hand column white and grey bars in left-hand versus right-hand facets, Figure 5.4.5), $F(1,197) = 4.483$, $p = .035$, $\eta^2 = .022$, in this case showing the coherence of low trust and undermining initial evidence (i.e. supporting suspicions provoked by low trust; Priester & Petty, 1995) results in greater confidence that the *opposite* of the belief is true, replicating the effects found in Experiment 7, despite the confidence being misplaced in this instance.

Ratings of Trust. An ANOVA was conducted to assess the effect of the *trust* manipulation on posterior ratings of trust. This analysis demonstrated that irrespective of group and experienced evidence, those in high trust groups rated the source as significantly more trustworthy ($M = 68.15$, $SD = 27.356$) than those in low trust groups ($M = 40.35$, $SD = 27.336$), $F(1,401) = 103.896$, $p < .001$, $\eta^2 = .206$. Following the protocol of Experiment 7 (see 5.3.2), analyses were conducted on the impact of evidence on ratings of trust, and comparisons to pre-test data. However, akin to the analysis of trust ratings in Experiment 7, the present analysis should be considered tentative.

Evidence on Trust. To assess the role that experienced evidence played on trust ratings, the question was initially restricted to look at the impact of having *refuted* the belief put forward by the source. To do this, two posterior measures were used; the binary preferences and probability estimates. Using the former of these, those who had refuted the belief (assessed using the posterior binary preference for the belief disease – with refutation as selection of the non-belief medicine option) rated the trust of the source significantly lower (M = 41.17, SD = 30.582) than those who adhered to the belief (M = 59.24, SD = 29.265), $F(1,401) = 29.417$, $p < .001$, $\eta^2 = .069$. However, unlike in Experiment 7, in which the communicated belief was fallacious, in this case, the reduction in trust was not driven by the high expertise conditions.

Using the posterior probability estimates, a similar overall effect is found, whereby trust ratings are correlated with posterior probability distribution judgements, r = .295, N = 402, $p < .001$, indicating those who are more inclined towards the belief indicated medicine also rate the trust in the source as higher, in line with both the findings of Experiment 7, and expectations informed by the Bayesian source credibility model (see 5.1.2).

Pre-test Comparison. By comparing Experiment 8's posterior ratings of trust to the pre-test ratings of trust, it was possible to determine the impact of the various conditions (and first-hand evidence in general) on the trustworthiness of a source that has communicated a valid belief. It is worth briefly reiterating that in Experiment 7, high trust sources received higher trust ratings having communicated a falsehood, whilst low trust sources were appropriately penalized for the error. Consequently, it is expected that high trust sources should once again benefit (especially given the validity of the belief) from communication. However, it is of interest whether low trust sources are able to benefit from the validity of their communication.

In theory, as all groups are receiving valid beliefs, trust ratings should increase given the opportunity to evaluate the belief first-hand. Although there is an overall significant increase in trust ratings across groups, $F(1,589) = 12.843$, $p < .001$, $\eta^2 = .021$, such an effect is driven by increases for high trust sources (pre-evidence, $M = 54.35$, $SD = 33.169$; post-evidence, $M = 68.15$, $SD = 27.356$), $F(1,290) = 13.686$, $p < .001$, $\eta^2 = .045$, whilst low trust sources see no significant improvement in trust ratings (pre-evidence, $M = 35.94$, $SD = 28.113$; post-evidence, $M = 40.35$, $SD = 27.336$, $p = .194$). Furthermore, by splitting this analysis into high and low initial evidence groups, such effects are shown to occur irrespective of *initial evidence*, in line with both the findings of Experiment 7, and the predicted pattern of credibility cues subsuming the role of *initial evidence* (see 5.1.3). To explain further, the increase in trust ratings as a consequence of high trust sources occurs both in supportive initial evidence conditions, $F(1,189) = 13.246$, $p < .001$, $\eta^2 = .066$, *and* in undermining initial evidence conditions, $F(1,188) = 6.991$, $p = .009$, $\eta^2 = .036$. Conversely, those in low trust conditions show no significant increases in trust in *either case*. Although tentative, such a finding suggests a difficulty in repairing trustworthiness, in line with loss of reputation findings in advice taking (Yaniv & Kleinberger, 2000).

Ratings of Expertise. An ANOVA was conducted to assess the effect of the *expertise* manipulation on posterior ratings of expertise. This analysis demonstrated that irrespective of group and experienced evidence, those in high expertise groups rated the source as significantly higher in expertise ($M = 63.6$, $SD = 25.523$) than those in low expertise groups ($M = 24.43$, $SD = 27.357$), $F(1,401) = 220.422$, $p < .001$, $\eta^2 = .355$. In line with the analysis of trust ratings, and following the protocol of Experiment 7 (see 5.3.2), analyses were conducted on the impact of evidence on ratings of expertise, and compared to pre-test data. However, as mentioned in previous analyses, the present analysis should be considered exploratory, and results taken as tentative as a consequence.

Evidence on Expertise. To assess the role that evidence played on expertise ratings, the question was initially restricted to look at the impact of having *refuted* the belief put forward by the source, much in the same way as trust ratings were assessed. To do this, the two posterior measures of binary preferences and probability estimates were used. Interestingly, the ANOVA using binary preference posterior of either accepting or refuting the belief as the independent variable found significantly higher levels of expertise ratings in those who accepted the belief (M = 46.8, SD = 33.202) as opposed to refuted it (M = 37.18, SD = 31.123), F(1,401) = 6.853, p = .009, η^2 = .08, which was further supported by the probability estimate correlation with expertise ratings, r = .129, N = 402, p = .009.

Pre-test Comparison. By comparing Experiment 7's posterior ratings of expertise to the pre-test ratings of expertise, it was possible to determine the impact of the various conditions (and first-hand evidence) on the expertise of a source that has communicated an erroneous belief. However, there were no significant changes to expertise ratings as a consequence of first-hand experience, either overall (p = .912), or when breaking down into high (p = .149) and low (p = .789) expertise sub-groups. A Bayesian T-test of this overall effect of first-hand experience on expertise ratings found a Bayes Factor of .099, indicating strong evidence for the null effect of experience ratings being affected by evidence when sources communicate valid beliefs.

5.4.3 Discussion

The purpose of Experiment 8 was to replicate the general pattern of results found in Experiment 7, but instead of using an erroneous belief, using a valid belief.

Specifically, in both experiments, participants were given a belief that one of the two possible medicines, to treat one of the two diseases, was better. In Experiment 7, the indicated medicine was in fact sub-optimal (i.e. worked 40% of the time, whilst the alternative worked 60% of the time), whilst in Experiment 8, the indicated medicine was in fact optimal (60% effective). Such a change allowed for the assessment of source credibility factors when they (potentially) hamper the uptake and maintenance of a valid belief, whilst Experiment 7 demonstrated source credibility factors, in line with predictions informed by the Bayesian source credibility model (see 5.1.2 and 5.1.3) could assist in the uptake and maintenance of a falsehood, despite sustained, first-hand evidence exposure.

Consequently, as in Experiment 7, choice data revealed that not only did high trust lead to significantly more belief-medicine choices (which were in this case optimal, rather than suboptimal), but such an effect similarly overruled the "gatekeeping" effect of undermining initial evidence found in previous research (Chapters 3 & 4). This finding supports the predicted impact of reliable sources, and replicates the first principal effect of Experiment 7, but instead of belief adherence resulting in sub-optimal choices, it instead resulted in optimal choices.

The second principal finding of Experiment 7 was also replicated in Experiment 8, whereby beliefs from low trust sources resulted in choices for the opposite medicine than that indicated by the belief. Such a finding once again supports an account of belief processing in light of the source credibility cues with which it is communicated. Additionally, despite low trust sources communicating a valid belief in Experiment 8, the participant's suspicion ("The source is likely lying, so I shall choose the opposite"), if confirmed by initial evidence, resulted in significantly more choices made in confirmation of the suspicion, corroborating the predicted influence of initial evidence when source reliability is low. Thus, the mirror effect found in Experiment 7 is also found in Experiment 8, only this time – given that the belief is in fact valid in this case – the confirmation of participant's assumption resulted in significantly more sub-optimal choices[29]. Furthermore, as in Experiment 7, if the suspicion is instead *undermined* by initial evidence, then once again, there is no significant impact of belief. This effect corroborates the prediction that in cases of uncertainty regarding the belief, whether using unknown sources in previous research (Chapters 3 & 4), or suspicions that the source is lying (as in the present work), then initial evidence plays a gatekeeping role.

Moving to posterior measures, much like Experiment 7, main effects of *belief* and *initial evidence* across all measures, although interactions between *belief*, *trust*, and *initial evidence* do not reach significance. This latter finding is attributed to the "washing out" of these manipulation factors (which all occur at the start of the experiment) by the time posterior measures are taken (and the fact beliefs are congruent with the optimal medicines, which people move towards over the course of the task). However, groups who have received beliefs from high trust sources and had such beliefs confirmed by initial evidence, once again show significant retention of belief biasing effects across posterior measures. Additionally, groups in which there is congruency between *belief* (whether the belief itself in high trust conditions, or the anterior assumption in low trust conditions) and *initial evidence* show significantly greater confidence in their posterior binary preferences.

As in Experiment 7, the role of expertise is once again restricted to posterior confidence, wherein groups that receive a belief from an "expert" are significantly more confident in their posterior judgements. This finding is supported by the significant positive correlation between expertise rating and degree of posterior confidence. However, once again expertise played no significant role in all other choice and

[29] An analysis of the low trust, and undermining initial evidence group of participants (N=100) reveals a significant effect of *belief*, *F*(1,99) = 5.964, *p* = .016, with belief disease choices significantly more sub-optimal than control disease choices.

judgement data. A finding that is attributed to the dominating effects of prolonged first-hand evidence exposure, and the (expected) mitigated impact of *expertise* (see 5.1.2).

To put these findings within a credibility context, trust implies the motive of the source, and thus the directionality for first choices. As such, although normatively, low trust sources should simply be ignored, trust dictates whether subsequent evidence experienced is either confirmatory or contradictory. For example, a low trust (i.e. distrusted) source's belief is suspected of being the opposite ("The source said A is better, but is likely lying, so I suspect B is in fact better."), and initial evidence is then perceived as either confirming or contradicting this suspicion. In either case (trusted or distrusted), the perceived access the source has to information (i.e. expertise), is secondary to the combination of trust and repeated, experienced evidence. However, given the source expertise's influence on posterior confidence, expertise may assist in tightening the confidence interval around the final choice and probability judgements – assuming trust and/or evidence factors have already validated the belief.

Regarding how participants rate the trust and expertise of the belief source, having experienced first-hand evidence, the effects found bear a close resemblance to those found in Experiment 7 (given that in Experiment 8, all sources have provided a valid belief). When comparing trust ratings to pre-test values (i.e. ratings of the source and belief prior to any evidence-exposure), there is once again a disparity between high and low trust sources. High trust sources receive significantly higher ratings of trust after evidence exposure; an effect that is even greater when the belief has also received initially supporting evidence. Conversely, low trust sources (although trending positively given their communication of a valid belief) receive no such significant increase in their trust ratings, despite having the most to gain from telling the truth. This difficulty in "repairing" estimations of trust has a parallel in the impression formation literature, in which early, negative impressions of an individual have been demonstrated as hard to overcome (Anderson, 1965; Mann & Ferguson, 2015) and in advice taking, the ease with which advisors may lose reputation, relative to regaining it (Yaniv & Kleinberger, 2000). This raises interesting questions regarding possible asymmetries in reliability updating, and further work is recommended in assessing the situational or evidence based limitations of this (e.g. how much supporting evidence would be required to "forgive" an untrustworthy source?).

Although Experiment 7 demonstrated a reduction in expertise ratings among expert sources (attributed to a realization of the ambiguity of experienced evidence, and the generally recognised falsehood of the belief), Experiment 8 found no changes to expertise ratings. Such a finding is likely situation dependent, but likewise raises questions regarding reliability updating. For example, much like the implication of the low trust source (lack of) updating, further work is suggested to investigate possible asymmetries in the degree of updating in light of evidence. In particular, are the present sub-optimal degrees of reliability updating artefacts of the present paradigm, in which beliefs are generally accepted as valid, but the evidence with which to update is somewhat ambiguous? Or is this finding more systematic across other domains of belief updating? Further, are such deviations from expectancy a function of situational importance (i.e. when decisions matter more, is there increased engagement in updating processes?)?

5.5 General Discussion

The capacity to communicate beliefs regarding action-outcomes, or the environment in general, yields a substantial advantage to humans as a species. Through second-hand transmission of knowledge, it is possible to learn without the need for direct experience. However, not all communications are created equal; some, whether through misunderstanding or ill-intent, convey misleading information, which often incurs a cost to either the recipient themselves, or society as a whole (Gilovich, 1993; Vyse, 2013). Further, environments in which such communications are assessed may differ in the (perceived) degree of noise, and thus diagnosticity of experienced evidence. Fortunately, humans as reasoners can also incorporate cues to a belief's validity (Briñol & Petty, 2009; Hahn et al., 2009), including source trustworthiness (Siegrist et al., 2005; Twyman et al., 2008) and expertise (Goodwin, 2011; Sniezek & Van Swol, 2001), under the category of source credibility (Chaiken & Maheswaran, 1994; Hahn et al., 2009, 2012; Harris et al., 2015; explored in social psychology as warmth and competence, respectively; Cuddy et al., 2011; Fiske et al., 2007). Trustworthiness has been defined as a form of motivational likelihood (i.e. how likely is the source motivated to help or deceive me?), whilst expertise can be thought of as competence (i.e. how likely is the source to possess accurate knowledge?). The present work has sought to implement these two factors into an existing paradigm in which a communicated belief is assessed against sustained, probabilistic first-hand evidence (Chapters 3 & 4). This previous research has found a pivotal role of initial evidence in either consolidating or refuting the communicated belief, given that it is communicated by an unknown source.

In this way, initial evidence acts to not only update the initial assessment of the belief being true, but by proxy updates the likelihood of the source being credible (see 3.6 and 4.5 for further explanation), from which one can draw a parallel to the importance of *confidence* in the source in advice efficacy (Earle et al., 2010; Harvey & Fischer, 1997; Siegrist et al., 2005; Twyman et al., 2008). Given this "consolidation" of a belief, subsequent evidence is then integrated in a biased manner, whereby evidence that should refute the belief is underweighted relative to confirmatory evidence (i.e. a confirmation bias). Thus, the belief is maintained, despite its inaccuracy (see Chapters 3 & 4).

The two experiments in the current work used pre-tested trustworthiness (high or low) and expertise statements (high or low) that accompanied the communicated belief and were manipulated between-subjects. The belief recommended one of the two medicines for one of the two possible diseases with which each patient (trial) presented, and as such the difference between the "belief" disease and the control (did not receive any belief information) disease was the within-subjects effect of *belief*. Both diseases were identical in terms of evidence; each having a dominant medicine option (cured 60% of the time) and a sub-optimal option (cured 40% of the time), and 50 trials were experienced for each disease (alternating between the two). Beliefs were then either supported or undermined by *initial evidence,* a between subject manipulation that was yoked across diseases (i.e. if the belief-indicated option was the sub-optimal medicine for the belief disease, and the initial evidence condition was supporting, then *both* the sub-optimal medicine in the belief disease *and* the sub-optimal medicine in the control disease cured the disease on their first two trials, whilst both optimal medicines failed to cure), following previous use of the paradigm (Chapter 4). The sole difference between Experiments 7 and 8 was the validity of the belief communicated. Specifically, all beliefs in Experiment 7 recommended the sub-optimal medicine, falsely indicating it was the superior choice, whilst all beliefs in Experiment 8 correctly recommended the optimal medicine. Taken together, the two experiments demonstrate several important effects in the uptake and maintenance of beliefs from second-hand sources in which source credibility cues are present, regardless of whether such beliefs are supported by evidence.

The first of these findings is that not only do high trust sources result in the greatest degree of belief uptake and adherence, but that such uptake and adherence occurs irrespective of whether initial evidence supports or undermines the belief. This finding supports the prediction informed by the Bayesian source credibility model

(Hahn et al., 2009, 2012; Harris et al., 2015; see 5.1.3), in that when source cues indicate the source is reliable (and as a consequence, increase confidence in the belief being valid), this subsumes the role initial evidence plays in validating both belief and (by proxy) source. Further, this occurred irrespective of the validity of the belief itself (i.e. Experiment 7, which involved the communication of a falsehood, and Experiment 8, which involved the communication of a valid belief, both demonstrated the same impact of high trust sources), and taken with the overruling of initial evidence effects, draw a sharp distinction from predictions of the Elaboration Likelihood and the Heuristic-Systematic Models (Briñol & Petty, 2009; Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984). These models instead predict that the impact of source credibility cues should be subsumed by exposure to evidence, when there is the capacity and motive to evaluate it (as in the current paradigm).

This finding also distinguishes the current work from previous research into belief uptake and maintenance that has found initial evidence plays a gatekeeping role in validating beliefs from *unknown* sources (Chapters 3 & 4). The present data also fit with findings on the impact of confidence in the source on the efficacy of advice (Harvey & Fischer, 1997; Twyman et al., 2008), in cooperation (Earle et al., 2010), risk communication (Siegrist et al., 2005), and argumentation (Harris et al., 2015), along with the elicitation of positive emotions and behaviour as a consequence of perceived warmth in social psychology (Fiske et al., 2007). However, the current work demonstrates the impact of high trustworthiness on confirmation bias effects over prolonged, first-hand evidence integration, with belief adherence occurring whether the belief was valid *or not*. This finding has implications for the potency of perceived trustworthiness over “truth” in applied settings including politics and marketing.

The second effect of interest is the interplay between low trustworthiness and initial evidence. The first element of this is those receiving a belief from a low trust source choose as if the opposite is true (i.e. they make the assumption that because the belief comes from an untrustworthy / ulterior motivated individual, then the opposite of the belief is likely to be true, and choose accordingly). Secondly, this suspicion (the source is lying and the opposite from the belief-recommended medicine is actually better) results in significantly more choices in favour of it, providing it has been consolidated by initial evidence. This fits with the prediction informed by prior work on unknown sources (Chapters 3 & 4, but see 5.1.3), that uncertainty surrounding the source (and by proxy, validity of the belief) results in reliance on initial evidence to act as a validator (in this case, of suspicion, or the anterior of the belief). Accordingly, in either situation (initial evidence confirms the suspicion, or not), recipients fail to take advantage of the truthful belief (5.4.2). Such a finding is termed Cassandra's Curse, as the low trust status of the source prevents the recipient from believing the truth. An additional implication is that beliefs are processed in light of the source credibility context from which they are communicated – a finding that fits with the Bayesian source credibility model (Bovens & Hartmann, 2003; Hahn et al., 2009, 2012; Harris et al., 2015) in which belief content and source (reliability) should inform one another.

Further, the interplay *between* credibility cues, belief content interpretation, and evidence evaluation of said interpretation, raises challenges for Elaboration Likelihood and the Heuristic-Systematic Model (Briñol & Petty, 2009; Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984) explanations. The impact of source cues on belief content interpretation, which in turn is then conditionally dependent upon an order effect, does not readily fit a dual process account. Further, both the resulting high levels of belief adherence, irrespective of belief validity, and the accuracy incentives and prolonged evidence exposure (motive and opportunity), are not readily incorporated. It should however be noted, that proponents of such models could argue that source cues are playing a role (rather than being ignored as shallow cues) due to insufficient task engagement. Although we feel that the learning effects present in the task undermine such an argument, further research in the manipulation of motivation is suggested to provide a more robust test of model predictions. Lastly, such an interpretation of belief content in light of source cues does helps rule out explanations of the effect of belief communications as simply an automated response to the mentioning of one medical option, akin to anchoring effects (Epley & Gilovich, 2001).

The third finding of interest, although tentative given the across-study comparison, is in how credibility is updated in light of prolonged exposure to first-hand evidence that can verify the communicated belief (and thus the sources credibility along with it). Unsurprisingly, and in accordance with research in advice taking (Schöbel et al., 2016; Twyman et al., 2008), when a trustworthy source gave valid information (Experiment 8), sources received an increase in their rating of trust. However, such increases were limited to sources that were already deemed highly trustworthy, whilst those who had communicated the identical, valid belief, but were low in trustworthiness, received no such significant increase in their trust ratings (part of Cassandra’s Curse). This is likely attributable to the increased scepticism and lower levels of belief uptake of participants in low trust groups. Further, even when communicating an unsupported or erroneous belief (Experiment 7), high trust sources in fact *benefited* from such a communication, receiving higher ratings of trust after evidence evaluation. This effect is attributable to the biased integration process that resulted from the high trust source belief transmission, including the overlooking of undermining initial evidence. Taken together, these two effects suggest an evaluative asymmetry in the updating of source reliability, something that has been alluded to in social psychology in terms of asymmetries in the stability of positive versus negative oriented warmth traits (Tausch et al., 2007), but is an interesting and important topic for further research.

Additionally, although this is a novel demonstration of a source benefiting from dissemination of a falsehood *despite* the ability to verify the belief against first-hand evidence, given that recipients believe they have been told the truth, the consequent positive updating fits with the general findings in social psychology (Cuddy et al., 2011; Kenworthy & Tausch, 2008), although this has yet to be fleshed out formally in terms of process or cognitive models.

Conversely, low trust sources are penalised for the communication of false information, with significant reductions in trust ratings. This is attributed to the increased scepticism in the belief given its untrustworthy source, resulting in either biased assimilation in the opposite direction, or (in the case of the belief still receiving some initial support and refuting the recipients' suspicion) no impact of belief whatsoever. Taken together with the findings regarding high trust sources, such polarizing effects demonstrate the dominating effects of pre-decisional cues (Nurek et al., 2014) and in particular source trustworthiness over the validity of the belief being communicated. This has implications for impression formation and stereotyping (Anderson, 1965; Mann & Ferguson, 2015; Smith, 2014), marketing (Ha & Hoch, 1989; Hoch & Ha, 1986; Klayman & Ha, 1987), politics (van Erkel & Thijssen, 2016), and placebo research (Colagiuri, Livesey, & Harris, 2011; Wager & Atlas, 2015). In the case of placebo effects, emphasis to create the impression of credibility on the part of the medical practitioner prescribing the placebo should lead to greater levels of placebo effects in patients – something that has already been preliminarily identified under "social cues" (Wager & Atlas, 2015).

Interestingly, expertise was found to play a limited role in the belief uptake and maintenance process in the present paradigm. In line with previous research in advice taking, a relationship was found between the perceived expertise of the source, and confidence ratings (Sniezek & Van Swol, 2001), with expert sources provoking higher confidence in poster binary preferences. However, both choice data and posterior probability judgements were otherwise found to be impacted primarily by trust and initial evidence. One explanation for this finding is the dominating role of first-hand evidence in such an integrative task. In other words, because belief recipients experience a prolonged amount of first-hand evidence, reliance on the expertise of the source is then underweighted, in line with an ego-centric discounting account (Yaniv & Kleinberger, 2000). Unlike trust, which dictated choice (and subsequently, judgement) directionality, the impact of expertise is subsumed by first-hand evidence. Exploring whether such a (lack of) impact of expertise is situational (i.e. a consequent of the belief content, environment noise, and evidence diagnosticity) or systematic, is an interesting question for further research, with potential implications for the formulation of descriptive models of belief (and source reliability) updating.

Several limitations should be taken into account with the results discussed here. First of all, as previously indicated, some of the subtler principal findings did not extend into posterior measures. In particular, effects such as the mirrored bias when suspicions are confirmed by initial evidence, although trending in the correct direction, did not quite reach significance. Possible reasons for this include the degree of learning that has taken place before posterior measures are taken. Although this should be taken into account when considering the central findings are based primarily on choice data, participants have experienced a large amount of evidence by the time posteriors are taken (50 trials per disease). As such, when extrapolating to real world parallels, the central findings still bear some validity (given the unlikely capacity to wash-out Source Trustworthiness x Initial Evidence interactions with large quantities of evidence in real world settings). However, some of the main effects, such as initially supported beliefs from high trust sources resulting in the greatest degree of bias, did carry through into posterior measures (irrespective of belief validity). Not only this, confidence measures and general main effects of *trust* supported the effects found in choice data, as well as previous findings on the greater degrees of bias in posteriors leading to higher confidence (Chapters 3 & 4).

The second limitation pertains to the online nature of the tasks in the present work. MTurk has known shortcomings, such as the lack of experimental control (Goodman, Cryder, & Cheema, 2013). However, several studies have shown strong replications of effects found in laboratory experiments using the site (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). One advantage of using such a method is its greater ecological validity, as participants show better representation of demographics relative to the majority of psychological study populations (Henrich, Heine, & Norenzayan, 2010). The fact participants did not see the belief manipulation as experimentally driven added real world applicability, and distinguishes this work from more artificial, lab based assessments of integrative confirmation bias (Staudinger & Büchel, 2013). Further, the online format of a "comment section" has growing real world relevance, given such a context represents an ever-expanding and abundant source of communicated beliefs in the modern world.

Before moving to the final, broader applications of the present research, it is worth summarising the implications to source credibility specifically, given its central role within the present work. Firstly, the finding of high trust sources subsuming the role of initial evidence (and generally provoking higher levels of belief uptake and maintenance) cues fits with predictions inspired by the Bayesian source credibility model (Hahn et al., 2009, 2012; Harris et al., 2015), and contradicts the expectations of competing, dual process accounts (Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984). The latter instead predicting the subsuming of source cues, given motive and opportunity to evaluate evidence (although further research is suggested to explore the role of these two latter components). Secondly, the finding that those receiving beliefs from low trust sources require initial evidence to consolidate their suspicion indicates the role uncertainty plays in belief acquisition. This effect bears a parallel to work on primacy in the formulation of a hypothesis from first hand evidence (Allahverdyan & Galstyan, 2014; Anderson, 1965; Dennis & Ahn, 2001; Hogarth & Einhorn, 1992), and fits the prediction based on the findings of previous research using beliefs of uncertain (source) validity (Chapters 3 & 4). However, this effect, given its dependence on order of presentation, does not fit easily within a Bayesian world.

One possible way of reconciling the initial evidence findings with source credibility theory is through confidence intervals (or the width of the probability density function) surrounding the nodes for the probability of the belief, and associated nodes for source credibility elements. In particular, when confidence intervals are wide (i.e. there is low confidence for the belief and associated source nodes), in such cases as either absent (Chapters 3 & 4) or uncertain / suspicious (low trust sources in present work), initial evidence acts to narrow these intervals, assuming coherence between evidence and expectancy. Put another way, given the suspicion a belief is false (and thus the anterior of the belief is true), the initially low confidence in this assertion rapidly increases in confidence as early experiences independently confirm this suspicion (and as a consequence, the probability of the source being untrustworthy is similarly updated). Conversely, if source cues indicate a source is credible (and thus a belief is likely valid), then confidence is already high before initial evidence is experienced, subsuming initial evidences’ consolidating role. It should however be noted that such a mechanism implies asymmetry in the initial level of confidence in the suspected (in)validity of the belief when accompanied by trustworthy versus untrustworthy cues. Such a supposition is not unreasonable given the singular interpretation for a high trust source (the belief is likely true, to the best of the source’s abilities), as opposed to myriad interpretations of a low trust source (the belief is a lie, the source is double bluffing and in fact the belief is true, etc.). Such effects are also likely to depend on the perceived clarity provided by evidence, among other situational factors. As such, further work is needed in the exploration of order effects in belief (and associated reliability) updating, with the present work serving as an interesting challenge to behavioural applications of source credibility models.

Similarly, the asymmetries found in reliability updating (i.e. the failure of low trust sources to capitalise on valid communications in terms of updated trustworthiness, and the general conservatism in expertise updating relative to trust) pose interesting challenges to models of source credibility. In particular, such asymmetries (notably given the tentative nature of these findings) require further work to determine whether the issue is situation specific (i.e. a function of the specific instantiation of evidence diagnosticity, or expected environmental noise) or something more systematic. In determining potential mechanisms or boundary conditions of such effects, further work should address not only evidence clarity factors, but engagement (via motive manipulation) based explanations. One immediate suggestion is to elicit confidence judgements from participants regarding the perceived reliability (i.e. expertise and trust) over the consolidation process, to help determine possible differences in confidence intervals.

Several literatures and domains that share common elements with the present work are worth implicating in future avenues of research. One such area is placebo effect research, which has already identified the value of authority cues in increased placebo uptake (Wager & Atlas, 2015). However, as indicated by previous research on erroneous belief uptake and maintenance (Chapters 3 & 4), work on placebo effects has not yet integrated elements of first-hand evidence order effects, which could further facilitate placebo responses. Similarly, work on persuasion in politics (Calvert, 1985; Lewandowsky et al., 2012; Madsen, 2016) and consumer research (Ha & Hoch, 1989;

Hoch & Ha, 1986; Mandel, Petrova, & Cialdini, 2006; Metzger & Flanagin, 2013) both deal with attempts to “persuade” recipients of a particular message, whether to vote for a candidate or buy a product. In both cases, research that explores the conjunction of source credibility cues to belief validity, *and* first-hand evidence integration in light of said belief, is of critical importance. For example, the finding that sources believed to be high in trust can not only lead people to misinterpret evidence in a confirmatory manner (even overlooking the impact of initially undermining evidence), but that the recipients’ resultant genuine (but erroneous) belief that they have been told the truth can then benefit the source even further, is particularly worrisome.

CHAPTER 6: AGENT-BASED MODELLING: FALSEHOOD PROPAGATION ACROSS SOCIAL NETWORKS

Over the preceding chapters, empirical evidence has been presented that argues a process account of confirmation bias in the uptake and maintenance of erroneous beliefs, despite exposure to prolonged first-hand evidence. Throughout, the societal implications of fallacious belief adoption have been alluded to with regard to the internet age, and the potential impact of an increasingly interconnected world. Put another way, the ever-increasing interconnectedness of the online and telecommunications world can be thought of as a "double-edged sword" for knowledge transference: On the one hand, the global population has never had greater access to information than it does today. On the other, such access has arisen in tandem with a democratisation of the mediums of communication – any individual with access to the internet can create websites or post information to forums with an agenda uncoupled from truth or veracity. It is easiest to highlight the present communicative capacities by drawing a contrast with the pre-internet age.

Prior to the internet, to communicate to others was only available to the average citizen of the 20th Century (i.e. those without a television or radio show) along two dimensions, in a limited manner. If one wanted breadth of communication, reaching many people at once in an unfiltered manner (i.e. no editors, censors, or communication experts), then one required a soapbox and as many people as were within earshot. If one wanted depth (or rather, *distance* – whether temporal or physical) of communication, then one had to phone or write a letter to each recipient. With either of these communication avenues, reaching a large number of people would take a great deal of time and effort and/or skill. Conversely, to communicate on a truly large scale in terms of both depth and breadth, required mediums such as television and newspapers. For these, especially in terms of the dissemination of fact and opinion, there was required some degree of expertise, verification, or at least deliberation, otherwise known as gatekeeping (Dylko, Beam, Landreville, & Geidner, 2012). This is not to say that erroneous information was not accidentally or intentionally communicated, but that – as far as the human condition allows – there was a greater likelihood of validity. Importantly, the democratisation of public debate is not intended to be argued as a solely negative phenomenon; the ability to communicate and influence the discourse (rather than a handful of TV or newspaper producers / editors) carries with it merit in a modern democracy.

To draw this in contrast with the present age of online communication, anyone with a smartphone or computer can send emails, tweets, and other messages, post on blogs, social media, and forums readily and repeatedly. This unfiltered information can then reach (whether directly from the source, or passed on by others) an audience of millions. This newfound capacity poses a particular problem. An erroneous belief, whether through ill-intent or misinterpretation, can be sent and received, prior to any form of verification, by a mass audience on the other side of a screen. This leads to potential super-additive, aggregate behaviours. Put another way, the structures people generate as individuals (networks) result in non-linear system-wide behaviours.

Agent-Based Models (ABMs) are simulations of individual agents, which interact within a synthetic environment (Gilbert, 2008). Agents have behaviours encoded into them, including rules for interaction, and as the simulation runs, system-wide behaviours are generated (termed micro-motives and macro-behaviours, respectively; Schelling, 2006). For example, by encoding individuals with a probabilistic preference for culturally similar neighbours (micro-motive), and the capacity to move, Robert Axelrod (1997) was able to demonstrate how communities (from local neighbourhoods, to cities and nations) gradually become culturally polarized (macro-behaviours). In this way, the aggregate consequences of individual-based findings may be explored. Simple, pattern-oriented ABMs can therefore test possible societal “patterns”, given a set of cognitive assumptions and behaviours encoded at the individual level.

Importantly, the purpose of the present work is not to provide a complete statistical treatment or computational model of the learning / biasing mechanisms of the previous chapters. Rather, it is instead aimed to provide a descriptive exploration of the societal patterns (and their implications) of the belief uptake and maintenance findings taken from an individual level, and extrapolated to a network of interconnected individuals. In so doing, one can simulate the impact of an interconnected, heterogeneous, dynamic system of individuals on societal level behaviours. To do this, the present chapter uses a series of ABMs in which agents are placed within an interconnected network and ascribed rules for belief uptake and evidenced-based evaluation (based on the preceding empirical work), along with propagation behaviours. Where necessary, model architecture is supported by relevant literatures to validate the model architecture.

ABMs have been used in this fashion in impression formation (Smith, 2014; Smith & Collins, 2009), information cascades (Cui, 2016), intergroup dynamics of prejudice (Gray et al., 2014), and opinion dynamics (Dandekar, Goel, & Lee, 2013; Duggins, 2016; Hegselmann & Krause, 2002) within social psychology. It is the latter literature that is of most relevance to present model development. This work has typically focused on the role of communication among heterogeneous individuals over time, to develop illustrative models of the processes behind convergence and polarization in cultures (Axelrod, 1997), attitudes (Jager & Amblard, 2005), and political opinions (Duggins, 2016), as well as group (in a cultural sense) stability over time (Carley, 1991), consensus effects (DeGroot & DeGroot, 1974; Hegselmann &

Krause, 2002), and social differentiation via structure-based interaction (Mark, 1998). However, such simulations typically focus on belief development through *repeated* agent interaction (e.g., the polarization of subjective opinions across a culture). The role of the present research is instead to focus on the interaction of communicated beliefs and subsequent evidence-evaluation within an interactive, societal level system. Moreover, along with explicit manipulation of parameters within the learning process, an additional aim of the present work is to explore the interplay between increasing interconnectivity among individuals, evidence-based belief propagation decisions, and latterly, social psychology variables.

In summary, the purpose of this chapter is to take the principles of belief transmission and evidence evaluation processes explored in the preceding empirical chapters (taken from the aggregation of isolated individuals), and extrapolate to an interactive, societal level. In other words, the ABM process makes it possible to provide an exploratory account of how such belief uptake and maintenance processes interact with the structural elements of, and behaviours within, interactive networks. Outcomes of interest, following from the proposed implications of online communications that have been raised within this thesis (see 2.1), include the degree of belief penetration within a society (i.e. the belief's "success"), its rate of spread, and possible polarization outcomes, which cannot be inferred from the individual level alone. However, before discussing the model architecture, it is necessary to give a brief overview of the ABM technique.

6.1 Agent-Based Modelling

ABMs have been used in areas of the natural sciences (Pogson, Smallwood, Qwarnstrom, & Holcombe, 2006), medicine (Segovia-Juarez, Ganguli, & Kirschner, 2004), and ecology (Blanchart et al., 2009; Gimblett, 2002; Janssen & Ostrom, 2006), as well as the simulation of human-based systems (Bonabeau, 2002; Epstein, 1999, 2006), including economics (for a brief overview, see Farmer & Foley, 2009), and epidemiology (Frias-Martinez, Williamson, & Frias-Martinez, 2011), as well as the aforementioned areas of social psychology. An incredibly flexible tool, almost any system can be specified within an ABM using three components:

Agents. The individual actors within a model. In the present work, they represent individual learners within a network. Cognitive rules (learning processes), simple behaviours (propagating a belief to a neighbour) and values (initial priors and belief states) are encoded into agents based on an assigned category (in the present work, all agents are considered "learners"). Agents are entirely autonomous from both each other, and the environment, enacting behaviours and updating values in accordance with their "micro-motives" and interactions with others.

Patches. Represent the environment of the simulation. Patches may similarly be encoded with rules and values, and show the same autonomy. In the present model, patches are not needed.

Links. Represent connections between agents. These, much like the preceding elements, can be created or destroyed, take values and enact behaviours. In the present models, links form the network connections between agents, creating the structure of the network, and are used to carry "belief" messages from one agent to another.

Once an ABM has said architecture in place, it is then possible to run dynamic, adaptive simulations (in the present work, of how beliefs spread across a network) over time, incrementing across parameters of interest. The general process for model creation starts with devising the specific research question: For example, "What will be the effect on behaviour [X] as a result of intervention [Y]?" The research question then constrains, through abstraction and idealisation, the creation of the simulation environment and the agents within it, such as the values each agent attains – whether psychological (e.g., trustworthiness, need for cognition), biological (e.g., height, weight, age) or physical (e.g., wealth, location) as well as the rules for behaviour within the system.

Importantly, in regard to the motivation for the present chapter, the dynamic and interactive element of the modelling technique allows for structures to emerge that are computationally intractable and inherently unpredictable if trying to explore agent decision making and reasoning in isolation. Further, one of the most powerful aspects of ABMs is the capacity to introduce *heterogeneity* to these values and rules among agents and environment. For example, agent values and behavioural rules can be drawn from specified distributions, or allocated purposefully according to validated criteria, and similarly for areas of the environment (patches) – such as resource allocation – and for links – such as strength and proximity (a thick, short link could represent a strong, close link between the two agents, like a family bond).

The second aspect of ABMs is the capacity to observe how these elements (agents, environments, and links) – once set up – change and deviate from one another, *over time* (i.e. models are diachronic, rather than synchronic). This capacity is necessary to observe behaviour, which would not be possible in a static system. The dynamic element, combined with the heterogeneity in the system, allows for the interactions between all the elements within the system to "grow" societal level behaviours (Epstein, 1999, 2006). Importantly, these simulations do not rely on linearity, and further can demonstrate behaviours that are in constant flux (i.e. do not reach an equilibrium, contrary to game theoretic and standard economic assumptions). The simulation sequentially provides each agent at a given time point the opportunity to act in a random order, based on a Markov process. Elements in the system can update their respective values, affect others, enact behaviours, adapt and modify these behaviours, as well as both create new agents (allowing for assessments of evolutionary advantage over time) and be removed from the system (i.e. die). In this way, system elements such as positive or negative feedback between agents, or between agents and the environment, can result in interesting *emergent behaviours* (i.e. behaviours that cannot be predicted from the initial conditions of the model).

All of the properties of the model in the initial conditions, whether agent and environment values or system properties such as number of agents (or proportions of agent types) can be manipulated. This, in combination with the capacity to record any variables of interest (both in snapshot and dynamically), allows for meaningful observation (see Gilbert, Hawksworth, & Swinney, 2009, for an illustrative example of the incrementing of parameters (mortgage policies) to observe system wide behaviours (housing market bubbles)). Importantly, in much the same way as more traditional computational modelling techniques used in psychology, the principle of “garbage in, garbage out” applies. More specifically, for the outcomes to be of merit, the architecture of the model needs to be validated against the relevant psychological literature and empirical data (and preferably against multiple known aggregates, to diminish the possibility of model over-fitting). This has an additional wrinkle in ABMs; one must be careful when designing the structure of the model (both agent behaviours and environmental structure) that resultant behaviours are not the result of over-fitting the architecture (i.e. the model design forces the desired outcome).

In summary, ABMs provide a useful tool for extrapolating from individual-based, empirical findings, to systems in which multiple agents may interact with each other and the environment over time. This allows for controlled simulation of dynamic, aggregate behaviours, which are not readily inferable from base conditions.

6.2 Modelling Rationale

The previous empirical work in this thesis has focused on the processes of the individual belief recipient, exploring the ways in which erroneous beliefs can find support through misinterpretations of first-hand, probabilistic evidence. However, although the importance of this empirical work, when considering the availability of information and ease of communication in the more interconnected world of the internet age, has been alluded to (see 2.1), further support is required to take such claims seriously. In this regard, the current chapter serves as a proof of concept for the application of belief-evidence interactions found in the individual, to a group or societal level. It is important to make clear, once again, that the present work, in serving as a proof of concept, is exploratory, rather than a complete statistical treatment of the topic. As a consequence, as is commonplace in ABM in affiliated topics including opinion dynamics (Duggins, 2016), and social psychology (Gray et al., 2014; Smith, 2014), results are discussed in terms of directionality, rather than hypothesis testing.

Importantly, by extrapolating individual-based results to an environment in which multiple agents exist and are interconnected, it is possible to demonstrate emergent behaviours, such as non-linear cascades of belief-spreading (akin to information cascades, see Cui, 2016, which show detrimental power law effects), and belief segregation (found in cyclical opinion dynamic simulations, see Duggins, 2016; Ngampruetikorn & Stephens, 2015). In this way, it is possible to demonstrate conditions under which erroneous beliefs are “corrected” (and thus fail to propagate across a society), or thrive. Further, through the interactive, heterogeneous elements of the simulation space (whether structural, such as physical location in the network, or evidence based, such as the generation of probabilistic evidence) it is possible to explore outcomes that are implied by the empirical work (see sections 3.6 and 4.5), such as the degree of penetration, speed of spread, and clustering or polarization behaviours.

6.3 General Model Outline

In outlining the general model for the present work, it is necessary to subdivide the model into layers of cohesive functionality. Starting at the more elementary levels (where the patches and links are specified in terms of environment), then moving through components such as the representation of time points, the behaviours of individual learner (agents) and selection of values, before finally coming to a holistic picture of the model – as seen through the prism of the behaviours of interest. Within each layer, necessary abstractions and idealisations are explicitly indicated with reference to the original material upon which the function is based, whether empirical work contained within this thesis, associated literatures, or real world examples.

6.3.1 Environment

As previously indicated, the environment selected for this work is a social network. Such a network is constructed of “nodes” (agents), which are a representation of individuals within the network, and links between said nodes, which are representative of the relationships between individuals. In this way, physical space between agents (and thus the length of the link) is representative of the closeness of the relationship. Thus, in the setup of the model, links are randomly assigned from an agent to their closest, unlinked neighbour outwards. The simulation takes place within a 2D space (see Appendix D for an example set-up of the space). As such, the only role patches play within the present models is in providing the 2D space onto which agents are randomly assigned (and affixed). Given the network-based functionality of the model (i.e. beliefs move across links of any given length), the x and y dimensions of the simulation space do not play an active role beyond the link assignment protocol. This formulation has previously been used in various literatures investigating information cascades (Cui, 2016), and social network pruning (Ngampruetikorn & Stephens, 2015).

Both the number of agents within the network, and the number of links between agents are explicitly parameterised to reflect the chosen target of simulation.

When determining the environment architecture, it is necessary to specify the role time or “ticks” play within the simulation. Ticks can represent whatever time point is meaningful to the target of the model (e.g., 1 tick may represent a nanosecond in a particle simulation, or 1 day of trading in a stock market simulation). Such a representation is created through the way in which behaviours within the model are defined, requiring coherence across behaviours. For the purposes of this model each “tick” represents the time taken for an agent to evaluate a belief and decide whether to communicate it onward. Additionally, given this dynamic element of the model, it is necessary to specify termination criteria for the simulation. Such end-states include the reaching of an equilibrium point, a constantly recursive or dynamic state of flux, or when all change within the system has subsided. Based upon the previous empirical work presented within this thesis, each agent may only receive one belief, then make an evaluation and decide to pass it on (such an idealisation is explained further in the belief evaluation section that follows). Given this “one-shot” nature of belief transfer, the end-state for the simulation is when no more beliefs are being transferred across the network, whether due to complete saturation of the belief, or belief isolation via refuting agents / nodes.

Finally, it is necessary to specify the number of agents in the simulation, and the number of links between agents. The number of agents within the simulation is set to 1000. However, the number of links between agents (i.e. the degree of interconnectivity across the network) is explicitly manipulated within the first model, as the capacity to communicate (unlike the total number of agents, which is a proxy for the total simulation time) is likely to impact the belief transfer effects under investigation (Duggins, 2016).

6.3.2 Belief Evaluation: Learning Model and Evidence

Briefly, it is important to note that all learning (i.e. from the receipt of a belief, to evidence-based evaluation of said belief, to the decision to propagate) takes place within any and all agents autonomously of one another, as soon as they are "activated" via receipt of a communication.

Belief as a Prior. Upon receiving a belief from a neighbour, the first logical question to answer is how is the initial interpretation of said belief is represented in the model? In answering this, it is first necessary to address the purpose of said belief. In accordance with the preceding work in this thesis, the communicated belief is considered a directional, generic (Cimpian et al., 2010) statement regarding one of two options being superior. As such, the belief can be constrained in terms of a probability distribution between 0 (the belief is completely false / the opposite option is the optimal choice) and 1 (the belief is completely true / the indicated option is the optimal choice). Thus, before receiving any information regarding the two options, the agent's prior belief can be idealised as P(H) = 0.5 (i.e. the probability of H, the option to be indicated by the belief as optimal, is in fact optimal, is neutral), an approach that has been used in previous Bayesian studies (Harris et al., 2015).

From this point, one is left with the unenviable task of translating an unquantified (subjective) statement (the belief) into a quantified parameter (the prior). Such a task is not readily solvable, and is likely to require an extensive mathematical treatment. However, given the supposition that the model is based upon previous research included within this thesis, the effect of a "belief" on the formulation of the prior can be sensibly idealised from empirical data. Accordingly, priors were elicited via a short experiment run online using 100 US participants on Amazon Mechanical Turk. Using the same general method as Chapter 4, participants were instructed regarding the treatment of patients (trials) presenting with one of two diseases, each with two possible medicines. Participants, having seen the belief manipulation (the same as that used in 4.2 and 4.3), immediately made probability estimate judgements (i.e. an indication of their prior, having experienced only the belief). In this way, it was possible to validate the choice of prior against empirical data, as previous experiments had only elicited such probability estimates *after* experiencing evidence (for a full account of the experiment, including the phrasing for how priors were elicited, see Appendix C). Consequently, the mean probability estimate indicated a deviation from the neutral point (0.5) of .1148 towards the belief indicated option. It is this value that is either added to (in the case of receiving the belief), or subtracted from (when the opposite, refuting belief is passed on) to form the prior.

Evidence Exposure. Upon receiving the belief (which is transferred via a link between the recipient, and a propagating agent), in line with the empirical work within this thesis, agents are then exposed to probabilistic evidence. For the sake of fidelity to the experimental work upon which the present work is premised, agents receive probabilistic evidence that, over the course of 50 trials, probabilistically favours the alternative (60%) over the belief indicated (40%) option (see 4.2 for empirical method description). Further, to reflect the initial evidence manipulations of the same work, agents are either exposed to two initial trials of supporting or undermining evidence (with a 50% likelihood). An idealisation is made to simplify the presentation of evidence, wherein trials are either positive (1), meaning the trial supports the belief, or negative (0), meaning the trial undermines the belief.

Learning Model. To determine how the combination of belief and trials of first-hand evidence are integrated, it is necessary to represent the updating of this belief given first-hand evidence, within an agent. To do this, a reinforcement-learning model was incorporated into each agent. This model was based upon reinforcement-learning models used in the assessment of learning in fMRI versions of an abstract-symbol categorisation task that shares broad methodological similarity to Chapters 4 and 5 (Decker et al., 2015; Doll et al., 2011, 2009; Staudinger & Büchel, 2013). The key parameters of the classic reinforcement-learning model, shown in equation 3 below, are the prediction error (δ), which when multiplied by the learning parameter (β), and added to the value associated with the belief prior for the current trial (Q(t)), lead to an updated belief value (Q(t + 1)).

$$Q(t+1) = Q(t) + \beta\delta(t) \tag{3}$$

However, in addition to these parameters, following previous models of instruction based bias (Doll et al, 2011, 2009), a parameter was added to reflect the confirmatory bias in updating. This additional parameter is conditional upon whether the direction of the prediction error is either in a supportive (positive), or an undermining (negative) direction.

$$Q(t+1) = Q(t) + \beta(\alpha_I\delta_+(t) + \delta_-(t)/\alpha_I) \tag{4}$$

In equation 4, the asymmetry towards confirmatory events is represented by the α_I parameter, which differentially affects positive and negative events. In this way, positive (belief congruent) prediction errors (δ_+) are multiplied by the bias parameter (α_I), whilst negative (belief-incongruent) prediction errors (δ_-) are divided by the bias parameter. Positive prediction errors are set to zero in negative trials, and vice versa. Such a model has been demonstrated as a superior fit to the standard model when looking at instruction based effects in evidence integration (Decker et al, 2015) on individuals. This confirmation bias parameter is explicitly incremented within the simulation process, as $1 \leq \alpha_I \leq 1.4$ (iterated in .01 increments. In this way, when α_I is set to 1, there is no confirmation bias influence on learners. This incrementation (rather than using a fixed bias parameter) is to prevent the results of the simulations in effect being “baked in” given the ingredients. Put another way, instead of confirmation bias being a fixed constant, incrementation allows for the exploration of how the impact of this parameter is exacerbated or diminished by other structural and behavioural elements of the model.

Finally, in line with the large main effect of initial evidence found in experiments in Chapters 4 and 5, a primacy parameter is fitted to the model. This additional parameter is a linear function based on the number of experienced trials (equation 5). Such an implementation, wherein *t* represents the ordinal position of the trial, and *C* represents a multiplying constant, fits previous work on the parameterisation of primacy effects in serial judgements (Anderson, 1965; Hogarth & Einhorn, 1992).

$$\alpha_P = \frac{1}{1+(t-1)C} \tag{5}$$

This parameter is then fitted to the model in equation 6, such that all prediction errors, irrespective of direction, are multiplied by the primacy parameter. Given that $0 \leq \alpha_P \leq 1$, and the specification of α_P as presented in equation 5, this results in a gradual underweighting of evidence with trial position.

$$Q(t+1) = Q(t) + \alpha_P \beta(\alpha_I \delta_+(t) + \alpha \delta_-(t)/\alpha_I) \tag{6}$$

Bayesian models of learning have gained popularity in cognitive psychology (Breen, 1999; Bröder & Schiffer, 2003; Dave & Wolfe, 2003; Harris et al., 2015; Lagnado et al., 2007; Lee & Wagenmakers, 2013; Sloman & Fernbach, 2011; Yu & Lagnado, 2012). However, a framework that focuses on order effect dependencies in confirmation bias, within a persuasion (belief-transfer) context, is yet to be demonstrated as an appropriate fit (Allahverdyan & Galstyan, 2014). This is not to preclude the adaption, use and comparison of such models within future work. However, given the purpose of the present work to focus on societal level interactions of individuals, the representation of individual cognitive apparatus is, for now, more parsimoniously represented by a readily adapted reinforcement-learning model.

6.3.3 Propagation of Belief

Upon the completion of the evidence evaluation by an agent, the agent is left with a posterior value pertaining to their judged validity of the communicated belief. For the sake of parsimony, an idealisation is made to reflect a binary account of either ending in a state of believing (Posterior > .5) or having refuted (Posterior < .5), and bears a parallel to the posterior binary preference measure used in the empirical work (see Experiment 3.3 onward). This simplifies the relationship between the degree of belief (and confidence in this belief) and the propensity to then communicate said belief. Such an idealisation has validity, given the exploratory, proof of concept focus of the present work, rather than specific hypothesis testing. However, such a relationship can be incorporated into later, more complex models. In the preliminary models explored in the present work, a simple decision tree is instead followed for the propagation of belief.

Given an agent is in a state of conclusion (i.e. the evidence integration process has finished), the agent is then in a position to communicate the belief (or in some model iterations, the refutation of the belief) to neighbouring agents. To do this, the agent asks:

1) Are any neighbours (agents with whom I share a (relationship) link) yet to receive any belief (state zero)? If yes:
2) Which “untouched” agent is closest in proximity to me?

Assuming a target agent meets this criterion, the propagator then passes a belief onward to the target, and does so in the order of closest to most distant relationships (in this sense physical proximity in the model is representative of relationship proximity, and is the aforementioned sole influence of environment within the model). In the case where a propagating agent no longer has any neighbours yet to receive a communication (whether from the propagator in question, or another linked propagator), then the propagator ceases attempting to communicate (as agents are fixed to their randomly allocated x and y coordinates). Such a constraint reflects the one-shot nature of the belief-evidence integration experiments in the preceding chapters upon which the architecture of the present models is validated. Further, the present models are constrained to a single "starting point" of the belief. Although systems in which multiple progenitors exist are of interest, with particular regard to competing belief content, such explorations are left to further work.

6.3.4 Behaviours of Interest

Given the premise of belief transmission across a network, several variables are of interest for the exploratory, proof of concept nature of the present work. These variables are chosen for the insight they can provide into network-level behaviours, so as to make use of the novel, diachronic, interactive, heterogeneous system. Further, these variables provide the most suitable summary description of the behaviours, and in doing so, address the motivation behind the present work (namely the assessment of overall levels of erroneous belief uptake, speed of spread throughout the network, and level of segregation among different believers).

Firstly, the degree of penetration of a belief across the network (taken as a percentage of "untouched" agents out of the starting total remaining when the simulation end conditions are met) provides one indicator of the "viral capacity" of the belief under the given conditions. Secondly, the percentage of total number of agents who believe the communicated belief (termed "believers"), along with the percentage of those who have refuted the communicated belief (termed "refuters") gives a sense of the success of the belief as an overall statistic for the given simulation conditions. In particular, the relative proportions of these two percentages are of interest (i.e. of those exposed to a belief, how many ended in a state of belief or refutation). Such proportions can be tentatively compared to the empirical data of Chapter 4. Given that the models do not yet include elements of source credibility, the posterior binary preferences for the belief disease, from the joint analysis of Chapter 4 (see 4.4), where N = 646, provides the most suitable proxy for comparison. This judgement forced participants to specify which of the two options they preferred having experienced first-hand evidence (i.e. the same that an agent in the simulation has to make in declaring a state of believing or refuting); either the belief indicated option, or the optimal (non-belief) option. It should be noted that this comparison does compare model outputs in which the end state may involve a proportion of those who are never exposed to the belief, to an outcome based on all individuals having been exposed. Despite this caveat, it is still of interest to compare outcomes of an interactive, non-linear system, to one in which exposure is systematic and in parallel. Consequently, the proportion of those defined as "believers" for the purposes of this comparison from the joint analysis data is 46.6% (301 out of 646).

Thirdly, the rate of propagation, taken as the percentage increase in those who have had a belief communicated to them from one time point to the next, is a measure that can provide information regarding both the peak and mean "spreading rate" for each given set of conditions. Such a capacity to assess speed of spread is a unique consequence of the diachronic nature of ABMs, and captures an element of the "viral capacity" of information-spread across networks.

Finally, the degree of clustering, taken as the mean number of "likeminded" linked agents (i.e. for each agent, how many linked neighbours match their belief state), is a useful indicator of societal patterning in belief adoption. This measure allows for the assessment of possible segregation-like behaviours across a social network, a behaviour that has been shown to take place as a consequence of "pruning" of contradictory opinion holders in social networks (Ngampruetikorn & Stephens, 2015). The present models do not allow for pruning behaviour, instead focusing on segregation as a consequence of a one-shot cascade of belief evaluations (and propagations) across a network (i.e. given equally rational agents with neutral priors, does the stochastic nature of the belief-evaluation process, in combination with neighbour-to-neighbour propagation, result in segregated groups of "believers" and "refuters" even before any cyclical behaviours like pruning occur).

Exploratory predictions are presented below, in relation to the specific model manipulations and constraints.

6.4 Model 1

The purpose of the Model 1 is two-fold. Firstly, it is to provide a general assessment of the basic model construction's validity in yielding sensible outcomes given the constraints proposed above. In brief, agents are placed within a static system, and connected via links to nearby agents (6.3.1). Each agent acts as a learner, who is activated upon receiving a belief from a linked neighbour. Activated agents then evaluate the belief in light of first hand evidence, using a simple learning model that incorporates the empirical work of the thesis (6.3.2). Having evaluated the belief, agents then pass on the belief (or refutation) to neighbouring agents (6.3.3). The network-level behaviours of interest are the degree of spread across the network, the peak rate of spread, and the amount of segregation among different belief types. Secondly, and linked to the primary purpose, the objective is to test several questions (i.e. condition manipulations) which are not only of interest given the novelty of societal extrapolation, but of critical importance for the further development of the model (and the additional complexity / functionality this entails). These questions can be broken down into the manipulation of three parameters; whether refutation beliefs are also propagated (binary), the average number of links between agents (i.e. the degree of interconnectivity on the network), and the degree of confirmation bias in the evidence assessment.

Refuter Actions. The first of these manipulations, whether agents who have concluded that the belief is false (Posterior < .5) decide to propagate such a prior to their neighbours (via links), is of interest for several reasons. Firstly, research into pseudoscience and illusions of causality has suggested, based on experimental data taken on an individual level, the benefit of presenting null information in the mitigation of fallacious belief uptake (Matute et al., 2011). Although such recommendations are based on the individual reasoning level, it is an intriguing extrapolation to extend this principle to communications across a network. For example, given the premise of a typical pseudoscientific belief (e.g., ingesting ground rhino horn will increase virility), and assuming all agents are of equal standing in terms of source credibility (i.e. no single agent is in a more authoritative position than any other), then the decision to pass on null information is not necessarily obvious. Specifically, by passing on a belief regarding this association, irrespective of the direction, creates an association in the mind of the recipient that otherwise would have gone unformed. As a consequence, given the recipient now has this information, they may evaluate the evidence for themselves and come to a different, more harmful conclusion (e.g., via a different prior regarding the efficacy of these types of treatments). The recipient may then pass on such a belief to another acquaintance, and so the belief, instead of dying with the original refuter, goes on to spread throughout the network, despite the refuters good intentions.

Secondly, the importance of communicating null information bears a parallel to other domains (and in so doing, brings further relevance to the manipulation of this parameter), including science and science journalism. Many are familiar with the “file drawer problem" – the underreporting of null effects (Rosenthal, 1979) – and similarly in journalism, the tendency in various media outlets to report scientific findings in an irresponsible manner (Bell, 1994). To explain the latter more fully, given the premise of an exciting sounding scientific finding (e.g., the latest "link" between a certain food and cancer), those journalists who read the original article (evidence evaluation), and conclude there is not enough evidence to substantiate writing an article *move on* to write about something else. Meanwhile, the journalist who concludes (erroneously) that the effect is worth reporting (e.g., by never bothering to fully investigate / understand the original paper), do so under the impression it is true. Thus, readers are only exposed to the (false) positive, whilst the "truth" may only become newsworthy if the exposure of the "lie" becomes worthwhile (i.e. the belief is widespread enough that reporting its falsity becomes of interest). Taken together, both file drawer problems and selective reporting are examples of natural instances in which there is an asymmetry in the communication of "believing" vs "refuting" information.

Interconnectivity. Given the rise of social media and the democratisation of knowledge via the internet (Dylko et al., 2012; Kahn & Kellner, 2004), the question of how this ever-increasing interconnectivity among individuals impacts the spread of beliefs, and how such a growing trend interacts with behaviours with communication decisions (such as passing on null information) is of increasing importance. Through explicit manipulation of the number of links between agents, it is possible to demonstrate, for example, the relationship between the degree of interconnectivity and both the rate and degree of belief spread across a network.

Confirmation Bias. Finally, it is of interest given the prior empirical chapters, to assess the impact of the confirmation bias parameter within the learning model, on the "success" of the erroneous belief. Although to some degree, the likely impact (higher levels of confirmation bias leading to higher amounts of belief uptake) is straightforward, the explicit iteration through this parameter helps provide validation for the model against empirical data. To explain further, it is of interest to iterate through the confirmation bias parameter in each of the system conditions to find a value that approximates the effect sizes found in the preceding experimental chapters. In this way, more complex models can adopt a “default” confirmation bias value when iterating across a number of other variables, simplifying the model outcomes in a sensibly constrained manner.

One can draw a number of exploratory predictions of the interplay of these factors, which follow logically from the model constraints. Firstly, that in general, as confirmation bias on the part of individual learners increases, the amount of “believers” in the system will increase, to the point of complete saturation. Secondly, as interconnectivity increases, beliefs have more available routes through which to spread, and thus rate of spread will be higher. Thirdly, when refuters remain silent, higher levels of belief uptake are expected than in refuter active systems. Finally, belief segregation increases with the dominance of one belief type (i.e. if the system is full of believers, then logically the number of “likeminded” individuals will be higher than in more equally proportioned systems). These exploratory predictions, although logically apparent, help assess whether the model, in general, is working *as expected*. This does not preclude the observation of behaviours not immediately apparent from the model setup.

Several questions are of particular interest, including:

1) How do the actions of refuters and the increasing levels of integration in the system affect the amount of confirmation bias required by individual learners for majority erroneous belief uptake in a system?

2) Does increasing interconnectivity lead to increased levels of belief rejection, holding levels of confirmation bias constant?
3) Is confirmation bias even required on the part of individual learners under certain system conditions (e.g., highly interconnected, but refuter silent networks)?

6.4.1 Model Setup

Model 1 is set up in a manner indicated by the general model outline, such that the total number of agents in the simulation is 1000, the initial starting point of the belief is a single agent occupying the centre-point of the simulation, and the end state occurs when there is no change in the proportion of agent belief states between two adjacent time points (whether through complete saturation or pre-mature belief "death"). The specifications for the belief evaluation process and propagation decisions also follow the outline set out in the general model, including the specification of the belief initial impact on the recipient (i.e. the prior) being formulated as 0.6148 for belief recipients, and 0.3852 for refutation recipients (see 6.3.2 and Appendix C for details).

The independent variables of interest – the number of links between agents, the action of belief "refuters" in either passing on their refutation belief or staying silent, and the confirmation bias parameter – are specified accordingly:

Interconnectivity. The number of links between neighbours within the model is generated in such a way that a degree of heterogeneity is allowed among agents, in that the number of links an agent has is drawn from a normal distribution with a mean of the manipulation value (3, 5, or 7 links). In this way, a more accurate reflection of the target system is created (every individual does not have precisely the same number of relationships, whether on or offline). The degree of interconnectivity is arbitrarily split into three levels; low interconnectivity is defined as a "3 link" system, as this is the minimum value before the network becomes fractured, medium interconnectivity is defined as a “5 link” system, and high interconnectivity is defined as a “7 link” system.

Refuter Actions. For the purposes of the basic model, an abstraction is made regarding the decision to propagate the agent’s conclusion. As previously indicated in the general outline, agents’ final belief-states are dictated solely by the value of their posterior (i.e. post-evidence evaluation). Specifically, agents are categorised as “believers” if their posterior is above neutral ($P > .5$), or “refuters” if their posterior is below or equal to neutral ($P \leq .5$). The decision to then pass on the belief is automatic for the former (assuming there is a linked neighbour who is yet to be communicated to by any agent: an “untouched” agent), whilst the decision for the refuter to do so is manipulated as either *on* or *off*. This idealisation can be adjusted in subsequent, more complex models to look at, for example, the relationship between belief confidence and the propensity to communicate, as well as incorporating social factors such as conformity. For now, given the “proof-of-concept” nature of the basic model, refuter actions are manipulated as a 2-level (binary) factor.

Confirmation Bias. The confirmation bias parameter within the learning model (α_I) is iterated through for each of the 6 (3 levels of interconnectivity by 2 levels of refuter actions) model types. The parameter is iterated from 1 (no confirmatory impact) through to 1.4 (severe confirmatory impact), in levels of .01. This specification allows for the detection of thresholds of successful belief spread within each model type. In this way, it becomes possible to demonstrate differences in the degree of confirmation bias required for the spread of erroneous beliefs across the different model types.

Behaviours of Interest. Finally, the three behaviours of interest, as indicated in the general model outline, are used for the basic model. The first of these is the percentage of the total agents for each belief-state (believers, refuters, or untouched) at the end of the simulation, which indicates the "success" of the belief across the system. The second is the peak rate of spread (taken as the difference in the number of untouched agents between two adjacent time points, as a percentage of the total agents in the system), which is used as a measure of the speed of belief penetration within the system (i.e. the viral capacity of the belief). Lastly, the degree of clustering (the percentage of linked neighbours with the same belief-state as the agent, averaged across all agents with an activated (i.e. believer or refuter) belief-state) is used as a measure for the degree of belief segregation within the system, or the degree to which (after an initial spread of a belief) people are exposed to those of an opposing belief state.

Finally, Table 6.4.1 below provides a summary of both the manipulated system specifications and the behaviours of interest for the basic model.

Table 6.4.1: Summary of Independent Variables and Behaviours of Interest for Model 1

Independent Variables	**Values**	**Increment**	**Levels**
Interconnectivity	3 links, 5 links, 7 links	2	3
Refuter Actions	Active, Silent	n/a	2
Confirmation bias	1 < x < 1.4	0.01	41
Behaviours of Interest	**Scale**	**Description**	
Success of Belief	0-100	Percentage of agents (out of total), split by belief-state ("believers", "refuters", and "untouched") at simulation end point.	
Peak Rate of Spread	0-100	Peak percentage change in "untouched" to belief-activated agents, across a single time point.	
Degree of Clustering	0-1	Mean proportion of likeminded (belief-state matching) linked-neighbours at simulation end point, across all agents.	

Each simulation set-up (Confirmation Bias (41) x Interconnectivity (3) x Refuter Actions (2); resulting in 246 different simulation set-ups) was independently run 100 times, resulting in a total of 24,600 runs using the RNetLogo package in R. In this way, dependent variables were taken as the mean across the 100 runs of each simulation set-up (for an example GUI demonstration of a model set-up, see Appendix D).

6.4.2 Results

Success of Belief. Looking first at the percentages of belief uptake across the system, there are several interesting general trends of which to take note. Primary among them, as evidenced in Figure 6.4.1, is the expected increase in percentage of believers as the system moves across the confirmation bias parameter space, as individuals become more susceptible to erroneous belief uptake. Secondly, when refuters are silent (right-hand column in Figure 6.4.1), the overall levels of refuters in the system remains substantially lower than in active refuter systems. This once again demonstrates the model behaving as expected, wherein successful refutation cannot rely on first-hand integration alone.

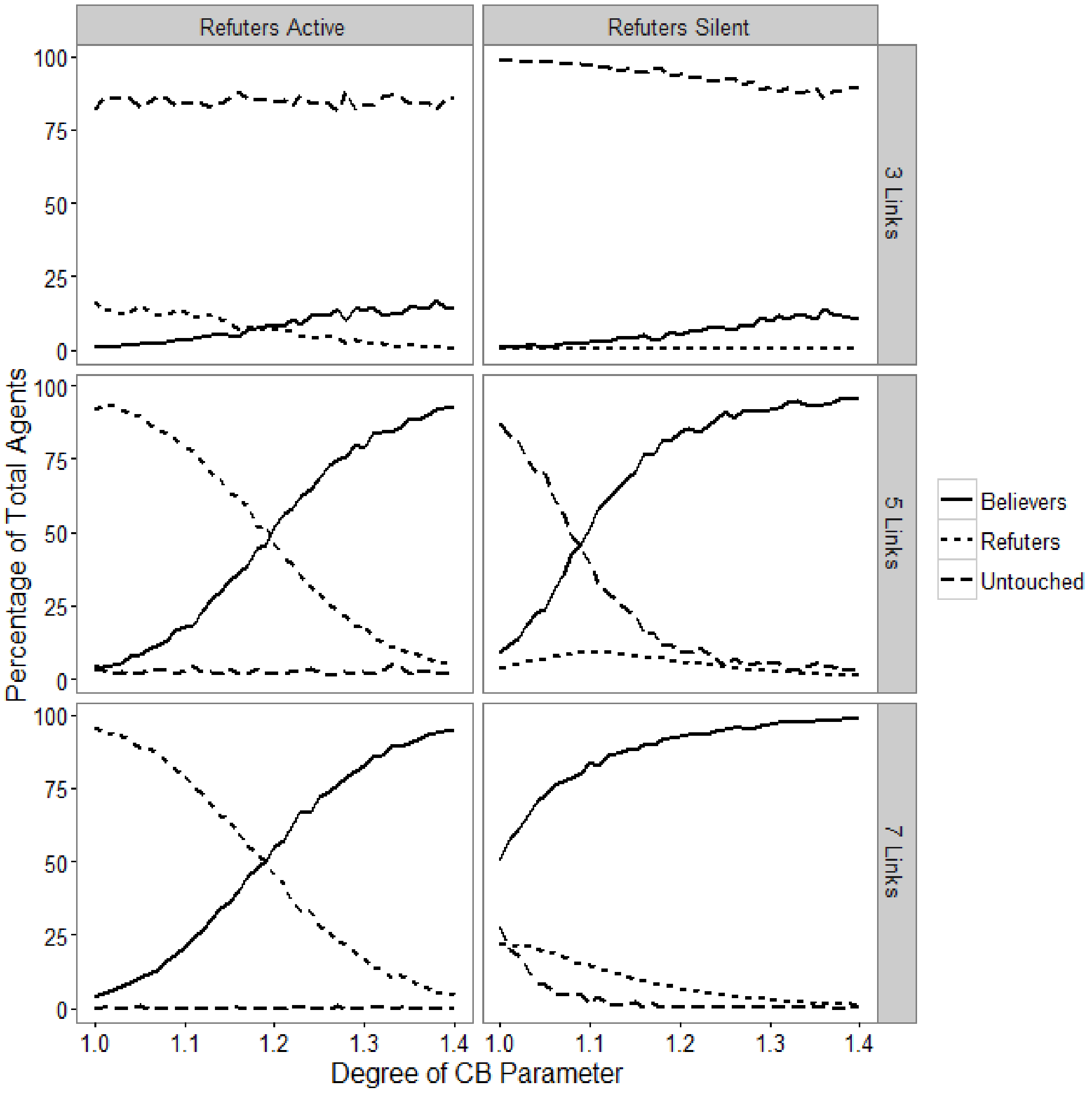


Figure 6.4.1. Model 1: Percentage of Agents at Simulation Termination Point, split by Belief-state. Values are averaged across 100 iterations.

Thirdly, as a system becomes more interconnected (going down rows in Figure 6.4.1), the more saturated (i.e. the smaller the proportion of “untouched” agents in the system) it becomes, whether with refuters or believers. However, when looking at the interaction of interconnectivity with refuter actions, there are several effects of interest. For example, in less interconnected systems, silent refuters in effect *cauterize* the spread of belief (as evidenced by the greater percentage of untouched agents, irrespective of the confirmation bias parameter, in refuter silent, 3 link systems, relative to refuter active systems of the same interconnectivity). Such an effect is attributed to the inability of beliefs to propagate *around* the “blocking” refuting agents in low interconnected systems. Inversely, the more interconnected the system, the more harmful the refuter action of remaining silent becomes, as the asymmetry of only beliefs (not refutations) being passed on as priors means a highly interconnected system can always bypass the refuter “blocks” (as evidenced by the substantially greater numbers of believers in refuter silent, 7 link systems than refuter active versions (bottom row of Figure 6.4.1), irrespective of confirmation bias level).

This leads to the final effects of interest; the impact of confirmation bias on different systems. As has been previously noted, there is a logical impact across all systems of greater degrees of confirmation bias resulting in higher percentages of believers in the system. However, when looking at the threshold points in each system, where believers start to become equally as prevalent as refuters when confirmation bias increases, there is an interaction with refuter action and interconnectivity. More specifically, in refuter active systems, interconnectivity does not change the required confirmation bias parameter value for the threshold to be met, only the degree of overall saturation. Across all levels of interconnectivity, the required confirmation bias parameter value remains at approximately 1.2, in refuter active systems. Conversely, in refuter silent systems, the required confirmation bias parameter value to meet the belief equals refuter threshold *decreases* as interconnectivity increases, with generally lower values required than refuter active system counterparts. Such a finding bears special relevance to the importance of null information communication across the increasingly prevalent highly interconnected networks, such as social media.

Comparing the proportion of believers found when a group of isolated individuals are presented with the belief (i.e. the empirical data of Chapter 4), to these interactive, integrated systems, there are several tentative, but worthwhile elements to note. This proportion of believers (indicated by posterior binary preferences in the joint analysis; 4.4.1) from the data suggests approximately 46% of belief recipients will take up the belief (i.e. adopt the state of a "believer"). In comparison, as a system becomes more interconnected, the degree of confirmation bias required of the individual learners to surpass this threshold is reduced severely in the case of the refuter silent systems. In contrast, when refuters are active, attaining (and surpassing) this empirically-based threshold suggests a required confirmation bias parameter of CB ≥ 1.18.

Peak Propagation Rate. Turning to the peak rates of spread (whether refutation or belief communications), there are several effects of interest. Firstly, the relationship between interconnectivity and rate of spread bears out in the expected direction across all systems, whereby the greater the level of interconnectivity, the faster the spread of belief through the system. Such a finding confirms the exploratory predictions of the model (6.4) and earlier assertions in the present work regarding the ease and speed of belief spread in the modern world (see 2.1).

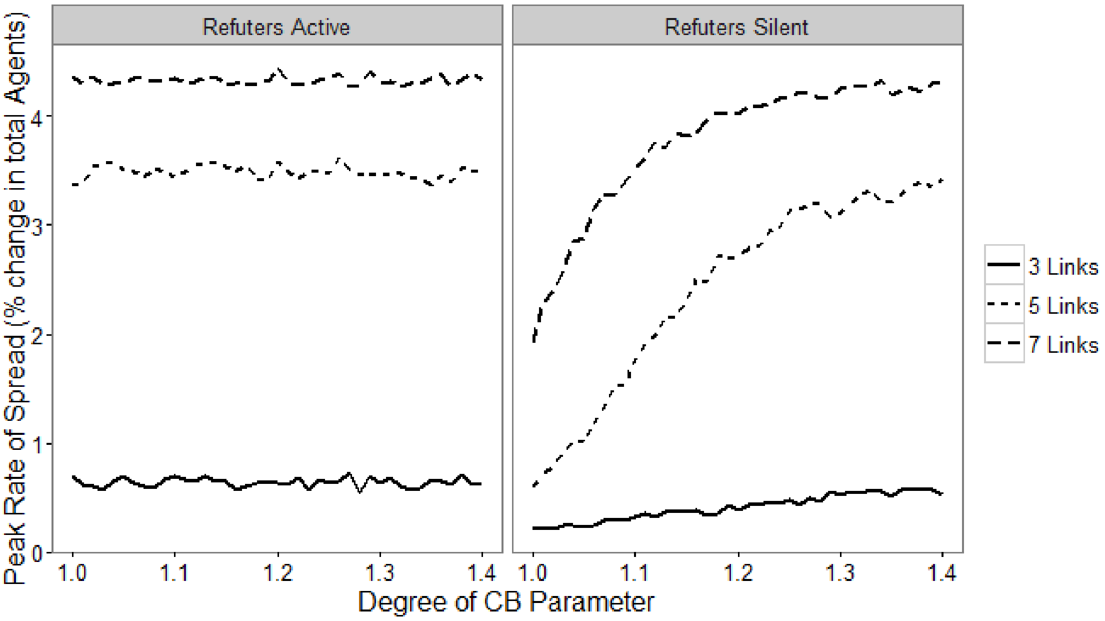


Figure 6.4.2. Model 1: Peak Rate of Spread across simulation. Averaged across 100 iterations.

Secondly, as evidenced in the comparison between columns in Figure 6.4.2, rate of spread is *dependent* upon the confirmation bias parameter spread in refuter silent systems, whilst refuter active systems show no such dependency. This upward trend in silent refuter systems is a likely demonstration of "believer-led" propagation, which requires higher confirmation bias values for greater success (i.e. to fight cauterization within the network).

Degree of Clustering. Finally, turning to the degree of clustering in the system (defined as the percentage of linked neighbours of an agent with the same belief-state, averaged across all "activated" (those who have received any communication) agents), as shown in Figure 6.4.3, there are several effects of interest.

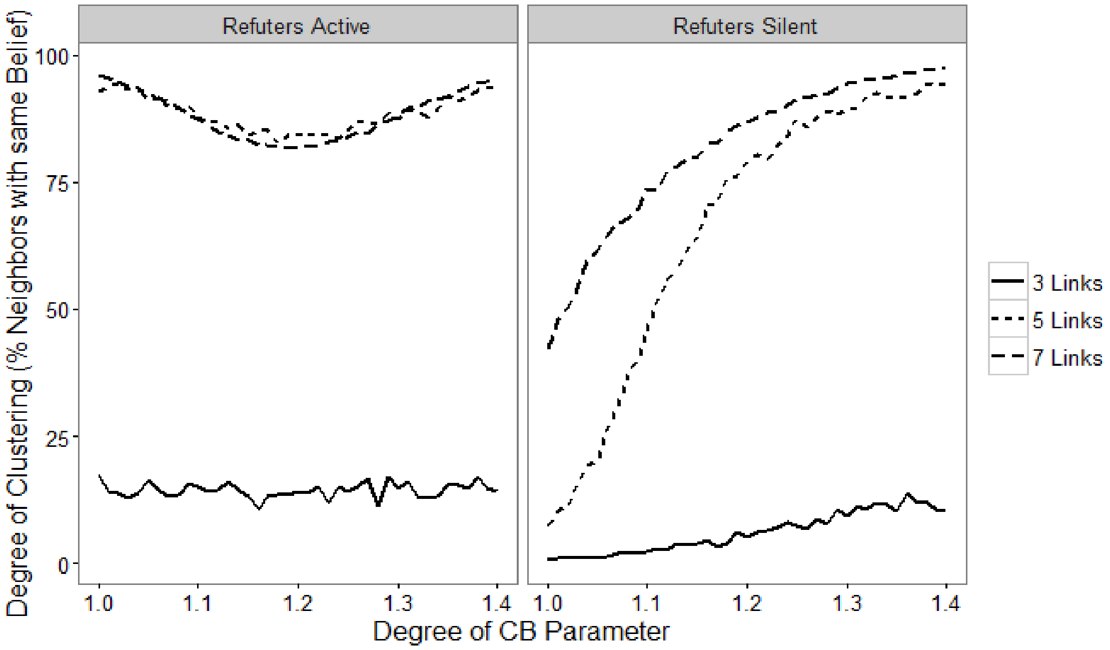


Figure 6.4.3. Model 1: Degree of Clustering, measured by average percentage of same-belief-state neighbours across all believing and refuting agents.

Firstly, much like the rate of spread in the system, the degree of clustering in refuter silent systems is dependent on the confirmation bias parameter space and the degree of interconnectivity in the system. Once again, this is attributable to the dependency on believers to propagate across a system, which in turn is dependent upon the degree of confirmation bias. To explain further, given the cauterizing effect of silent refuters in low interconnected systems, the relative proportion of like-minded neighbours has to remain low. To draw a metaphor from medieval townships, in a small walled town (cauterized belief), there is a higher proportion of houses that will share proximity to the wall, rather than just other houses. However, in a large city that has expanded beyond its original walled boundaries (more interconnected systems), the relative number of houses sharing proximity with walls is lower, and will decrease further the larger the town gets. In this way, as the believers are able to more successfully spread (via increased degrees of confirmation bias), and the relative numbers of refuters (pieces of wall) become lower, increasing the believer driven clustering.

Secondly, active refuters show two interesting trends relating to the degree of interconnectivity in the system. Given clustering's dependency on percentages of active agent belief-states (Figure 6.4.1), in low interconnected systems, across the confirmation bias parameter space, there are relatively equally low proportions of believers and refuters. In other words, the low total number of active agents in low interconnected systems means that within that low total, it is more commonplace for refuters and believers to live alongside one another. In comparison, given the similarity in proportions of believers and refuters in medium and highly interconnected systems, not only is there no perceptible difference in the degree of clustering between these two systems, but the perceptible "dip" in the degree of clustering (when the confirmation bias parameter approximates 1.2) can be attributed to the threshold of equal believer and refuter proportions towards the middle of the confirmation bias parameter space. To put this another way, as the confirmation bias parameter value approaches 1, medium and highly interconnected refuter active systems are dominated by refuters (and thus have a high degree of clustering). Conversely, as the confirmation bias parameter approaches 1.4, the domination of believers in the system results in a mirrored increase in clustering.

Taken together, these findings are generally in line with exploratory speculations (6.4), whilst the sensitivity of clustering to increasing interconnectivity serves as a demonstration of the unique insights ABMs can provide into societal level behaviours.

6.4.4 Discussion

The purpose of Model 1 has been to not only verify the architecture of the model, but to explore the consequences of increasing interconnectivity, the actions of refuters, and the degree of confirmation bias in evaluating beliefs on the system-wide behaviours of belief penetration, speed of spread, and degree of clustering.

Increasing levels of interconnectivity in the system, in line with exploratory speculations, was found to lead to greater levels of communication saturation, increased rates of spread, and higher levels of belief clustering. The findings neatly demonstrate the double-edged sword of society's increasing interconnectivity with the rise of the internet and social media. Furthermore, where once the actions of refuters in remaining silent may have *cauterized* the spread of (in this case fallacious, but could similarly be applied to more truthful) beliefs, such action in the more interconnected systems inherent to the modern day not only facilitate the spread of fallacious beliefs, but further lead to such beliefs requiring lower levels of confirmation bias on the part of the individual reasoner to reach dominant belief-states across the system. This is drawn into sharper contrast by tentative comparison to the aggregated, individual-based (i.e. non-interactive) empirically-based proportions. Specifically, as refuter silent systems become more interconnected, they quickly surpass the proportions expected from the aggregated set individual belief recipients. Such comparisons are of course tentative given the idealisations regarding the propagation decisions of agents, but act as an initial demonstration of the unique insights ABM techniques can provide in informing individual-based data.

In regard to the adaption of the basic model to incorporate more complex components in belief-spread across a network, several interesting threshold points and system choices have been demonstrated to be of key interest for future model architecture and manipulations. Firstly, given the implication of the current findings regarding the challenges facing more interconnected systems, and their relevance (and closeness) to the target systems (e.g., social networks), further models should focus on more highly interconnected systems. Secondly, refuter actions play a critical role in belief uptake across a network, and remain of interest when manipulating interconnectivity.

6.5 Model 2: Incorporating Social Conformity

Building upon Model 1, interesting questions remain regarding the additional psychological effects at play in the public declarations of beliefs across an interconnected system. In particular, the role of social conformity, the pressure to conform to the perceived majority opinion (Asch, 1955; Cialdini & Goldstein, 2004; Wood, 2000), in the adoption of (erroneous) beliefs is of interest. Although it is acknowledged that incorporating social conformity may somewhat “beg the question” in exacerbating erroneous belief uptake, we argue useful and novel insights may be provided by its inclusion. For example, in a highly interconnected system, as the degree of social conformity among individuals is increased, how does this affect the respective amount of confirmation bias required to adopt a fallacious belief on the part of the individual reasoner? Further, how extensive is the impact of conformity on the degree of clustering among individuals of differing belief-states? This latter effect bears direct relevance to other models of information cascades (Cui, 2016), social network pruning (Ngampruetikorn & Stephens, 2015) and opinion dynamics (Duggins, 2016). To explain further, making the assumption that individuals tend to create links or networks with those they know / are familiar with, and given one can only conform with those one is aware of, interconnected systems immediately create localised conforming pressures. This, by the aforementioned assumption, asymmetrically favours (via perceived representation) the local belief-state (or opinion) over the (unseen) distal or global belief-states (or opinions) of the non-local members of the overall system.

The addition of social conformity is further based on elements drawn from preceding sections of the thesis. The attention of a belief recipient to how many of their neighbours have either accepted or refuted the belief can be thought of as a cue to the belief's validity. Consequently, the notion of truth values in propositional learning (2.1.1.1 in this thesis, but see Mitchel et al., 2009) provides a social cognitive theoretical framework for the incorporation of these cues to the assessment of the belief's validity. In brief, when a belief appears to be believed by the majority (and further, by neighbouring, more proximal relations of those closer (and visible) on the network), this may be taken as evidence of the truth value being valid. Further, the preceding empirical work has touched on the impact of such cues (i.e. inferred cues, rather than first-hand evidence based cues) to the belief uptake process, such as the change from one to several comments in the belief manipulations of Experiment 3.2 moving to 3.3, and its noted impact on increased belief uptake.

Similarly, a clearer demonstration of the impact of inferred cues on belief validity comes from the source credibility (Chapter 5; sections 5.3 and 5.4). Here the impact of source trustworthiness (inferred from a statement regarding the source's motivations for communicating a belief; cooperation or deception) was found to be incorporated by participants when evaluating the belief. In particular, when source cues indicated the belief to be valid (i.e. the source was inferred to be credible), levels of belief adoption increased. Given this prior exploration of inferred cues to credibility in the belief adoption processes, when focusing on the individual reasoner, it is of interest to simulate the impact of this social influence on interconnected systems, and how this impact relates to aforementioned factors including the actions of refuters and the degree of confirmation bias among reasoners.

6.5.1 Social Conformity

Given the broad social psychological literature on various forms of conformity among individuals, it is necessary to abstract the potential instantiations of the phenomena to a sensible idealisation within the model. To achieve this, the principles of belief adjustment in light of conformity pressures (Cialdini & Goldstein, 2004; Duggins, 2016), were incorporated into the model in the following way:

Each agent, at the time of belief evaluation, is aware of the relative proportion of their linked-neighbours with each belief-state, and the agents' belief threshold (the value against which an agents' posterior is evaluated as either resulting in a "believer" or "refuter") is adjusted accordingly, so that if surrounded by believers, an agent's threshold for believing is lowered, making their own assertion of belief more likely. Given the exploratory nature of the model, belief-states being publicly viewable by an agent's neighbours (which is within the bounds of sensibility given the online, social media target system) is a suitable idealisation. Such awareness is implemented within the model using the following equation:

$$S = (\sum R - \sum B) \tag{7}$$

In equation 7, the degree of belief conformity in an agent's linked-neighbours (S) is taken as difference between the sum of the refuting neighbours (R) and the sum of the believing neighbours (B), in accordance with principles of social influence in that greater numbers of one opinion should hold proportionately greater sway over the reasoner (Latané, 1981). However, for the sake of simplicity, the current instantiation of social conformity does not incorporate Latané's source credibility factors (relative age differences, socio-economic status, likely future power of recipient, and so forth), which is left to future models. Returning to the difference component, the calculation results in either a positive, negative, or neutral conformity value, this can then be directly applied to an agents' own belief threshold (θ) as follows:

$$\theta_1 = \theta_0 + S\delta\gamma \tag{8}$$

In this way, as equation 8 demonstrates, the final belief threshold (θ_1) is equal to the original, neutral threshold (θ_0) plus the degree of belief conformity in neighbouring agents (S), multiplied by a constant (δ) and the individual agents' sensitivity towards conformity (γ). This last parameter is thus manipulated within the model iterations of 0.5 from 0 (no impact of conformity elements on belief thresholds) to 1 (conformity in the system). Accordingly, equal proportions of belief-states among neighbours will result in no change to the belief threshold. However, when an agent has more believing than refuting neighbours, this will result in a negative value, which in turn reduces the belief threshold, increasing the likelihood of the agent adopting the belief themselves, whilst more refuters than believers makes adoption of the belief less likely.

Given the directional, exploratory nature of the present work, and in relation to the findings of Model 1, several speculations are of interest:

Firstly, although conformity is expected to generally increase the degree of erroneous belief uptake, it is of interest how this impacts confirmation bias requirements on the part of the individual, and secondly, how such an impact may be exacerbated or diminished by the actions of refuters. This is especially interesting given the effects of interconnectivity on increasing levels of belief uptake demonstrated in Model 1 (i.e. given the impact of increasing the number of linked neighbours on belief uptake, who provide paths for beliefs to be spread, when linked neighbours are able to start exerting more direct pressure as well).

Secondly, conformity is expected to increase the degree of clustering and rate of spread, but similarly how such an impact is mitigated or exaggerated by the actions of refuters is of interest.

6.5.2 Model Setup

The Model 2 is also set up in a manner indicated by the general model outline, such that the total number of agents in the simulation is 1000, the initial starting point of the belief is a single agent occupying the centre-point of the simulation, and the end state occurs when there is no change in the proportion of agent belief states between two adjacent time points (whether through complete saturation or pre-mature belief "death"). The specifications for the belief evaluation process and propagation decisions also follow the outline set out in the general model and Model 1, with the exception of the social conformity element acting upon the belief threshold.

Unlike Model 1, in the current model the number of links between agents is held constant at 7 (i.e. representative of a highly interconnected system). This restriction is made primarily given the focus of the present work on the consequences of highly interconnected systems, and secondarily is a useful simplification for keeping the number of models tractable. Given the similarities to Model 1, in that refuter actions and confirmation bias is manipulated in the same manner, and the behaviours of interest are identical, Table 6.5.1 below provides a summary of the independent variables and behaviours of interest for Model 2.

Table 6.5.1: Summary of Independent Variables and Behaviours of Interest for Model 2

Independent Variables	**Values**	**Increment**	**Levels**
Social Conformity	0, 0.5, 1	0.5	3
Refuter Actions	Active, Silent	n/a	2
Confirmation bias	1 < *x* < 1.4	0.01	41
Behaviours of Interest	**Scale**	**Description**	
Success of Belief	0-100	Percentage of agents (out of total), split by belief-state ("believers", "refuters", and "untouched") at simulation end point.	
Peak Rate of Spread	0-100	Peak percentage change in "untouched" to belief-activated agents, across a single time point.	
Degree of Clustering	0-1	Mean proportion of likeminded (belief-state matching) linked-neighbours at simulation end point, across all agents.	

Social Conformity. The social conformity parameter (i.e. the degree to which agents are susceptible to conforming pressures), as represented by γ in equation 8 above, is manipulated over three levels of interest (creating a three-level variable). These are set within the model as either 0 (no impact of social conformity / baseline measure), 0.5 (social conformity has a mitigated impact), and 1 ("full" conformity within the system).

For Model 2, it is expected that the dependent variable of greatest interest will be the degree of clustering in the model space. Given the architecture of the model, and previous findings of Model 1, the conditions during which agents with differing belief-states are in roughly equal proportions are of critical interest. Conditions under which one belief-state dominates (e.g., when confirmation bias parameters that are either very low – resulting in mostly refuters, or very high – resulting in mostly believers) are consequently less impacted by conformity pressures (everyone is already conforming via other processes).

As specified above, and unlike Model 1, interconnectivity is held constant for all systems at 7 links. Each simulation set-up (Confirmation Bias (41) x Social Conformity

(3) x Refuter Actions (2); resulting in 246 different simulation set-ups) was run independently 100 times, resulting in a total of 24,600 runs using the RNetLogo package in R. In this way, dependent variables were taken as the mean across the 100 runs of each simulation set-up. Given the exploratory, proof of concept nature of the model, a directional approach is once again adopted in the discussion of results.

6.5.3 Results

Success of Belief. Looking first at the percentages of belief uptake across the system, there are a number of interesting general trends of which to take note. As in Model 1 and now evidenced in Figure 6.5.1, the percentage of believers increases as the system moves across the confirmation bias parameter space. Similarly, when refuters are silent (right-hand column in Figure 6.5.1), in line with the expectations and findings of Model 1, the overall levels of refuters in the system remains substantially lower than in active refuter systems.

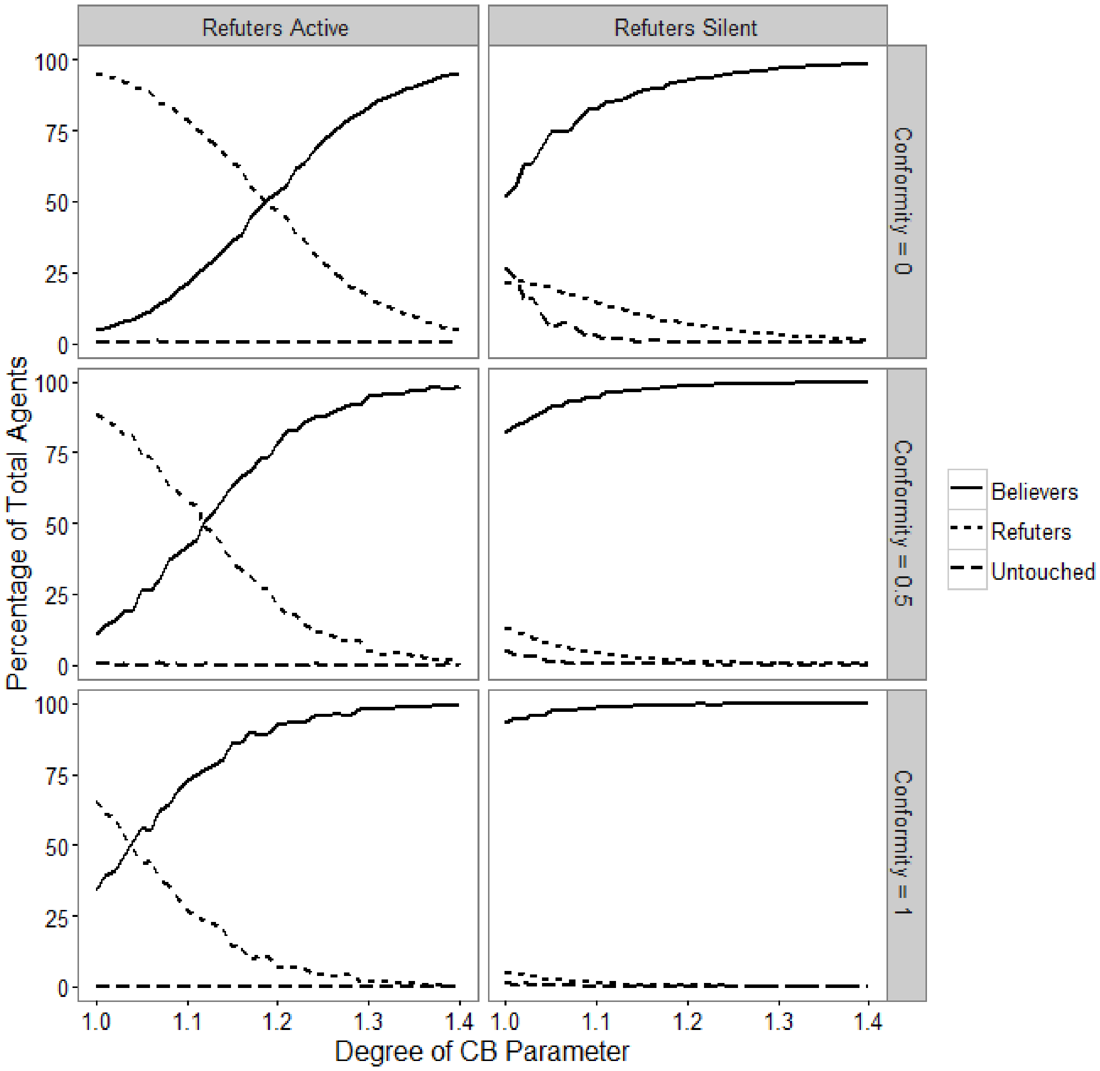


Figure 6.5.1. Model 2: Percentage of Agents at Simulation Termination Point, split by Belief-state. Values are averaged across 100 iterations.

Turning to the role of conformity, there is a clear impact of increasing conformity in the system on the percentage of believers. Specifically, as conformity increases, the necessary amount of confirmation bias for the majority in the system to be believers (i.e. meeting the majority threshold) decreases, as demonstrated by the leftward shifting cross-over point of believers and refuters in the active refuter (left-hand) column of Figure 6.5.1. Furthermore, this trend is increasingly severe in silent refuter systems (right-hand column). In particular, from no conformity (conformity = 0; top row of Figure 6.5.1) to only a mitigated level of conformity (conformity = 0.5;

middle row of Figure 6.5.1), the system requires no confirmation bias on the part of individual learners for the proportion to believers to become greater than 75%.

Further, the role of conformity is drawn into sharper contrast via tentative comparison to the 46% proportion of believers derived from an aggregation of individual belief recipients (taken from 4.1.1). To surpass this threshold, conformity is not required given the integrated level of the system when refuters are silent. However, when refuters are active, the degree of confirmation bias required within individual learners to surpass this threshold systematically decreases as conformity increases. Although tentative, taken with the overall findings, this indicates not only the potential dangers of conformity in interconnected systems, but the exacerbating impact conformity has when refutations are not made public.

Peak Propagation Rate. Turning to the peak rates of spread (whether refutation or belief communications), there is a selective impact of social conformity.

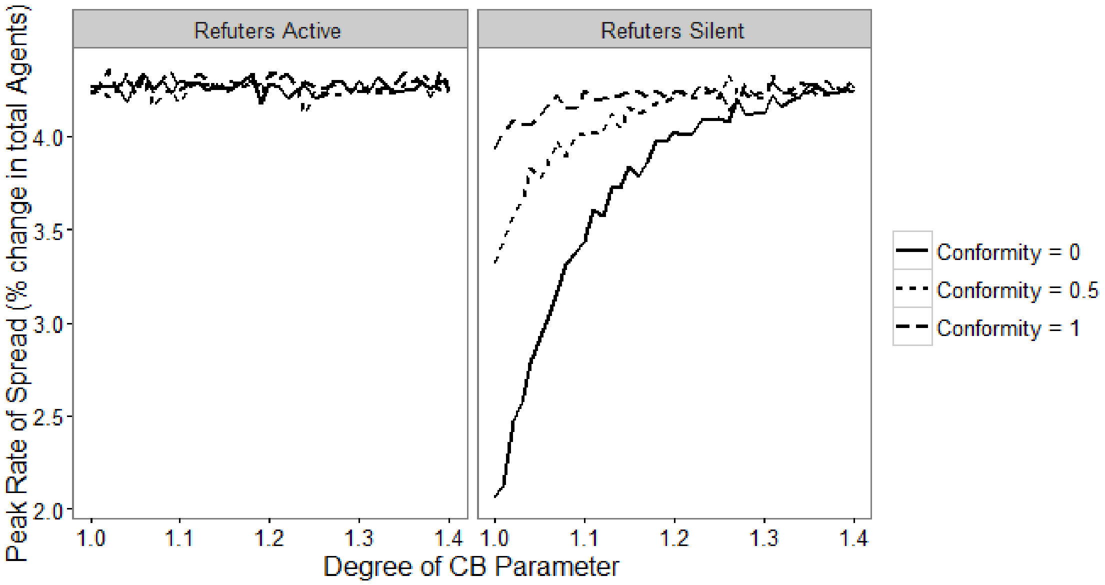


Figure 6.5.2. Model 2: Peak Rate of Spread across simulation. Averaged across 100 iterations.

As evidenced in the comparison between columns in Figure 6.5.2, rate of spread is once again *dependent* upon the confirmation bias parameter spread in refuter silent systems, whilst refuter active systems show no such dependency. However, social conformity also only affects the peak rate of spread in refuter silent systems, whereby more conforming systems propagate beliefs faster for any given degree of confirmation bias.

Degree of Clustering. Finally, turning to the degree of clustering in the system (defined as the percentage of linked neighbours of an agent with the same belief-state, averaged across all "activated" (those who have received any communication) agents), as shown in Figure 6.5.3, there are several effects of particular interest given the manipulation of social conformity across systems.

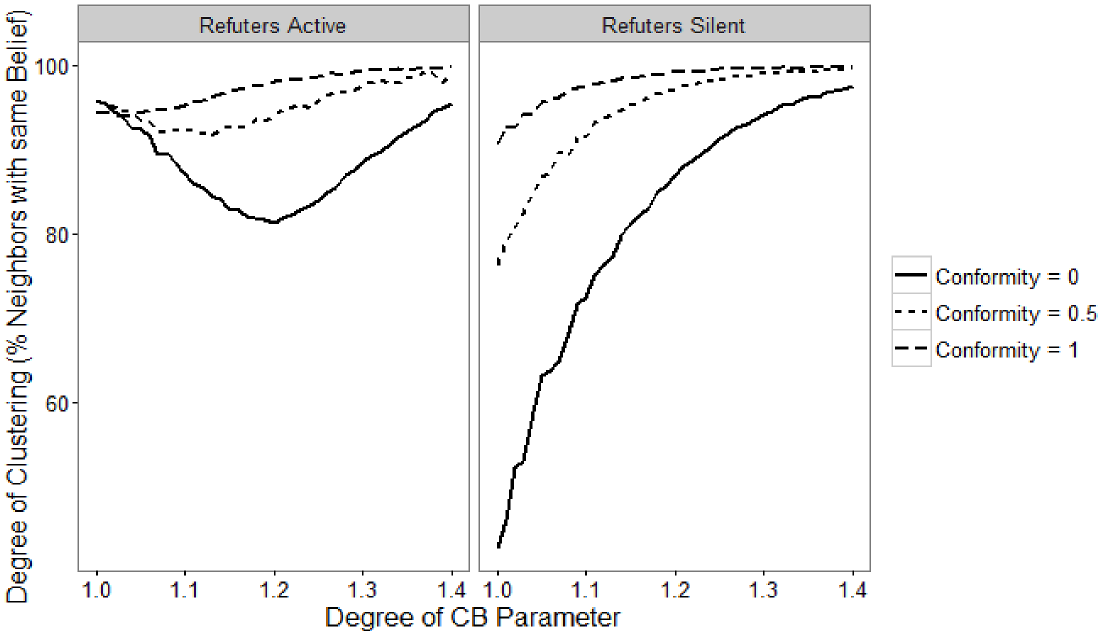


Figure 6.5.3. Model 2: Degree of Clustering, measured by average percentage of same-belief-state neighbours across all believing and refuting agents.

Overall, increasing conformity across systems increases the degree of clustering, in line with expected trends (6.5.1), homophily in networks (Dandekar et al., 2013) and opinion dynamics (Duggins, 2016). However, the form of this increase depends on the (in)action of refuters. When focusing on silent refuter systems (right-hand panel of Figure 6.5.3), increasing conformity increases the degree of clustering across all values of confirmation bias. Such an effect is a consequence of increasing conformity in these systems resulting in outright dominance of believers (right-hand column of Figure 6.5.1) regardless of confirmation bias parameter values. Turning to refuter active systems, there is a particular interest in the degree of clustering in a system when proportions (left-hand column of Figure 6.5.1) of believers and refuters are roughly equally represented. Specifically, when conformity is increased within a system the degree of clustering within each system at the equal proportion point becomes increasingly segregated. The fact that the most conforming line is almost flat indicates the changing proportions have almost no impact on clustering.

6.5.4 Discussion

The central objective of Model 2 has been to demonstrate the impact of introducing social conformity pressures to belief propagation processes across a highly interconnected system. Through a comparison to systems in which conformity plays no influence (conformity = 0 systems are functionally equivalent to Model 1, 7 link systems), it has been possible to demonstrate several important insights of conformity and its interplay with both the actions of refuters and levels of confirmation bias in the system.

Overall, increasing conformity within a system leads to increased proportions of believers within a system, and the consequent lower requirements for confirmation bias on the part of any given reasoner within the system to achieve a majority of believers, when compared to a model that does not incorporate conformity (6.4, and conformity = 0 systems in the present model). Breaking this down by refuter actions, this trend is exacerbated in silent refuter systems, to the point where confirmation bias is no longer required for global erroneous belief acquisition. Similarly, when turning to peak rates of spread across the system, conformity plays little role in the speed of information (belief or refutations) spread in refuter active systems, but through the aforementioned severity of conformity's impact on believer proportions, rapidly increases the rate of spread (in this case of almost exclusively erroneous beliefs) in refuter silent systems. Thus, not only are silent refuter systems more prone to (wide-spread) erroneous belief propagations as conformity increases, but such propagation occurs at faster and faster rates.

Finally, turning to the impact of conformity on the degree of clustering within systems, there is a clear impact in increasing clustering across both active and silent refuter systems, on any given confirmation bias parameter value. Of most interest, however, is the impact of conformity in refuter active systems, where conformity can lead to clustering approaching a ceiling, even when the number of believers and refuters in the system are equally prevalent. Such a demonstration, especially given that all agents have the same belief-neutral prior, illustrates the potency of localised influence within an interconnected network, and its cascading effects out across the system. This finding bears particular relevance to work on social network pruning (Ngampruetikorn & Stephens, 2015) and opinion dynamics (Dandekar et al., 2013; Deffuant, Neau, Amblard, & Weisbuch, 2000; Del Vicario et al., 2016; Duggins, 2016), which demonstrate agents with established beliefs / opinions are prone to enact behaviours that preserve said beliefs. Most notably it demonstrates, with minimal assumptions, that from base conditions of no prior beliefs across agents in a network, and a singular starting point of a new belief, segregation of belief states can occur prior to cyclical agent interactions. In other words, the inherent structure of a social network (or interconnected system), through the differential influence of proximal vs. distal agents on the network, interact with conforming pressures to create immediate and wide-spread segregation of belief-states.

6.6 General Discussion

Over the course of the present chapter, the aim has been to demonstrate, through the use of Agent-Based Models, how erroneous beliefs can be propagated across a connected network of individuals. Models were designed with individual learners represented by agents, who were ascribed general rules for receiving, evaluating, and propagating beliefs. These 1000 agents were randomly allocated across, and fixed to, a 2D space. Agents were then connected via links to manipulated (6.4) or fixed (6.5) numbers of neighbouring agents, forming an interconnected network. From a singular starting point, beliefs were then allowed to propagate through the network, with individual agents acting autonomously over time.

Through these models, the focus has been on demonstrating the interplay of structural (increasing interconnectivity between agents), social (refuter actions and social conformity) and cognitive (confirmation bias processes within evidence evaluation) factors on the success of a belief (measured by its proportion of believers throughout the system), the rate of belief spread, and the degree of clustering (taken as a measure of how segregated agents with opposing belief states are) within the system. Although the present work is exploratory in nature, we feel it adds insight through the proof of concept application of individual-based belief uptake and adherence findings to interactive, dynamic, societal level systems. Consequently, there are several, interrelated findings of interest that pertain to other, simulation-based literatures including information cascades (Cui, 2016; Martin, Hofman, Sharma, Anderson, & Watts, 2016) and opinion dynamics (Dandekar et al., 2013; Duggins, 2016; Ngampruetikorn & Stephens, 2015), as well as the real-world implications of an increasingly interconnected world, and the actions of individuals within it (e.g., the propagation of "fake news" across social media).

The first finding of interest is the impact of increasing interconnectivity on belief propagation. Unsurprisingly, as a system becomes more interconnected, the overall amount of communication across the system increases, and does so at faster and faster rates, and results in larger homogenous groups (i.e. clustering) of individuals sharing the same belief-state. However, the actions of refuters (those who have evaluated the communicated belief and believe it is false) in either remaining silent or actively communicated the refutation to neighbours plays a huge impact as interconnectivity increases.

In systems of low interconnectivity, the decision of refuters to stay silent, in the case of the correct evaluation of a false belief, does carry a benefit to the global population. This is because refuters in effect *cauterize* the potentially harmful belief. To elaborate further, in a low interconnected system, there are fewer paths for beliefs to be communicated across. As such, the decision of a refuter to remain silent in effect *blocks* that pathway to the remainder of the network. If instead the refuter passes on their refutation, it now opens up the possibility that individuals within that (otherwise untouched) branch of the network, having now been presented with the question of a potential belief being true or not, may evaluate the evidence and come to a different conclusion, despite the recommendation to refute. Thus, the global number of (harmful) believers ends up being higher in the latter condition. Conversely, as a system becomes more interconnected (e.g., via the rise of the internet and social media), the action of refuter silence becomes maladaptive. In particular, as demonstrated in Model 1, as a system becomes more interconnected, there are more potential pathways for beliefs to travel between agents, and thus the aforementioned cauterizing effect of refuter silence can be bypassed.

The importance of refuter action is illustrated further by the manipulation of social conformity within said interconnected systems. Although in refuter active systems, conformity leads to lower confirmation bias requirements on the part of each individual reasoner for a majority of the population to become believers, this consequence of conformity in silent refuter systems is far more extreme. In effect, a highly interconnected system with silent refuters is already, as mentioned above, highly susceptible to the spread of erroneous beliefs (with relatively low levels of confirmation bias required on the part of an individual reasoner for believers to quickly reach majority status). However, as conformity is introduced to such a system, not only do rates of spread and degree of clustering increase, any requirements for confirmation bias on the part of individual reasoners is in effect removed. Such a powerful effect of conformity in these systems can be attributed to both the difference in *perceived* conformity among an agent's neighbours, and the snowballing effect of one-sided communication in an interconnected, broadly conformist system. Further, the finding provides a concise demonstration of the subjectivity of human reasoning (i.e. a direct, motivational impact on the individual), exacerbated by man-made structures (i.e. online networks). A critical implication of this is the demonstrable importance of openly and rapidly communicating corrective assertions in the internet era.

Conformity within an interconnected network, even when refutations are communicated, along with decreasing the necessary degree of confirmation bias in individual reasoners for believers to reach a majority, has a critical impact on the degree of clustering when proportions of believers and refuters are at their most equal. To explain further, when conformity is not present in the system, the increased equality of the two belief-states results in lower levels of clustering (i.e. as both belief-states are more evenly represented, more individuals are neighbours with agents of the opposite belief-state). However, as conformity is introduced, the intermingling of agents of differing belief-states does not increase with the increasingly equal proportions of the two groups. Put another way, from a single starting point, a belief spread among initially neutral, equally rational agents, results in two almost entirely segregated communities of refuters and believers. Such an outcome is attributed to the way in which a network structure and social conformity interact, whereby the proximity dictated structure of “visible” neighbours (we tend to “friend” those we know / frequent sites we like or are familiar with) and “invisible” distal members of the network is exacerbated by the pressures to conform to the *perceived* majority. Associated work on opinion dynamics, has shown opinion polarization (Dandekar et al., 2013) and segregation effects (Ngampruetikorn & Stephens, 2015) are common consequences of interactions among groups of individuals over *repeated interactions*. This finding, that the *inherent structure* of modern interconnected networks, along with a straightforward instantiation of social conformity, leads to immediately segregated groups of differing belief states despite neutral starting positions, has damning implications for the resolution of belief discrepancies in the internet era.

The present work does carry several caveats given the abstractions and idealisations necessary to extrapolate from the target system to a tractable simulation. For example, the abstractions regarding the iterations of social conformity, as an example of the general decision making process regarding model architecture, are a necessary simplification of what is otherwise a complex phenomenon. Given the exploratory nature of the present work, so as to demonstrate a proof of concept, the present simplifications are in this case not only an opportunity for further (more complex) modelling, but warrant more in depth statistical treatment. Whether this is through the incorporation of more complex learning models, such as the use of a Bayesian alternative, or more pertinently, through the expansion of the model architecture to incorporate interesting psychological elements associated with the behaviour at hand, including source credibility and confidence components.

One such example of further work, to help bring further fidelity of the model to the target system, is to incorporate these source credibility elements. At present, the model assumes a flat hierarchy of influence (i.e. each agent is equally capable of influencing their equally plentiful neighbours). By adding social hierarchy, for example by drawing the number of links an agent has from a positive-tailed gamma distribution (resulting in a few, highly connected agents) and tying this to social conformity via Latané’s (1981) principles of social influence (i.e. the popularity, and thus visibility (represented as the number of links) of the agent across the network is also a marker of credibility), can add further insight to belief propagation effects. Such an instantiation could shed light on the role of internet-based news media (we frequent media sites that we value as a source of news) and modern social media hierarchies (we tend to follow those we value and perceive others as valuing).

Lastly, with regard to the incorporation of cognitive architecture into an ABM, an interesting and pertinent topic for the use of ABMs is worth highlighting. As demonstrated in the present work, ABMs as a methodological tool, allow for novel insights not readily available to Equation-Based Modelling (EBM) and experimentation (see Parunak, Savit, & Riolo, 1998, for a review of the differences between ABM and EBM). Namely, the in-depth simulation of multiple, contingent agents (and thus the capacity for *interaction*) over time, gleans insights into aggregate or emergent behaviours, and provides validation of implemented cognitive architecture on the part of the individual. In this way, ABMs provide a way of examining cognitive models, simulating experiments (and otherwise unethical interventions), testing evolutionary arguments, and optimising (through incremented simulation) behavioural interventions. Consequently, we argue ABMs can provide a useful complement to the cognitive psychologist’s toolbox.

CHAPTER 7: GENERAL DISCUSSION

The purpose of this thesis has been to investigate how communicated beliefs may be adopted and maintained, without long-term objective support from evidence. The material covered in the investigation of this topic can be addressed in two pairs of questions: Firstly, can communicated beliefs, by their mere presence, lead to integrative confirmation bias effects that maintain them, despite being erroneous? Secondly, and related to the first, when motivational and evidence exposure explanations (which are typically associated with belief-maintenance biases) are accounted for, do these effects remain? Thirdly, in the uptake of such beliefs, what role do early experiences play in the acceptance or rejection of communicated beliefs? And finally, is the role of early experiences in belief uptake a consequence of anonymous or unknown sources (i.e. given knowledge regarding the reliability of a source, early experiences no longer play a role in uptake)?

This thesis has been motivated by the surprising prevalence of not only erroneous, but often harmful beliefs[30]. In particular, it is driven by current ecological factors, namely the advent of mass communication technologies - most notably the internet - and the increased capacity to not only communicate beliefs, but search for evidence and (in)validate them. In particular, the quandary remains of how despite this availability of information, erroneous beliefs such as homeopathy, and climate change denial, among many others, are still prevalent (Vyse, 2013). The questions and general approach of this research has been an attempt to provide a more holistic account of erroneous belief uptake and maintenance processes. In doing so, it is argued that evidence structure (both in terms of order and diagnosticity) and source cues alone can

[30] For example, the United States has experienced resurgences in measles outbreaks, linked to the decision of parents not to vaccinate their children, as a consequence of the belief that such vaccines cause autism or are generally harmful. Source: The Centers for Disease Control and Prevention, http://www.cdc.gov/measles/cases-outbreaks.html.

result in confirmation biases in evidence integration, providing an explanation for the adoption and maintenance of pseudoscientific, superstitious, or generally misinformed beliefs, despite the availability of evidence.

The three empirical chapters (3-5) have sought to address these questions, each building off the last, and their results are summarised below.

Firstly, as evidenced in Chapter 3, through the use of a probabilistic reversal learning paradigm, it was possible to demonstrate an ongoing learning account of communicated belief biasing effects. Not only did this paradigm illustrate the impact of a novel, generic belief manipulation, but it was demonstrated that long-lasting, integrative biasing effects depended upon evidence order. Furthermore, these effects, both in choice and posterior judgement data, occurred in spite of the presence of counterfactual information, and despite accuracy motivations. Although this work provided an initial answer to questions regarding the mechanism behind belief maintenance, questions remained regarding the role of initial evidence in the uptake of beliefs. In light of this, a new paradigm was designed for Chapter 4, to explore the possible limits of initial evidence as a gatekeeper to belief uptake (using a more ecologically valid initial evidence manipulation to do so) and to answer questions regarding the impact of counterfactuals.

This paradigm folded the control (non-belief) condition within-subjects, improving the power and resolving potential selectivity issues. Further, this paradigm focused on more short-term fluctuations of evidence and their role in belief consolidation effects, and explored ways of disrupting such an effect via explicit intervention. Consequently, large main effects of both beliefs and initial evidence (primacy) were found in both choice and posterior measures. However, beliefs and initial evidence were also found to *interact*, demonstrating the aforementioned consolidation effect. Such an effect was linked to the validating role initial evidence played when a belief was communicated in the absence of cues (such as source credibility) that would otherwise indicate validity. It should be noted that although in comparison to the large main effects of beliefs and initial evidence, the interaction was comparatively modest; we argue that such consolidation effects may be exacerbated by the motivational and selective exposure elements this work had sought to reduce.

Furthermore, although counterfactual presence was found to facilitate learning (i.e. the presence of counterfactuals led to higher proportions of optimal choices over time), it did not impact on the consolidation effect, whilst explicit intervention did demonstrate some tentative impact on interrupting the consolidation effect, supporting an explicit truth value account of early belief-validation processes. Up until this point, the source of the belief had purposefully been kept neutral (via an absence of cues regarding credibility, including any indication of the belief originating from the experimenter) so as to focus on evidence factors in belief adoption and maintenance. As such, questions remained regarding whether the role of initial evidence in belief consolidation was due to the absence of other means for assessing the likely validity of a belief (such as the reliability of the source).

To address this, Chapter 5 introduced an explicit manipulation of the perceived trustworthiness and expertise of the source of the belief. A manipulation was built into the pre-existing paradigm from Chapter 4, in such a way that the interplay of source cues and initial evidence effects on belief uptake could be assessed. By manipulating these factors, it was not only demonstrated that communicated beliefs were interpreted in light of the source cues, but that when these cues suggested the belief was more credible (i.e. the source was perceived as highly trustworthy), then the role of initial evidence as a belief (and source) validator was subsumed. Conversely, when cues indicated the credibility of the source (and thus the belief) to be suspicious (i.e.

perceived to be untrustworthy), then initial evidence was once again critical to the consolidation of the suspected belief (i.e. the suspicion that the source was lying and thus the opposite was true), with such effects occurring *independently* of the actual truth of the belief itself. This suggests that evidence and sources were kept (in the mind of the reasoner) as independent. Further, it was tentatively demonstrated that high trust sources may profit (both in terms of the degree of belief uptake, and subsequent improved ratings of trust), irrespective of the truth of communication, if available evidence is ambiguous enough. Conversely, low trust individuals were damned (by the same metrics) regardless of whether they told the truth or not.

Lastly, these findings, as empirical evidence of the impact of communicated beliefs on evidence integration in the individual, were extrapolated to an Agent-Based Model of belief-propagation across networks in Chapter 6. This work sought to demonstrate the implications of the aforementioned individual processes as a network of individuals becomes more interconnected, as well as the impact of refuter actions (passing on refutation or remaining silent), increasing social conformity, and the degree of confirmation bias in the individual’s learning processes. This impact was assessed in terms of the “success” of the belief – or its degree of penetration across the network, the speed of spread, and degree of clustering. These models illustrated not only the growing importance of active refuters as a system becomes more interconnected, but that interconnectivity, in conjunction with factors such as social conformity, results in high levels of belief penetration, segregation among believers and refuters, and increasing rates of spread. Such effects emphasise the importance of this work as a whole, demonstrating the exaggerated impact of beliefs (even when evidence is available to test the belief) when moving from an individual to an interactive, heterogeneous, societal level.

From a theoretical standpoint, the present work moves away from previous characterisations of erroneous belief maintenance as due to individual differences in susceptibility based on demographics like gender and age (Bal et al., 2014; Dömötör, Ruíz-Barquín, & Szabo, 2016; Emme, 1941), or cultural specificity (Lepori, 2009; Tsang, 2004; Woo & Kwok, 1994), instead shifting towards work that has focused on erroneous belief as a natural by-product of the environment (Abbott & Sherratt, 2011; Foster & Kokko, 2009; Fudenberg & Levine, 2006) and learning errors (Beck & Forstmeier, 2007; Catania & Cutts, 1963; Matute, 1995; Ono, 1987; Rudski et al., 1999; Skinner, 1948). The present work has sought to take a considered view, demonstrating that it is a combination of the *context* (in terms of both evidence and source cues), *cognitive constraints,* and *internal motivations of the reasoner* that are pivotal to the adoption and maintenance of fallacious beliefs.

The latter of these bears particular relevance to the discussions of the rationality of such beliefs. More explicitly, an individual is not always motivated to be accurate (Kruglanski & Ajzen, 1983), but rather has what may broadly be termed as a directional motivations (Kunda, 1990). This latter category of motivations include (but are not limited to) social conforming pressures (Asch, 1955; Cialdini & Goldstein, 2004) and the maintenance of positive self-concept (Pyszczynski & Greenberg, 1987). Both of these can be thought of as influencing the utility of possible outcomes. For example, if presented with evidence that potentially undermines one's positive self-concept, the utility (or in this case, cost) of "adjusting down" to a more negative self-concept for the sake of accuracy is outweighed by the utility of concluding the evidence is invalid in some manner, and thus retaining a positive self-concept (such a mechanism has recently been reformulated in the Meaning Maintenance Model; see Heine et al., 2006; Proulx & Inzlicht, 2012). In this way, while arguments can be made regarding the (ir)rationality of such motivations (although perhaps with some degree of difficulty), this thesis suggests that nominally “fallacious” beliefs retained by individuals through a motivated reasoning process should not be dismissively termed irrational. Instead this work has sought to demonstrate that it is the contrivance of structural (e.g., evidence clarity), social (belief transference, source credibility elements), and cognitive (reasoning capacity, pattern match/mismatch asymmetries) circumstances that can mislead an otherwise adaptive knowledge acquisition process.

7.1 Conclusions

Over the course of the thesis, the central purpose has been to take a considered approach that focuses on the role of second-hand, communicated beliefs in biasing the interpretation of subsequent first-hand evidence. With reference to the reviewed literature looking at beliefs as propositional statements (Houwer, 2009; Mitchell et al., 2009), the disentanglement of possible motivational (Klein & Kunda, 1989; Kruglanski & Freund, 1983; Kruglanski & Ajzen, 1983; Kunda, 1990) and cognitive (Allahverdyan & Galstyan, 2014; Hahn & Harris, 2014; Jonas et al., 2001; Klayman, 1995; MacDougall, 1906; Nickerson, 1998; Nurek et al., 2014; Phillips & Edwards, 1966; Pohl, 2004; Yu & Lagnado, 2012) accounts of biasing effects, the empirical chapters that followed have sought to tie these elements together, providing novel insights into the question at hand, both individually and as a collective whole. Accordingly, there are a number of overarching conclusions that can be drawn from the present work.

Firstly, beliefs can be conceptualised as propositional statements[31] regarding (in this case) an action (“select option A”) and an outcome (“win”). This way of conceptualising a belief, as forwarded in the social cognition literature, has a readily applicable terminology associated with it. Specifically, a proposition, which can be thought of as an explicit statement regarding an association between (in this case) an

[31] The notion of beliefs as propositional statements has been discussed at length in the tradition of Analytics in Philosophy, see Martinich & Sosa, 2001, for a companion reference.

action and an outcome, implies a truth value (De Houwer, 2009). Given that a proposition is a statement *about* the world; such a statement can therefore be evaluated in terms of its likely truth (*its* truth being only a *representation* of the association's strength). Evidence for the conceptualisation of a belief as a propositional statement can be found most notably in the low trust conditions of the source credibility experiments (5.3.2 and 5.4.2). These conditions demonstrate that when presented with a belief from an untrustworthy individual, the belief is interpreted by participants in light of this cue (i.e. "the source is saying A is better, but I suspect them of lying, so perhaps B is in fact better") resulting in a reflection of the standard belief-consolidation effect (consolidation of the *suspicion*). Thus, the presented belief is interpreted by a recipient, in light of its context of communication (given the reliable source, the belief is interpreted as [H] / given the unreliable source, the belief is interpreted as [¬H]), as a working hypothesis, or recipe for potential action (i.e. information pertaining to an action-outcome relationship), otherwise known as a proposition. Further, such an interpretation demonstrates that participants can disentangle a belief from its source.

However, generic beliefs provide an additional challenge to the evaluation of a truth value. When faced not only with probabilistic, ambiguous evidence (which in itself can make belief verification difficult), but a belief that is generic (i.e. a broad, unquantified / non-statistical statement; Cimpian et al., 2010; Leslie, 2008), the additional ambiguity of interpretation (i.e. difficulty in falsification), makes accurate evaluation of a belief's truth value especially difficult to a (human) reasoner. This is because not only is the signal (or "true state" of the environment) difficult to detect, but the application of any given signal is also ambiguous (akin to "looking for a needle in a haystack", when you are not really sure what a needle looks like). Such situational factors are consequently influential when inferring systemic implications of the discussed effects, and further work is encouraged to explore possible boundary conditions. In particular, work investigating relevance and process accounts of situational (and potentially temporal) factors is of interest.

Further support for an account of communicated beliefs as generic, propositions possessing a truth value, and the implied difficulty of refutation, comes from the retention of erroneous beliefs in Chapters 4 and 5, given an insufficiently large refutation period (i.e. the large main effects of the belief). However, despite such large main effects, it was demonstrated that the truth value could still be evaluated against the objective distribution (and consequently the impact of belief on choices and judgements reduced), by the interaction with initial evidence (further evidenced in Chapter 3).

This leads to the second consequence regarding the conceptualisation and interpretation of a communicated belief: the role of consolidation. This can be thought of in terms of an order effect in the evaluation of the truth value (Mitchell et al., 2009; Schwartz, 1982), wherein early information plays a more critical role in the evaluation of truth than later evidence. The resurgence of contrast codes in the empirical work of Chapter 3, and the belief by initial evidence interaction term in later chapters suggest it is not simply the imposition of primacy onto the belief (i.e. purely *additive* effects), but an interaction that implies evaluation of belief *in light of* initial evidence. Such a finding is highly pertinent to the developing framework of propositional learning (Houwer, 2009; Mitchell et al., 2009), with particular regard to the recent focus on instruction effects (Roswarski & Proctor, 2003; Van Dessel et al., 2015; Van Dessel, De Houwer, et al., 2016), given the latter's emphasis on how instruction shapes evidence interpretation via goal redirection. However, to explore a more explicit exposition of this mechanism, it is useful to turn to more cognitive explanations of such a process.

Beginning with the identified importance of initial evidence, in the case of anonymously sourced beliefs, the lack of credibility cues is argued to lead to such a validating role. However, when credibility cues are available, such as in 5.3.2 and 5.4.2, where beliefs have originated from a high trust source, then the role of initial evidence in ratifying a belief is no longer needed. This finding does not readily fit with the predictions of the Elaboration Likelihood and the Heuristic-Systematic Models (Briñol & Petty, 2009; Chaiken & Maheswaran, 1994; Petty & Cacioppo, 1984), which suggest that the impact of source credibility (either as a cue, or heuristic) should be subsumed by exposure to (contrary) evidence and the capacity to evaluate it (i.e. motive and opportunity; both present in the current paradigm). However, one cannot fully rule out these models, given not only their use of strong or weak arguments (which possess truth values), rather than objective evidence evaluation, but the reliance on the manipulation of motive (high or low elaboration likelihood, or task importance) to delineate between processes. Put another way, proponents of such models could argue that any retained use of source information was due to the task importance (or what we have termed as accuracy incentives) not being high enough. These discrepancies are discussed further in regard to implications and further work.

The interplay of source credibility and objective evidence does, however, more readily fit within pre-existing models of source credibility in argumentation and persuasion (Bovens & Hartmann, 2003; Hahn et al., 2012; Harris et al., 2015; Madsen, 2016), where the interplay between an argument, and the expertise and trustworthiness of its source are used to explain phenomena such as appeals to authority (Harris et al., 2015). These accounts argue the (possible) usefulness of source cues when an individual does not have access to first-hand evidence (a common occurrence in everyday society). In particular, propositions from a source perceived to be reliable have been found to carry more weight with recipients, in line with Bayesian expectations. The present work has extended this behaviourally, using sequential evidence-integration paradigms, to demonstrate that such added "convincingness" can lead to greater degrees of evidence misinterpretation to support an erroneous proposition (although further work is of interest to formally model possible expected deviation as a function of source reliability). Further, a belief from a high trust source may be considered to result in the recipient having a strong enough degree of confidence in the belief (which can be thought of as a function of the narrowness – and thus height – of the distributions of the prior for the belief as well as its source). Put another way, a high trust source will represent advice to the best of her ability, which makes the statement easier to interpret (i.e. the interpreter does not need to consider complicated motives such as lying).

This Bayesian model has been extended from the Bayesian network approach (Pearl, 2000, 2014), which has met with success in areas of reasoning including causal learning (Holyoak & Cheng, 2011; Lagnado & Sloman, 2004; Lagnado et al., 2007; Waldmann & Hagmayer, 2001). In reconciling such (typically) normative accounts of updating with the results found in the present work, two effects are of interest. The first is the aforementioned role of initial evidence in validating beliefs from sources that *lack* credibility cues. What is, in essence, an order effect, does not fit within a standard Bayesian model wherein evidence should carry equal weighting in the updating of a posterior, but models that incorporate additional "nodes" regarding the source provide a possible solution. If one assumes that the absence of source cues results in a (relatively) neutral / uninformed prior for a source's reliability, and a belief (or hypothesis) is communicated about an environment in which the recipient carries no priors, then the prior for the belief is only slightly informed (i.e. there is not much weight surrounding P(H)). In such a situation, initial evidence acts to update both the probability of the hypothesis being true, *and* the probability of the source being credible. In doing so, to take the example of refutation, initial evidence refuting the belief both shifts P(H) and updates the reliability of the source as increasingly *unreliable.*

If one considers each piece of evidence as an independent testimony (trials were explicitly noted as being independent from one another), where the perceived reliability is represented by the perceived diagnostic capability of a given trial, exposure to repeated (positive) evidence may lead to a coherence effect in updating P(H) and P(R|H). Such an effect would be attenuated by the perceived (un)likelihood of the evidence occurring (Bovens & Hartmann, 2003) independently (i.e. if perceived as unlikely, yet occurs anyway, a larger coherence effect occurs). However, this mechanism requires further work and formalisation, as the problem still rests at the formalisation of *order* in the process. Put another way, low confidence in both belief and source (i.e. wide confidence intervals on both probabilities) results in initial evidence having a disproportionate role in *shaping* these distributions (supported by the dismissal of beliefs pre-reversal in Chapter 3, and interaction of belief and initial evidence in Chapter 4). Thus, consolidation can be thought of in terms of reaching a degree of confidence in the belief, via narrowing the confidence interval. Such a potential mechanism to incorporate order effects is ripe for further research when validating beliefs (and sources) against sequential, bipartisan (i.e. lacking implied source elements) data.

In a similar manner, initial evidence was also found to be critical in the low trust conditions of Chapter 5 experiments (5.3.2 and 5.4.2). These conditions, in which sources were perceived to be low in trust, similarly relied upon initial evidence to confirm suspicions. Distinct from experiments in which the source trustworthiness, and as a consequence, reliability, is *unknown*, the low trust sources are indicative of being *unreliable*, and thus provoke active scepticism (Priester & Petty, 1995; Schul & Peri, 2015). In this way, the potential dishonesty of the source does not have the equivalent, but inverse, impact of a trustworthy source (i.e. recipients are not equally as convinced of / confident in the falsity of a belief from an untrustworthy source as they are the truth of a belief from a trustworthy source). Put another way, when cues indicate the source of a belief is untrustworthy; one becomes *suspicious* as a result (which one could consider an inversion of the truth-function). This does not however imply equivalent confidence in the opposite of the belief being true. As such, initial evidence plays a gatekeeping role in consolidating the *suspected* belief ("I was told A is better, but given the source is untrustworthy, I suspect that B is in fact better") – and in doing so updates the suspicion that the source is unreliable.

Conversely, when initial evidence is incoherent with the suspicion of falsehood indicated by the source's supposed untrustworthiness, then no such consolidation occurs (and thus no effect of belief is found in either (P(H) or P(¬H)) direction). This results in the suspicion being abandoned, and behaviourally, choice and judgement data reveals equivalent patterns to a context in which no belief was received (see 5.3.2 and 5.4.2). The role of initial evidence, in validating (or invalidating) beliefs from either unknown or suspicious (i.e. untrustworthy) bears a parallel to work in legal reasoning (Fenton, Neil, & Lagnado, 2013; Lagnado & Harvey, 2008), which has similarly forwarded a coherence-based model when explaining the power of the order of discrediting evidence on associated judgements of guilt in mock jury decisions. This proposal highlights the importance of eliciting confidence judgements regarding not only the probability of the belief being true, but extending to the reliability of the source in providing good and true information. In this way, a proxy can be gathered for the initial width of such probability distributions (i.e. formalising suspicion and lack thereof in terms of confidence), and the sequential impact of each (mis)matching piece of independent evidence on respective confidence nodes can be monitored (i.e. investigating mechanistic explanations of the aforementioned order effect on belief and reliability confidence).

The asymmetrical impact of initial evidence in the consolidation or refutation of a communicated belief (i.e. less evidence was required for consolidation than refutation, as indicated in Experiments in 3.2 to 3.5), bears a close parallel to the forwarded mechanism behind the belief *maintenance* process. Taking into account both cognitive and motivational elements, the present work has sought to forward an integrative account of confirmation bias in the evaluation of evidence to maintain beliefs. In accounting for possible directional motivational explanations (Kunda, 1990), communicated beliefs were incidentally communicated (in appearance, from other site users), resulting in the preclusion of experimenter demand effects (Hertwig & Ortmann, 2008) and affiliated suspicions (or increased confidence) regarding the purpose of the belief. Specifically, the motivation to adhere to a communicated belief that has been communicated seemingly by the experimenter (i.e. a position of authority, with likely knowledge regarding the task greater than the participants own) may muddle (and even dominate) more fundamental integrative processes, and remains an interesting avenue for further research. This speculation finds support from the source credibility experiment (5.3.2 and 5.4.2) conditions in which high trust sources dominate initial evidence consolidation and belief maintenance effects. Further, the dominance of credible sources fits tentatively with directional predictions from work in appeals to authority (Goodwin, 2011; Harris et al., 2015; Walton, 1997), in that information from sources deemed credible is more convincing to recipients (precluding the opportunity to evaluate). These models are, however, a-motivational (and instead forward a rational learner account), and future work would benefit from integrating Bayesian updating within an attentional and (directed) motivational framework.

In a similar manner, the use of performance (rather than belief maintenance) based incentives assisted in reducing directional motivation-based explanations for belief maintenance. Notably, this reduction occurred due to the presence of the belief being *incidental*, rather than either instructed as being pivotal to the task (i.e. participants were not *directed* to use or ignore the belief) or related to phenomena in which participants may have pre-existing motivations, whether based on prior opinions (Asch, 1955; DeMarzo, Vayanos, & Zwiebel, 2003; Festinger & Carlsmith, 1959) or being linked to internal consistency motivations, including the maintenance of a positive self-concept (Cialdini & Goldstein, 2004; Festinger, 1962; Heine et al., 2006; Proulx & Inzlicht, 2012). In other words, the experimental architecture allows for the assertion that biasing processes to maintain a fallacious belief occur *despite motivations to be accurate*.

In forwarding an integrative account of confirmation bias in the maintenance of erroneous beliefs, it is necessary to rule out purely selective evidence (or *input*) based explanations (Klayman, 1995; Klayman & Ha, 1987; MacDougall, 1906; Nickerson, 1998). By focusing on an integrative (or second order), explanation in this way, it is possible to discern whether there is a fundamental asymmetry in the evaluation (i.e. updating value) of confirmatory versus contradictory evidence. Given the existence of this asymmetry, one can forward that more selective, first order, accounts may (when evidence selection is in control of the reasoner) be hierarchically dependent upon the more fundamental integrative valuation difference (Klayman, 1995; MacDougall, 1906). Consequently, if an experiment allows for selective exposure explanations, it becomes immediately difficult to discern whether bias is due to integrative, and or selective processes. Thus, in early experiments (3.2 to 3.5) counterfactuals were present as a way of preventing selectivity in exposure (i.e. when making a choice between two options, participants were still shown the forgone option, preventing exposure to only selected cases). These experiments did find evidence for confirmation bias, despite the inability to selectively expose oneself to confirmatory-expected outcomes, thus suggesting an integrative explanation. However, despite attention to reversals indicating that participants were using counterfactual information, selective attention explanations could not be fully precluded (3.6).

To further clarify the distinction between first and second order accounts, counterfactual presence was then explicitly manipulated in experiments 4.2 to 4.4. This revealed, through a joint Bayesian analysis of the potential impact of counterfactuals, that although the presence of counterfactuals generally facilitated learning, the fundamental biasing interaction remained independent (supported by strong evidence for the null interaction with counterfactuals). Put another way, one would expect an interaction of the facilitated-learning effect of counterfactuals (counterfactual presence improves the proportion of optimal choices across blocks) with belief (and/or belief-initial evidence interactions). Specifically, control disease choices should have been affected by counterfactuals being present or absent (which they were), whilst belief disease choices – if attention selectivity is occurring – should have been relatively unaffected by the presence or absence of counterfactuals, as counterfactuals would be equivalently ignored in either condition. However, this was shown to not be the case, as belief disease choices were as equally affected by the presence or absence of counterfactuals as control disease choices (i.e. both showed the same facilitated learning effect). Alternatively, if selective attention was occurring in *both* control and belief disease choices, then one would not find the facilitated-learning effect through counterfactual inclusion (see 4.4).

To preclude primacy (Hogarth & Einhorn, 1992; Peterson & DuCharme, 1967) and conservatism (Edwards, 1961; Phillips & Edwards, 1966; Pitz, 1969; Pitz et al., 1967) explanations of the deviation caused by beliefs, in favour of an ongoing, integrative confirmation bias account (Gilovich, 1983; Klayman, 1995; MacDougall, 1906; Nickerson, 1998), two key effects can be highlighted as supportive evidence. The contrast code (those who are in the Belief Initially Supported (BIS) group are significantly different from both controls and Belief Initially Undermined (BIU) groups, which are equivalent) *reappears* post-reversal, despite groups having *converged* prior to reversal in experiments 3.2 to 3.5. This indicates that not only is learning ongoing, but that groups systematically diverge as a consequence of the previously consolidated belief. This implies that a consolidated belief can be maintained despite active engagement with contrary evidence. Further, the rapid abandonment of the belief in the BIU group indicates that the belief itself is not causing inertia in subsequent choices. In this way, there is sensitivity to initial evidence, followed by underweighting of contradictory evidence, which is not easily accommodated by a pure conservatism explanation. As such, we have forwarded the notion of an order effect (consolidation), dependent on confidence (which could instead be increased via source reliability cues), which establishes the "pattern", and an integrative confirmation bias mechanism that then maintains it. Although the present work makes comparisons between beliefs and either between- (Chapter 3) or within-subject (Chapters 4 & 5) control groups for determining belief-based deviations, further work is suggested below regarding model comparisons.

The forwarded integrative account of confirmation bias is leant recent neuroscientific support through pattern-match / mismatch updating asymmetries (Whitman et al., 2015). Specifically, when evidence is encountered that fits with the leading hypothesis (i.e. pattern), the updating signal is significantly stronger than when evidence does not fit the leading hypothesis. To think of this in terms of neural architecture, there is a neural representation of the proposed hypothesis, which is then strengthened by hypothesis-matching evidence. In contrast, there is no such neural architecture to update in the case of the *null*. Thus, the present work seeks to fit this forwarded explanation behind overweighted confirmatory evidence on updating at an integrative level. Tentatively, one can think of the aforementioned consolidation effect as the *establishment* of this pattern.

In summary, over the course of this thesis, evidence has been presented for an integrative confirmation bias account of belief maintenance (see Chapter 3). However, in the absence of cues indicating the credibility of a belief's source, such maintenance is dependent upon initial evidence in first consolidating the belief (Chapters 3 & 4). Conversely, when cues indicate a source as credible (Chapter 5), the consolidating role of initial evidence is subsumed (i.e. the belief is already considered valid). These latter consolidation effects have been related to the role of confidence in belief (and reliability) updating. Further, the present work has demonstrated not only that beliefs and sources may be considered independently of one another (see 5.5), but that there are promising avenues of research in integrating order effects, source credibility, and systematic updating deviations. Finally, through the use of Agent-Based Models, it has been possible to demonstrate that the effects of communicated erroneous beliefs, ascertained from empirical work using aggregated, isolated individuals, are likely exacerbated in the interactive structures of modern social networks.

7.2 Limitations

There are several limitations of importance when considering the work presented in this thesis. Although the preceding chapters (3 through 6) have highlighted limitations pertaining to the specific pieces of empirical or modelling work, this section serves to highlight a few of the broader, thesis level limitations of the work. Where appropriate, further work is proposed to resolve outstanding issues.

Firstly, it is necessary to address the lack of lab-based experiments. For this, it is worth re-iterating some of the reasoning for such a decision, as forwarded in the empirical sections of the thesis. The use of online experiments hosted on MTurk have been shown to not only replicate effects found under laboratory conditions (Buhrmester et al., 2011; Paolacci et al., 2010), but consistently demonstrate better demographic representation relative to typical, lab-based psychological experiments (Henrich et al., 2010). Although it has been noted that online studies do lack the experimental control of lab-based equivalents (Goodman et al., 2013), unless the experiment in question relies primarily on reaction time data (in which case internet-based programs can prove too noisy; for a review, see Gould, Cox, Brumby, & Wiseman, 2015), such an absence of control can be compensated for by larger sample sizes. In line with this solution, where possible power calculations were performed to assess the minimum sample size needed to detect effects within each online paradigm. Further, the online context not only allowed for the aforementioned motivational control (i.e. the removal of experimenter demand effects; Hertwig & Ortmann, 2008), but further provided a real world parallel to the ever-expanding internet based context of communication in the modern world. In this way, the artificiality of the laboratory setting is abandoned in favour of a more ecologically valid, methodologically and theoretically convenient alternative.

A second limitation is the difficulty in ruling out an attention selectivity (i.e. an input based bias) explanation. Although the present work, through probabilistic reversals and the manipulation of counterfactual presence, has sought to rule out such an explanation with some success, the possibility cannot be fully ruled out. An eye-tracking study is proposed as a potential solution to this, as the degree to which participants are attending to stimuli (i.e. both selected and counterfactual feedback) can be explicitly measured as a marker of attention. This solution does however pose its own difficulties, given the aforementioned issues with the artificiality of the laboratory setting.

The third and final limitation of the present work is the absence of model comparisons within the empirical chapters. The approach of the thesis has been instead to focus on differences in choices and judgements due to a communicated belief, relative to a control (belief-absent) baseline. Although we feel this approach carries merit for demonstrating the impact of these beliefs, and the factors at play in their uptake and maintenance, more in depth comparisons of the proposed deviations in comparison to normative model baselines (such as a standard Bayesian model) are of interest to provide greater insight into the proposed integrative bias account. Although the present work has sought to take a holistic approach, in which cognitive *and* motivational factors are considered, formal demonstration of the deviation from normative expectancies would lend additional credibility to the arguments set out, and is planned for further work. In particular, the promising incorporation of the Bayesian source credibility model (Hahn et al., 2009, 2012; Harris et al., 2015) to inform the predictions of Chapter 5 suggests an interesting avenue of model application to understanding deviations from normative predictions for both the consolidation and maintenance mechanisms currently proposed. However, it should be noted that future examination of the assumptions, methodology and predictions of various models (whether Bayesian (i.e. coherence based), or dual process accounts such as ELM, Petty & Cacioppo, 1984, and HSM, Chaiken & Maheswaran, 1994) are also needed to ensure tests properly and fairly subscribe to the predictions made by the models in contention.

7.3 Implications & Further Work

The findings put forward in the present work have implications for areas of research that involve the interaction between communicated information and evidence interpretation, such as impression formation and placebo effect. Further work is also suggested to advance the integration of the findings with present theories of source credibility, including the manipulation of motives and belief content. Finally, we touch upon the implications for real-world scenarios that involve erroneous belief acquisition and maintenance, including social networks. We now explore these implications and further work recommendations.

There are several areas of research that may benefit from work detailing the interplay between source credibility, communicated beliefs and order effects. Prior work on action-outcome learning, and more specifically the rise of propositional learning (De Houwer, 2009; Mitchell et al., 2009), has started to investigate the potential power of what those within the social cognition field term "instruction effects" (Van Dessel et al., 2015; Van Dessel, De Houwer, et al., 2016). This area of research investigates the way in which instruction (i.e. communication from another individual, typically the experimenter) influences the goals of the recipient. Thus, the "success" of the instruction is defined as the degree to which behaviour is shifted, in line with the predicted shift in goals. In this way, the factors that have been brought to bear in the present work, may also serve to impact this instructional efficacy. For example, if early experiences are manipulated to be congruent with the instruction-inferred goal state, does this facilitate a long-term shift in the participant's goals, in light of the "consolidation" of the instruction-inferred goal?

Linking this to phenomena such as placebo effects (Jensen et al., 2012; Schwarz & Büchel, 2015; Stewart-Williams & Podd, 2004; Wager & Atlas, 2015; Yarritu et al., 2015) and impression formation (Mann & Ferguson, 2015; Smith & Collins, 2009), these areas of research, which have yet to explore the interplay between evidence and communicated beliefs, could benefit from a more integrative approach that incorporates the cognitive elements outlined in the current work. For example, within placebo research, efforts have focused on determining effective additive factors, known as "cues to efficacy", which cover what can be thought of as source credibility elements (i.e. cues that facilitate the perception of the medical practitioner prescribing the placebo as authoritative / trustworthy; for a review see Wager & Atlas, 2015). It is briefly worth noting that such an impact of an authority in itself is readily reconcilable with current models of the rationality of following expert advice (Harris et al., 2015). Further cues to efficacy include the degree of perceived intervention (i.e. how much is "being done" medically), experience of expected side effects (known as active placebos, which mimic expected side effects of the medically active drug, such as using caffeine in a placebo to mimic elevated heart rates), and other content cues (e.g., the name of the "drug", the size and shape of the pill, and the number of required doses). However, despite the detailing of these potential contributing factors to rates of placebo responses, this area of research could benefit from a more integrated, social and cognitive approach which looks at the potential interactions between the belief formation factors and evidence-based factors, including order effects. Specifically, questions such as the potential impact of short-term, immediate perceptions of efficacy on belief adherence (and thus not only rates of placebo uptake, but placebo longevity) are of interest.

Regarding the role of belief content, interesting questions remain regarding its impact on the way in which evidence is then interpreted (and consequently, how the belief is either maintained or rejected). The present work has focused on beliefs as generic statements (Leslie, 2008; Prasada, 2000; Prasada et al., 2013) which provide, in this case, directional information (i.e. a hypothesis) as a prescription for action. The purpose of this implementation has been to draw a closer parallel to the informal form of successful pseudoscientific and superstitious beliefs communications (Gilovich, 1993), in that quantified parameters (which may allow for easier refutation) are missing, unlike in statistical statements (Leslie, 2008). In this way, beliefs as generic statements tend to benefit from openness to interpretation (Cimpian et al., 2010; Leslie et al., 2011). For example, while "A is better than B" provides a prescription for action, it does not provide any quantified information, making the interpretation more flexible (i.e. the belief could be interpreted as "*Eventually*, A is better than B", or "A is *only just* better than B", or "A is *sometimes* better than B"). Whereas a quantified alternative is more inflexible, providing an increasing number of potential conditions for refutation (and thus making the evaluation of the belief's truth value easier). To use a somewhat extreme, but illustrative example; "A wins *every time*.", similarly provides the prescription for action, but unlike a generic counterpart, provides a clear refutation criterion – if A *loses* once, the belief is false. For example, when considering the implications for belief encoding and its knock on impact to source evaluation, are statistical beliefs encoded in an equivalent way to more generic versions (e.g., "The person said X, but looking at the evidence I think they meant Y"), whereby the flexibility of interpretation relies primarily on the recipient, rather than the content? Alternatively, does the provided "falsification criteria" of the statistical belief result in penalisation not only to the updated validity of the belief (i.e. the belief is considered false, whilst its generic counterpart is still entertained as valid), but to the source of the belief as a consequence? In line with the earlier limitations raised regarding the situational versus systemic claims of the present work, the clarity and diagnosticity of evidence, in conjunction with belief content, is ripe for further exploration in determining boundary conditions for belief-favouring misinterpretation.

In a similar manner to the manipulation of belief content, directional and accuracy motivations (Kunda, 1990), which have been acknowledged but kept constant in the present work, are an intriguing avenue of research. The approach taken in the current work has been to rule out directional motivations behind belief adherence (through incidentally communicating a belief that is not integral to extraneous motivations of the recipient, such as opinions), and instead incentivising an accuracy motivation (through performance-based pay). The purpose of this has been to focus on a cognitive, second order bias account (Hahn & Harris, 2014; Klayman, 1995; MacDougall, 1906). More specifically, the inclusion of directional motivations is proposed to increase the proclivity for using sub-optimal cognitive strategies (Hahn & Harris, 2014; Kruglanski & Ajzen, 1983), such as adopting positive test strategies (Navarro & Perfors, 2011; Nickerson, 1998). Thus, in line with the purpose of stripping back possible biasing explanations, directional motivations were necessarily removed, as failing to do so would have obscured the possible mechanisms behind the belief adherence effect. However, further research into the explicit manipulation of these motivations, whether via direct incentive (Kruglanski & Freund, 1983) or public versus private exposure to decision outcomes / reasoning (Tetlock, 1983a, 1983b, 1985a), which have been shown to reduce primacy effects in impression formation when implemented (Kunda, 1990; Tetlock, 1983b), may be of interest in exploring incremental increases in sub-optimal cognitive strategy deployment.

Such a manipulation of motive is of further interest given the overruling impact of initial evidence, and the general impact of source trustworthiness on choices and judgements in Chapter 5. The interaction of source, implied motive and initial evidence elements is of interest to other areas of research that deal with the impact of sources on advice efficacy (Harvey & Fischer, 1997; Twyman et al., 2008), perception of risk (Siegrist et al., 2000, 2005), positive emotions and behaviours in social psychology (Cuddy et al., 2011; Kenworthy & Tausch, 2008; Tausch et al., 2007), persuasion (Briñol & Petty, 2009; Priester & Petty, 1995; Wood, 2000), and argumentation (Hahn et al., 2009, 2012; Harris et al., 2015; Walton, 1997). For example, the demonstrated relationship in the present work between the manipulation of source credibility, and the subsequent impact on evidence evaluation in itself already poses interesting challenges to models of source credibility (Chaiken & Maheswaran, 1994; Harris et al., 2015; Madsen, 2016; Petty & Cacioppo, 1984). However, of interest is the way in which motivation may impact this relationship. For example, when outcomes matter more, there may be increased reliance on trust and expertise factors (akin to work in advice taking on the impact of increased uptake of advice when forecasts carry more critical implications; Harvey & Fischer, 1997), leading to a stronger misleading influence of source credibility (which would contradict the predictions of dual-process accounts). Alternatively, such motivations may instead lead to a decrease in the misleading influence of source and/or initial evidence factors. Put another way, does increasing the utility attached to being *right* as a belief recipient lead to increased reliance on – or critical assessment of – both the belief itself, and as a consequence, the reliability of a source? Further research is required to synthesise, for example, the growing literature on Bayesian models of source credibility in the behavioural domain, with the deviations in learning exhibited as a consequence of source influences, and with motivational accounts more generally. Such an avenue of research is of interest to investigations into partisanship, internal versus externally declared motivations behind opinion development, and polarization effects alluded to in Chapter 6's Agent-Based models.

Further, as mentioned in the discussion of consolidation (7.1), questions remain regarding the process by which the belief is initially consolidated, leading to the consequent unleashing of the biasing effect. As has previously been mentioned, one possibility is a representation that focuses on an "informed prior" for a Bayesian learning process (Pearl, 2000, 2014). In such an instantiation, the belief is represented by a probability distribution, and this can be linked to a probability of the source being reliable, as forwarded by Bayesian models of source credibility (Hahn et al., 2012; Harris et al., 2015; Madsen, 2016). In instances where information is absent or uncertain regarding the source of a belief, and thus the belief's validity is similarly uncertain, initial evidence has a disproportionate impact on updating both elements (such as the experiments of Chapters 3 & 4), in that initially wide distributions are rapidly sharpened, increasing confidence. However, in cases where cues to credibility, such as trustworthiness (Schul & Peri, 2015b; Siegrist, Cvetkovich, & Roth, 2000; Twyman et al., 2008) and expertise (Goodwin, 2011; Harris et al., 2015; Sniezek & Van Swol, 2001; Walton, 1997) are present prior to evidence exposure (Chapter 5), these cues already act as evidence to inform this process (supplanting the role of initial evidence). Such a strategy may be argued as both reasonable and rational (a point that contradicts dual process accounts such as ELM and HSM). Thus, a more in depth investigation into the interplay of order effects, biased integration and the updating of belief and (in conjunction) source reliability elements, is likely to add insight to descriptive models of belief evaluation (i.e. uptake and maintenance) processes. One recommendation for further work is to elicit confidence in credibility nodes (source trustworthiness and expertise) during behavioural tasks, exploring the impact of evidence increments in (disproportionately) increasing confidence.

Returning to the implications for other areas of research, causal learning (Dennis & Ahn, 2001; Fernbach & Sloman, 2009; Holyoak & Cheng, 2011; Lagnado & Harvey, 2008), consumer research (Ha & Hoch, 1989; Hoch & Ha, 1986), decision making (Nurek et al., 2014; Schöbel et al., 2016), and illusions of control and causality (Langer, 1975; Rudski, 2004; Yarritu & Matute, 2015; Yarritu et al., 2013), can similarly benefit from incorporating elements of the presented findings. Although these areas have already well-established effects such as primacy (Anderson, 1965; Curley et al., 1988; Dennis & Ahn, 2001; Mantonakis et al., 2009; Peterson & DuCharme, 1967) and conservatism (Phillips & Edwards, 1966; Pitz, 1969; Pitz et al., 1967; Winkler & Murphy, 1973), typically such effects are associated with hypothesis *formation* processes (Hogarth & Einhorn, 1992). The present work seeks to explore the role of first-hand evidence order in the *evaluation* of a communicated hypothesis, integrating the necessary elements of source cues and belief content influences.

Moving to applied implications of the presented research, there are many instances in which individuals receive and maintain beliefs that are not supported by empirical evidence. Among these are some of the most critical issues facing humanity today. For example, the belief that climate change is a hoax[32], vaccines are unsafe and/or cause autism[33], and various forms of systematic misinformation – such as in the recent political campaign for Leaving the EU in the United Kingdom[34], are all instances in which beliefs are adhered to despite evidence that refutes them. Such beliefs typically involve the capacity to selectively search for evidence (Jonas et al., 2001; Lord et al., 1979), and directional motivations that interfere with an objective desire for accuracy (Hahn & Harris, 2014; Klein & Kunda, 1989; Kunda, 1990). However, the present work has demonstrated that erroneous belief maintenance can occur in the absence of such alternative explanations. Accordingly, the indicated sensitivity to initial evidence as a gatekeeper to this maintenance should also be taken into account when considering possible solutions.

More critically still, as evidenced in the Agent-Based models of Chapter 6, in a world of increasing interconnectivity, such erroneous beliefs can spread not only more successfully, but faster. Further, with simple assumptions regarding the social conformity elements inherent to social interactions (of which social interactions on a network, e.g., posting a status on a social media site, are not immune from; Duggins, 2016), individual reasoning processes are compromised (to the extent that an erroneous belief no longer requires any confirmation bias elements for successful adoption / maintenance), and polarization across the network is *immediately* extremely high (> 95%). To expound further, from a system of naïve agents, with the same learning

[32] Recent polling in the US indicates that despite an 8-year high from extensive public campaigns regarding the effects of climate change, and the past 5-year trend of consistently increasing warmer weather, 10% of Americans do not believe climate change to be real. Gallup poll data taken from: http://www.gallup.com/poll/190010/concern-global-warming-eight-year-high.aspx

[33] Recent polling from the US indicates 9% of Americans believe vaccines to be unsafe, with a further 7% admitting to "not knowing". Taken from a PewResearch poll: http://www.people-press.org/2015/02/09/83-percent-say-measles-vaccine-is-safe-for-healthy-children/

[34] For a summary list of evaluated claims, see: http://www.telegraph.co.uk/news/2016/06/22/eu-referendum-fact-checking-the-big-claims1/

apparatus, and neutral priors, the initial stochastic process of the first few “infected” individuals rapidly propagates into highly segregated groups *before* any cyclical interactions occur. Put in conjunction with work on social network pruning (Cui, 2016), which has found segregation tends to increase over time through severance of connections with those of differing opinions, this appears to paint the possibility of successfully refuting such erroneous beliefs in a somewhat depressing light. There is, however, a parallel recommendation to the potential early intervention point for avoiding erroneous belief uptake on an individual level (4.3) when dealing with increasing interconnected networks. The actions of refuters, where once inaction serves to cauterize the spread of belief in more sparsely connected systems, should instead be replaced with action as networks have become more interconnected. Given that the nature of a highly interconnected system results in eventual complete penetration of belief, it becomes logically imperative to communicate refutations to prevent a form of network based selective exposure. Further, this provides an opportunity to test how interconnected systems may be designed more efficiently to minimise belief solidification.

Accordingly, this thesis has aimed to shed light on the base structural (evidence) conditions and cognitive processes in erroneous belief acquisition and maintenance. This, in combination with the consideration of source cues, belief content, and motivational influences, is forwarded as the first steps in an integrative ecological model of belief uptake and maintenance.

REFERENCES

Abbott, K. R., & Sherratt, T. N. (2011). The evolution of superstition through optimal use of incomplete information. *Animal Behaviour*, *82*(1), 85–92. http://doi.org/10.1016/j.anbehav.2011.04.002

Addario, M. D., & Macchi, L. (2012). Pseudodiagnosticity : The Role of the Rarity Factor in the Perception of the Informativeness of Data, *3*(6), 489–493.

Aggarwal, M., Hyland, B. I., & Wickens, J. R. (2012). Neural control of dopamine neurotransmission: implications for reinforcement learning. *The European Journal of Neuroscience*, *35*(7), 1115–23. http://doi.org/10.1111/j.1460-9568.2012.08055.x

Allahverdyan, A. E., & Galstyan, A. (2014). Opinion dynamics with confirmation bias. *PloS One*, *9*(7), e99557. http://doi.org/10.1371/journal.pone.0099557

Anderson, N. H. (1965). Primacy Effects in Personality Impression Formation Using a Generalized Order Effect Paradigm. *Journal of Personality and Social Psychology*, *34*(1), 1–9. http://doi.org/10.1037/h0021966

Aronson, E. (1968). Dissonance theory: Progress and problems. In R. P. Abelson (Ed.), *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Arrowood, A. J., & Ross, L. (1966). Anticipated effort and subjective probability. *Journal of Personality and Social Psychology*, *4*(1), 57–64. http://doi.org/10.1037/h0023441

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, *193*(5), 31–35. http://doi.org/10.1038/scientificamerican1155-31

Axelrod, R. (1997). The Dissemination of Culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, *41*(2), 203–226.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? *Memory & Cognition, 32*(8), 1369–78.

Baeyens, F., Eelen, P., Crombez, G., & De Houwer, J. (2001). On the Role of Beliefs in Observational Flavor Conditioning. *Current Psychology, 20*(2), 183–203.

Baeyens, F., Vansteenwegen, D., De Houwer, J., & Crombez, G. (1996). Observational conditioning of food valence in humans. *Appetite*, *27*(3), 235–250. http://doi.org/10.1006/appe.1996.0049

Bal, B. S., Singh, D., Badwal, K. K., & Dhaliwal, G. S. (2014). Superstitions Behavior and Decision Making in Collegiate Athletes: An Illogical Phenomenon. *Advances in Physical Education, 4*(1), 1–5. http://doi.org/10.4236/ape.2014.41001

Balodis, I. M., MacDonald, T. K., & Olmstead, M. C. (2006). Instructional cues modify performance on the Iowa Gambling Task. *Brain and Cognition*, *60*(2), 109–17. http://doi.org/10.1016/j.bandc.2005.05.007

Bandura, A. (1977). *Social Learning Theory*. *General Learning*. New York, NY: General Learning Press.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, *7*(6), 525–553. http://doi.org/10.1016/j.psychsport.2006.03.001

Bar-Hillel, M., & Wagenaar, W. a. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454. http://doi.org/10.1016/0196-8858(91)90029-I

Barron, G., & Leider, S. (2010). The role of experience in the gambler's fallacy. *Journal of Behavioral Decision Making*, *129*(October 2009), 117–129. http://doi.org/10.1002/bdm

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *The Academy of Management Review*, *3*, 439–449. http://doi.org/10.1177/0022022108314549

Beck, J., & Forstmeier, W. (2007). Superstition and belief as inevitable by-products of an adaptive learning strategy. *Human Nature*, *18*(1), 35–46. http://doi.org/10.1007/BF02820845

Becker, D., & van der Pligt, J. (2015). Forcing your luck: Goal-striving behavior in chance situations. *Motivation and Emotion*, *40*, 1–9. http://doi.org/10.1007/s11031-015-9527-5

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–21. http://doi.org/10.1038/nn1954

Bell, A. (1994). Media (mis)communcation on the science of climate change. *Public Understanding of Science*, *3*, 259–275.

Benjamin, D. J., & Raymond, C. (2012). A Model of Non-Belief in the Law of Large Numbers, 1–106.

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science*, *33*(2), 206–242. http://doi.org/10.1111/j.1551-6709.2009.01010.x

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. *Cognition*, *108*(3), 652–61. http://doi.org/10.1016/j.cognition.2008.05.005

Billman, D., Bornstein, B., & Richards, J. (1992). Effects of expectancy on assessing covariation in data: “Prior belief” versus “meaning.” *Organizational Behavior and*

Human Decision Processes, *53*(1), 74–88. http://doi.org/10.1016/0749-5978(92)90055-C

Blanchard, S. J., Carlson, K. a, & Meloy, M. G. (2014). Biased predecisional processing of leading and nonleading alternatives. *Psychological Science*, *25*(3), 812–6. http://doi.org/10.1177/0956797613512663

Blanchart, E., Marilleau, N., Chotte, J. L., Drogoul, A., Perrier, E., & Cambier, C. (2009). SWORM: An agent-based model to simulate the effect of earthworms on soil structure. *European Journal of Soil Science*, *60*(1), 13–21. http://doi.org/10.1111/j.1365-2389.2008.01091.x

Blanco, F. (2017). Positive and negative implications of the causal illusion. *Consciousness and Cognition*, *50*, 56–68. http://doi.org/10.1016/j.concog.2016.08.012

Blanco, F., Barberia, I., & Matute, H. (2014). The lack of side effects of an ineffective treatment facilitates the development of a belief in its effectiveness. *PloS One*, *9*(1). http://doi.org/10.1371/journal.pone.0084084

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(90003), 7280–7287. http://doi.org/10.1073/pnas.082080899

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. http://doi.org/10.1016/j.obhdp.2006.07.001

Boureau, Y.-L., & Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology : Official*

Publication of the American College of Neuropsychopharmacology, *36*(1), 74–97. http://doi.org/10.1038/npp.2010.151

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Bradbury, J. W., & Vehrencamp, S. L. (1998). Principles of Animal Communication. *Animal Behaviour*. http://doi.org/10.1016/j.anbehav.2011.12.014

Brandone, A. C., Gelman, S. A., & Hedglen, J. (2015). Children's developing intuitions about the truth conditions and implications of novel generics vs. quantified statements. *Cognitive Science*, *39*(4), 711–738. http://doi.org/10.1007/978-1-4614-5915-6

Breen, R. (1999). Beliefs, rational choice and Bayesian learning. *Rationality and Society*, *11*, 463–479.

Briggs, P., Burford, B., De Angeli, a., & Lynch, P. (2002). Trust in Online Advice. *Social Science Computer Review*, *20*(3), 321–332. http://doi.org/10.1177/089443930202000309

Briñol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology, 20*(1), 49–96. http://doi.org/10.1080/10463280802643640

Bröder, A., & Schiffer, S. (2003). Bayesian Strategy Assessment in Multi-attribute Decision Making. *Journal of Behavioral Decision Making*, *16*(3), 193–213. http://doi.org/10.1002/bdm.442

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Burke, C. J., & Tobler, P. N. (2011). Coding of reward probability and risk by single neurons in animals. *Frontiers in Neuroscience*, *5*(October), 121. http://doi.org/10.3389/fnins.2011.00121

Calvert, R. L. (1985). The Value of Biased Information : A Rational Choice Model of Political Advice. *The Journal of Politics*, *47*(2), 530–555.

Canic, E. (2014). Serial-position effects in preference construction : a sensitivity analysis of the pairwise-competition model, *5*(August), 1–5. http://doi.org/10.3389/fpsyg.2014.00902

Carley, K. (1991). A Theory of Group Stability. *American Sociological Review*, *56*(3), 331–354.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical Physics of Social Dynamics. *Reviews of Modern Physics*, *81*(2), 591. http://doi.org/10.1103/PhysRevLett.93.098103

Catania, a C., & Cutts, D. (1963). Experimental control of superstitious responding inhumans. *Journal of the Experimental Analysis of Behavior*, *6*(2), 203–8. http://doi.org/10.1901/jeab.1963.6-203

Chaiken, S., & Maheswaran, D. (1994). Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Chambers, C. G., Graham, S. A., & Turner, J. N. (2008). When hearsay trumps evidence: How generic language guides preschoolers' inferences about unfamiliar things. *Language and Cognitive Processes*, *23*(5), 749–766. http://doi.org/10.1080/01690960701786111

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*(2), 262–275. http://doi.org/10.1037//0022-3514.71.2.262

Choi, J. A., Koo, M., Choi, I., & Auh, S. (2008). Need for Cognitive Closure and Information Search Strategy. *Psychology & Marketing*, *25*(11), 1027–1042. http://doi.org/10.1002/mar

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105–109. http://doi.org/10.1111/1467-8721.01242

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. http://doi.org/10.1146/annurev.psych.55.090902.142015

Cimpian, A., Brandone, A. C., & Gelman, S. a. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, *34*(8), 1452–1482. http://doi.org/10.1111/j.1551-6709.2010.01126.x

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*, 155–159.

Colagiuri, B., Livesey, E. J., & Harris, J. a. (2011). Can expectancies produce placebo effects for implicit learning? *Psychonomic Bulletin & Review*, *18*(2), 399–405. http://doi.org/10.3758/s13423-010-0041-1

Collins, E. C., Percy, E. J., Smith, E. R., & Kruschke, J. K. (2011). Integrating advice and experience: learning and decision making with social and nonsocial cues. *Journal of Personality and Social Psychology*, *100*(6), 967–982. http://doi.org/10.1037/a0022982

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *The Journal of Applied Psychology*, *92*(4), 909–927. http://doi.org/10.1037/0021-9010.92.4.909

Cone, J., & Ferguson, M. J. (2015). He Did What?: The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. http://doi.org/10.1007/s12671-013-0269-8.Moving

Cooper, J., & Fazio, R. H. (1984). A New Look at Dissonance Theory. *Advances in Experimental Social Psychology*, *17*, 229–266. http://doi.org/10.1016/S0065-2601(08)60121-5

Cornelius, C. A. (2015). *The Effects of Control and Uncertainty on Children's Supernatural Beliefs. (Doctoral dissertation).*

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. http://doi.org/10.1016/j.riob.2011.10.004

Cui, B. (2016). *From Information Cascade to Knowledge Transfer : Predictive Analyses on Social Networks*.

Curley, S. P., Yates, J. F., & Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes, 38*, 230–256.

Curley, S. P., Young, M. J., Kingry, M. J., & Yates, J. F. (1988). Primacy effects in clinical judgments of contingency. *Med Decis Making*, *8*(3), 216–222. http://doi.org/10.1177/0272989X8800800310

Custers, R., & Aarts, H. (2010). The unconscious will: how the pursuit of goals operates outside of conscious awareness. *Science (New York, N.Y.)*, *329*(5987), 47–50.

http://doi.org/10.1126/science.1188595

Damisch, L., Stoberock, B., & Mussweiler, T. (2010). Keep your fingers crossed!: how superstition improves performance. *Psychological Science*, *21*(7), 1014–20. http://doi.org/10.1177/0956797610372631

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5791–6. http://doi.org/10.1073/pnas.1217220110

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. http://doi.org/10.1037/0022-3514.44.1.20

Dave, C., & Wolfe, K. (2003). On confirmation bias and deviations from bayesian updating. *Internet Access: Http://www. Peel. Pitt. edu/esa2003/papers/wolfe_confirmationbias. pdf.[Prieiga per Internetą 2011 02 24]*.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior : A Psychonomic Society Publication*, *37*, 1–20. http://doi.org/10.3758/LB.37.1.1

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. http://doi.org/10.1111/spc3.12111

Decker, J. H., Lourenco, F. S., Doll, B. B., & Hartley, C. A. (2015). Experiential reward learning outweighs instruction prior to adulthood. *Cognitive, Affective & Behavioral Neuroscience*, *15*(2), 310–320. http://doi.org/10.3758/s13415-014-

0332-5

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98. http://doi.org/10.1142/S0219525900000078

DeGroot, M. H., & DeGroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, *69*(345), 118–121. http://doi.org/10.2307/2285509

DeKay, M. L., Miller, S. A., Schley, D. R., & Erford, B. M. (2014). Proleader and antitrailer information distortion and their effects on choice and postchoice memory. *Organizational Behavior and Human Decision Processes*, *125*(2), 134–150.

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). Modeling confirmation bias and polarization, 1–12. Retrieved from http://arxiv.org/abs/1607.00022

DeMarzo, P., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 909–968.

Dennis, M. J., & Ahn, W.-K. K. (2001). Primacy in causal strength judgments: the effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*, *29*, 152–164. http://doi.org/10.3758/BF03195749

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. http://doi.org/10.3389/fpsyg.2014.00781

Ditto, P. H., Jemmott III, J. B., & Darley, J. M. (1988). Appraising the threat of illness: a mental representational approach. *Health Psychology*, *7*(2), 183–201. http://doi.org/10.1037/0278-6133.7.2.183

Doherty, M. E., & Mynatt, C. R. (1990). Inattention to P (H) and to P (D|¬H): A converging operation. *Acta Psychologica*, *75*, 1–11.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111–121.

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, *31*(16), 6188–6198. http://doi.org/10.1523/JNEUROSCI.6486-10.2011

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. http://doi.org/10.1016/j.brainres.2009.07.007

Dömötör, Z., Ruíz-Barquín, R., & Szabo, A. (2016). Superstitious behavior in sport: A literature review. *Scandinavian Journal of Psychology*, *57*(4), 368–382. http://doi.org/10.1111/sjop.12301

Donald, M. (1993). Précis of Origins of the modern mind: Three stages in the evolution of culture and cognition. *Behavioral and Brain Sciences*, *16*, 737–791.

Duggins, P. (2016). A Psychologically-Motivated Model of Opinion Change with Applications to American Politics. *ArXiv*.

Dylko, I. B., Beam, M. A., Landreville, K. D., & Geidner, N. (2012). Filtering 2008 US presidential election news on YouTube by elites and nonelites: An examination of the democratizing potential of the internet. *New Media & Society*, *14*(5), 832–849. http://doi.org/10.1177/1461444811428899

Earle, T. C., Siegrist, M., & Gutscher, H. (2010). Trust, risk perception and the TCC model of cooperation. In *Trust in risk management: Uncertainty and scepticism in* *the public mind* (pp. 1–50). http://doi.org/10.4324/9781849773461

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, *62*(4), 385–94.

Emme, E. (1941). Supplementary study of superstitious belief among college students. *The Journal of Psychology: Interdisciplinary and Applied*, *12*(2), 183–184.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. http://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic : Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. http://doi.org/10.1111/j.1467-9280.2006.01704.x

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modelling*. Princeton University Press.

Erdelyi, M. H. (1974). A new look at the new look: perceptual defense and vigilance. *Psychological Review*, *81*(1), 1–25. http://doi.org/10.1037/h0035852

Ernst, E. (2002). A systematic review of systematic reviews of homeopathy. *British Journal of Clinical Pharmacology*, *54*, 577–582.

Ernst, E. (2007). Adverse effects of spinal manipulation: a systematic review. *J R Soc Med*, *100*(7), 330–338. http://doi.org/10.1258/jrsm.100.7.330

Evans, D. (2004). *Placebo: Mind over matter in modern medicine*. Oxford University

Press.

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, *460*(August), 685--686. http://doi.org/10.1038/460685a

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–60. http://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G * Power 3 : A flexible statistical power analysis program for the social , behavioral , and biomedical sciences. *Behaviour Research Methods*, *39*(2), 175–191.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102. http://doi.org/10.1111/cogs.12004

Fernbach, P. M., & Sloman, S. a. (2009). Causal learning with local computations. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(3), 678–93. http://doi.org/10.1037/a0014928

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. http://doi.org/10.1177/001872675400700202

Festinger, L. (1962). *A theory of cognitive dissonance*. Stanford University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Psychology*, *58*(2), 203–210. http://doi.org/10.1037/h0041593

Fiedler, K., & Freytag, P. (2004). Pseudocontingencies. *Journal of Personality and Social Psychology*, *87*(4), 453.

Fiedler, K., & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(4), 883–903. http://doi.org/10.1037/0278-7393.32.4.883

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 349–358.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239–260. http://doi.org/10.1037//0033-295X.90.3.239

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. http://doi.org/10.1016/j.tics.2006.11.005

Foster, K. R., & Kokko, H. (2009). The evolution of superstitious and superstition-like behaviour. *Proceedings. Biological Sciences / The Royal Society*, *276*(1654), 31–7. http://doi.org/10.1098/rspb.2008.0981

Francke, W., & Dettner, K. (2005). Chemical signalling in beetles. *Topics in Current Chemistry*, *240*, 85–166. http://doi.org/10.1007/b98316

Freund, T., Kruglanski, A. W., & Shpitzajzen, A. (1985). The Freezing and Unfreezing of Impressional Primacy: Effects of the Need for Structure and the Fear of Invalidity. *Personality & Social Psychology Bulletin*, *11*, 479–487.

Frias-Martinez, E., Williamson, G., & Frias-Martinez, V. (2011). An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information. *3rd International Conference on Social Computing (SocialCom'11)*, 49–56. http://doi.org/10.1109/PASSAT/SocialCom.2011.142

Frisch, K. von. (1950). *Bees: Their Vision, Chemical Senses, and Language.* Cornell University Press.

Frost, P., Casey, B., Griffin, K., Raymundo, L., Farrell, C., & Carrigan, R. (2015). The influence of confirmation bias on memory and source monitoring. *The Journal of General Psychology*, *142*(4), 238–252. http://doi.org/10.1080/00221309.2015.1084987

Fudenberg, D., & Levine, D. K. (2006). Superstition and Rational Learning. *American Economic Review*, *96*(3), 630–651. http://doi.org/10.1257/aer.96.3.630

Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*(5), 800–815.

Gaertner, S. L., & Dovidio, J. F. (2014). *Reducing intergroup bias: The common ingroup identity model*. Psychology Press.

Garcia-retamero, R., Hoffrage, U., Müller, S. M., & Maldonado, A. (2010). The influence of causal knowledge in two-alternative forced-choice tasks. *Open Psychology Journal*, *3*, 136–144. http://doi.org/10.2174/1874350101003020136

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. http://doi.org/10.1016/j.lmot.2013.03.003

Gaudreau, P., Blondin, J.-P., & Lapierre, a.-M. (2002). Athletes' coping during a competition: relationship of coping strategies with positive affect, negative affect, and performance–goal discrepancy. *Psychology of Sport and Exercise*, *3*(2), 125–150. http://doi.org/10.1016/S1469-0292(01)00015-2

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of*

Experimental Social Psychology, 40(4), 535–542. http://doi.org/10.1016/j.jesp.2003.10.005

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*(4), 650–669. http://doi.org/10.1192/bjpo.bp.115.000224

Gilbert, N. (2008). *Agent-Based Models*. SAGE Publications.

Gilbert, N., Hawksworth, J. C., & Swinney, P. A. (2009). An Agent-Based Model of the English Housing Market. *Artificial Intelligence*, 30–35.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*(6), 1110–1126.

Gilovich, T. (1993). *How we know what isn't so*. New York, NY: The Free Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295–314.

Gimblett, R. H. (2002). *Integrating geographic information systems and agent-based modelling techniques for simulating social and ecological processes*. Oxford University Press.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making, 26*(3), 213–224. http://doi.org/10.1002/bdm.1753

Goodwin, J. (2011). Accounting for the Appeal to the Authority of Experts. *Argumentation, 25*(3), 285–296. http://doi.org/10.1007/s10503-011-9219-6

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. E. M. (2015). Home is Where the Lab is: A Comparison of Online and Lab Data From a Time-sensitive

Study of Interruption. *Human Computation*, 45–67. http://doi.org/10.15346/hc.v2i1.4

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of “us and them” in 80 lines of code: modeling group genesis in homogeneous populations. *Psychological Science*, *25*(4), 982–90. http://doi.org/10.1177/0956797614521816

Greenbaum, C. W., & Zemach, M. (1972). Role-playing and change of Attitude toward the Police after a Campus Riot: Effects of Situational Demand and Justification. *Human Relations*, *25,* 87–99.

Greenwald, A. G., & Ronis, D. L. (1978). Twenty Years of Cognitive Dissonance: Case Study of the Evolution of a Theory. *Psychological Review*, *85*(1), 53–57. http://doi.org/10.1037/0033-295X.85.1.53

Ha, Y., & Hoch, S. (1989). Ambiguity, processing strategy, and advertising-evidence interactions. *Journal of Consumer Research*, *16*(3), 354–360.

Hahn, U., & Harris, A. J. L. (2014). *What Does It Mean to be Biased. Motivated Reasoning and Rationality. Psychology of Learning and Motivation - Advances in Research and Theory* (1st ed., Vol. 61). Elsevier Inc. http://doi.org/10.1016/B978-0-12-800283-4.00002-2

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

Hahn, U., Oaksford, M., & Harris, A. J. L. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian Argumentation* (pp. 15–38). http://doi.org/10.1007/978-94-007-5357-0

Hammerl, M., & Grabitz, H.-J. (2000). Affective-Evaluative Learning in Humans: A

Form of Associative Learning or Only an Artifact? *Learning and Motivation*, *31*(4), 345–363. http://doi.org/10.1006/lmot.2000.1059

Harmon-Jones, E., Peterson, H., & Vaughn, K. (2003). The dissonance-inducing effects of an inconsistency between experienced empathy and knowledge of past failures to help: Support for the action-based model of Dissonance. *Basic and Applied Social …*, *25*, 69–78.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: incorporating the role of coherence. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(5), 1366–1373. http://doi.org/10.1037/a0016567

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *39*(7), 1–38. http://doi.org/10.1111/cogs.12276

Harris, A. J. L., & Osman, M. (2012). The illusion of control: A Bayesian perspective. *Synthese*, pp. 1–10. http://doi.org/10.1007/s11229-012-0090-2

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin*, *135*(4), 555–588.

Hartmann, A., D’Ettorre, P., Jones, G. R., & Heinze, J. (2005). Fertility signaling - The proximate mechanism of worker policing in a clonal ant. *Naturwissenschaften*, *92*(6), 282–286. http://doi.org/10.1007/s00114-005-0625-1

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. http://doi.org/10.1006/obhd.1997.2697

Hegselmann, R., & Krause, U. (2002). Opinion Dynamics and Bounded Confidence. *Simulation*, *5*(3), 2. http://doi.org/citeulike-article-id:613092

Heider, F. (1958). The psychology of interpersonal relations. *The Journal of Marketting*, *56*, 322.

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: on the coherence of social motivations. *Personality and Social Psychology Review*, *10*(2), 88–110. http://doi.org/10.1207/s15327957pspr1002_1

Heller, D., Komar, J., & Lee, W. B. (2007). The dynamics of personality states, goals, and well-being. *Personality & Social Psychology Bulletin*, *33*(6), 898–910. http://doi.org/10.1177/0146167207301010

Hendrickson, A. T., Perfors, A. F., & Navarro, D. J. (2014). Adaptive information source selection during hypothesis testing. *36th Annual Meeting of the Cognitive Science Society (CogSci 2014).*

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61-83-135. http://doi.org/10.1017/S0140525X0999152X

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decision From Experience and the Effect of Rare Events in Risky Choice. *American Psychological Society*, *15*(8), 534–539.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: a methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383-403-451. http://doi.org/10.1037/e683322011-032

Hertwig, R., & Ortmann, A. (2008). Deception in Experiments: Revisiting the Arguments in Its Defense. *Ethics & Behavior*, *18*(1), 59–92.

http://doi.org/10.1080/10508420701712990

Higgins, E. T., & King, G. (1981). Accessibility of social constructs: Information-processing consequences of individual and contextual variability. *Personality, Cognition, and Social Interaction*, *69*, 121.

Hoch, S., & Ha, Y. (1986). Consumer learning: Advertising and the ambiguity of product experience. *Journal of Consumer Research*, *13*(2), 221–233.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55. http://doi.org/10.1016/0010-0285(92)90002-J

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–63. http://doi.org/10.1146/annurev.psych.121208.131634

Hommel, B. (1993). Inverting the Simon effect by intention: Determinants of direction and extent of effects of irrelevant spatial information. *Psychological Research*, *55*, 270–279.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2002). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(5), 849–878. http://doi.org/10.1017/S0140525X01000103

Houwer, J. De, Thomas, S., Baeyens, F., De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*. http://doi.org/http://psycnet.apa.org/doi/10.1037/0033-2909.127.6.853

Irwin, F. W. (1953). Stated Expectations as Functions of Probability and Desirability of Outcomes. *Journal of Personality*, *21*(3), 329–335. http://doi.org/10.1111/j.1467-

6494.1953.tb01775.x

Jager, W., & Amblard, F. (2005). Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, *10*(4), 295–303. http://doi.org/10.1007/s10588-005-6282-2

James, W. (1890). *The principles of psychology*. Read Books Ltd.

Janssen, M. A., & Ostrom, E. (2006). Empirically Based , Agent-based models. *Ecology And Society*, *11*(2), 37.

Jarvstad, A., & Hahn, U. (2011). Source Reliability and the Conjunction Fallacy. *Cognitive Science*, *35*(4), 682–711. http://doi.org/10.1111/j.1551-6709.2011.01170.x

JASP Team. (2016). JASP (Version 0.8.0.0).

Jensen, K. B., Kaptchuk, T. J., Kirsch, I., Raicek, J., Lindstrom, K. M., Berna, C., … Kong, J. (2012). Nonconscious activation of placebo and nocebo pain responses. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(39), 15959–64. http://doi.org/10.1073/pnas.1202056109

Jessup, R. K., & O’Doherty, J. P. (2011). Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the gambler’s fallacy. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(17), 6296–304. http://doi.org/10.1523/JNEUROSCI.6421-10.2011

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *J. Pers. Soc. Psychol.*, *80*(4), 557–571. http://doi.org/10.1037/0022-3514.80.4.557

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*, 59–99.

Joukhador, J., Blaszczynski, A., & Maccallum, F. (2004). Superstitious beliefs in gambling among problem and non-problem gamblers: preliminary data. *Journal of Gambling Studies*, *20*(2), 171–80. http://doi.org/10.1023/B:JOGS.0000022308.27774.2b

Kahn, R., & Kellner, D. (2004). New Media and Internet Activism: From the "Battle of Seattle" to Blogging. *New Media & Society*, *6*, 87–95. http://doi.org/10.1177/1461444804039908

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. http://doi.org/10.1016/0010-0285(72)90016-3

Kassajian, H. H., & Cohen, J. B. (1965). Cognitive Dissonance and Consumer Behavior. *California Management Review*, *8*, 55–64.

Keinan, G. (1994). Effects of stress and tolerance of ambiguity on magical thinking. *Journal of Personality and Social Psychology*, *67*(1), 48–55. http://doi.org/10.1037/0022-3514.67.1.48

Keinan, G. (2002). The Effects of Stress and Desire for Control on Superstitious Behavior. *Personality and Social Psychology Bulletin*, *28*(1), 102–108. http://doi.org/10.1177/0146167202281009

Kelman, H. C. (1958). Compliance, identification, and internalization three processes of attitude change. *Journal of Conflict Resolution*, *2*(1), 51–60. http://doi.org/10.1177/002200275800200106

Kenworthy, J. B., & Tausch, N. (2008). Expectations about the accuracy and stability of warmth versus competence traits: An intergroup analysis. *European Journal of Social Psychology*, *38*, 1121–1129. http://doi.org/10.1002/ejsp

King, L. A., Hicks, J. A., & Abdelkhalik, J. (2009). Death, life, scarcity, and value: an alternative perspective on the meaning of death. *Psychological Science*, *20*(12), 1459–62. http://doi.org/10.1111/j.1467-9280.2009.02466.x

Klayman, J. (1984). Learning from feedback in probabilistic environments. *Acta Psychologica*, *56*, 81–92.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 317–330.

Klayman, J. (1995). Varieties of Confirmation Bias, *32*, 385–418. http://doi.org/10.1016/S0079-7421(08)60315-1

Klayman, J., & Ha, Y. (1987). Confirmation , Disconfirmation , and Information in Hypothesis Testing. *Psychological Review*, *94*(2), 211–228. http://doi.org/10.1037/0033-295X.94.2.211

Klein, W. M., & Kunda, Z. (1989). Motivated person perception: Justifying desired conclusions. In *meeting of the Eastern Psychological Association*. Boston.

Kruglanski, A. W. (1980). Lay epistemo-logic--process and contents: Another look at attribution theory. *Psychological Review*, *87*(1), 70–87. http://doi.org/10.1037/0033-295X.87.1.70

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19*(5), 448–468. http://doi.org/10.1016/0022-1031(83)90022-7

Kruglanski, A. W., & Klar, Y. (1987). A view from a bridge: Synthesizing the consistency and attribution paradigms from a lay epistemic perspective. *European Journal of Social Psychology*, *17*, 211–241. http://doi.org/10.1177/1368431007084369

Kruglanski, a W., & Ajzen, I. (1983). Bias and Error in Human Judgement. *European Journal of Social Psychology, 13*(August 2015), 1–44. http://doi.org/http://dx.doi.org/10.1002/ejsp.2420130102

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, *53*(4), 636–647. http://doi.org/10.1037/0022-3514.53.4.636

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18*(2), 195–224. http://doi.org/10.1016/0010-0285(86)90012-5

Kusec, A., Tallon, K., & Koerner, N. (2016). Intolerance of uncertainty , causal uncertainty , causal importance , self-concept clarity and their relations to generalized anxiety disorder. *Cognitive Behaviour Therapy*. http://doi.org/10.1080/16506073.2016.1171391

Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review*, *15*(6), 1166–1173. http://doi.org/10.3758/PBR.15.6.1166

Lagnado, D. A., & Sloman, S. (2004). The Advantage of Timely Intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 856–876. http://doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In *Causal learning: Psychology, philosophy, and computation* (pp. 154–172).

Langer, E. J. (1975). Illusion of Control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. http://doi.org/10.1027/1618-3169/a000225

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356. http://doi.org/10.1037/0003-066X.36.4.343

Lee, M., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lee, P. M. (1989). *Bayesian Statistics: An introduciton*. John Wiley & Sons.

Lejarraga, T., & Müller-trede, J. (2016). When Experience Meets Description : How Dyads Integrate Experiential and Descriptive Information in Risky Decisions When Experience Meets Description : How Dyads Integrate Experiential and Descriptive Information in Risky Decisions. *Management Science, Articles i*, 1–19.

Lepori, G. (2009). Dark Omens in the Sky: Do Superstitious Beliefs Affect Investment Decisions? *Available at SSRN 1428792*, 1–52.

Leslie, S. J. (2008). Generics: Cognition and Acquisition. *Philosophical Review*, *117*(1), 1–47. http://doi.org/10.1215/00318108-2007-023

Leslie, S. J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, *65*(1), 15–31. http://doi.org/10.1016/j.jml.2010.12.005

Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, *41*(3), 397–404. http://doi.org/10.1037/h0035357

Levey, A. B., & Martin, I. (1975). Classical conditioning of human “evaluative” responses. *Behaviour Research and Therapy*, *13*(4), 221–226. http://doi.org/10.1016/0005-7967(75)90026-1

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. http://doi.org/10.1177/1529100612451018

Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology*, *6*(3), 245–254. http://doi.org/10.1037/h0021220

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. http://doi.org/10.1037/0022-3514.47.6.1231

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social …*, *37*(11), 2098–2109.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, *54*(248), 1–95.

Ludvig, E. a, & Spetch, M. L. (2011). Of black swans and tossed coins: is the description-experience gap in risky choice limited to rare events? *PloS One*, *6*(6), e20262. http://doi.org/10.1371/journal.pone.0020262

Lybbert, T. J., Barrett, C. B., McPeak, J. G., & Luseno, W. K. (2007). Bayesian Herders: Updating of Rainfall Beliefs in Response to External Forecasts. *World Development*, *35*(3), 480–497. http://doi.org/10.1016/j.worlddev.2006.04.004

MacDougall, R. (1906). On secondary bias in objective judg- ments. *Psychological Review*, *13*(2), 97. http://doi.org/10.1037/h0072010

Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 165–170).

Malinowski, B. (1948). *Magic, Science and Religion and Other Essays*. Glencoe, Illinois: The Free Press. http://doi.org/10.2307/2181808

Mandel, N., Petrova, P. K., & Cialdini, R. B. (2006). Images of Success and the Preference for Luxury Brands. *Journal of Consumer Psychology*, *16*(1), 57–69. http://doi.org/10.1207/s15327663jcp1601_8

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions?: The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823–849. http://doi.org/10.1037/pspa0000021

Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science, 20*(11), 1309–1312. http://doi.org/10.1111/j.1467-9280.2009.02453.x

Mark, N. (1998). Beyond Individual Differences : Social Differentiation from First Principles. *American Sociological Review*, *63*(3), 309–330.

Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. http://doi.org/10.1145/2872427.2883001

Martinich, A. P., & Sosa, D. (2001). *A Companion to Analytic Philosophy*. *Blackwell Companions to Philosophy*. Oxford, UK: Blackwell Publishers.

http://doi.org/10.1353/hph.2002.0101

Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement in humans. *Learning and Motivation*, *25*, 216–232.

Matute, H. (1995). Human reactions to uncontrollable outcomes: Further evidence for superstitions rather than helplessness. *The Quarterly Journal of Experimental Psychology*, *48*(2), 142–157. http://doi.org/10.1080/14640749508401444

Matute, H., Yarritu, I., & Vadillo, M. a. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*, *102*(3), 392–405. http://doi.org/10.1348/000712610X532210

Mayer, R. C., & Davis, J. H. (1999). The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment. *Journal of Applied Psychology*, *84*(1), 123–136. http://doi.org/10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & David, S. F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734.

Mayer, R. C., & Gavin, M. B. (2005). Trust in Management and Performance: Who Minds the Shop while the Emplyees Watch the Boss? *The Academy of Management Journal*, *48*(5), 874–888. http://doi.org/10.5465/AMJ.2005.18803928

Mertens, G., & De Houwer, J. (2016). Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology*, *113*(JANUARY), 91–99. http://doi.org/10.1016/j.biopsycho.2015.11.014

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220. http://doi.org/10.1016/j.pragma.2013.07.012

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *The Behavioral and Brain Sciences*, *32*(2), 183-98-246. http://doi.org/10.1017/S0140525X09000855

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–34. http://doi.org/10.1037/a0021110

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431–44. http://doi.org/10.1037/0022-3514.53.3.431

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*(6), 1133–1139. http://doi.org/10.3758/BF03193102

Ngampruetikorn, V., & Stephens, G. J. (2015). Bias, Belief and Consensus: Collective opinion formation on fluctuating networks. *arXiv Preprint arXiv:1512.09074*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. http://doi.org/10.1037//1089-2680.2.2.175

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice-Hall.

Nurek, M., Kostopoulou, O., & Hagmayer, Y. (2014). Predecisional information distortion in physicians ' diagnostic judgments: Strengthening a leading hypothesis or weakening its competitor ? *Judgment and Decision Making*, *9*(6), 572–585.

Ono, K. (1987). Superstitious behavior in humans. *Journal of the Experimental Analysis* *of Behavior*, *3*(3), 261–271.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. http://doi.org/10.2139/ssrn.1626226

Parunak, H. V., Savit, R., & Riolo, R. L. (1998). Agent-based modeling vs. equation-based modeling: A case study and users' guide. *Proceedings of Multi-Agent Systems and Agent-Based Simulation (MABS'98)*, 10–25. http://doi.org/10.1007/10692956_2

Pavlov, I. P. (1927). *Conditioned reflexes*. London: Oxford University Press.

Pearce, J. M. (2013). *Animal learning and cognition: an introduction*. Psychology Press.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Peirce, C. S. (1877). The Fixation of Belief. *Popular Science Monthly*, *12*, 1–15.

Pelham, B. W., & Swann, W. B. (1989). From self-conceptions to self-worth: On the sources and structure of global self-esteem. *Journal of Personality and Social Psychology*, *57*, 672–680. http://doi.org/10.1037/0022-3514.57.4.672

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*, 1042–5. http://doi.org/10.1038/nature05051

Peterson, C. R., & DuCharme, W. M. (1967). A Primacy Effect in Subjective

Probability Revision. *Journal of Experimental Psychology*, *73*(1), 61–65. http://doi.org/10.1037/h0024139

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*(1), 117–121. http://doi.org/10.1037/h0022023

Peterson, D., Elliott, C., & Song, D. (2009). Probabilistic reversal learning is impaired in Parkinson’s disease. *Neuroscience*, *163*(4), 1092–1101. http://doi.org/10.1016/j.neuroscience.2009.07.033

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*(10), 1915–1926. http://doi.org/10.1037/0022-3514.37.10.1915

Petty, R. E., & Cacioppo, J. T. (1984). Source Factors and the Elaboration Likelihood Model of Persuasion. *Advances in Consumer Research*, *11*, 668–672.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354. http://doi.org/10.1037/h0023653

Pittman, T. S., & D’Agostino, P. R. (1985). Motivation and attribution: The effects of control deprivation on subsequent information processing. In G. Weary & J. Harvey (Eds.), *Attribution: Basic issues and applications* (pp. 117–141). New York: Academic Press.

Pitz, G. F. (1969a). An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology/Revue Canadienne ...*, *23*(1), 24–33. http://doi.org/10.1037/h0082790

Pitz, G. F. (1969b). The Influence of Prior Probabilities on Information Seeking and Decision-making. *Organizational Behavior and Human Performance*, *4*, 213–226.

Pitz, G. F., Downing, L., & Reinhold, H. (1967). Sequential Effects in the Revision of Subjective Probabilities. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *21*(5), 381–393. http://doi.org/10.1037/h0082998

Pogson, M., Smallwood, R., Qwarnstrom, E., & Holcombe, M. (2006). Formal agent-based modelling of intracellular chemical interactions. *BioSystems*, *85*(1), 37–45. http://doi.org/10.1016/j.biosystems.2006.02.004

Pohl, R. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Hove & New York: Psychology Press.

Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Sciences*, *4*(2), 66–72. http://doi.org/10.1016/S1364-6613(99)01429-1

Prasada, S., Khemlani, S., Leslie, S. J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–422. http://doi.org/10.1016/j.cognition.2012.11.010

Priester, J. R., & Petty, R. E. (1995). Source Attributions and Persuasion: Perceived Honesty as a Determinant of Message Scrutiny. *Personality & Social Psychology Bulletin*, *21*, 637–654.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*, *111*(3), 781–799. http://doi.org/10.1037/0033-295x.111.3.781

Proulx, T., & Inzlicht, M. (2012). The five “A”s of meaning maintenance: Finding meaning in the theories of sense-making. *Psychological Inquiry*, *23*(4), 317–335. http://doi.org/Doi 10.1080/1047840x.2012.702372

Pruitt, D. G., & Hoge, R. D. (1965). Strength of the relationship between the value of an event and its subjective probability as a function of method of measurement. *Journal of Experimental Psychology*, *69*(5), 483–489. http://doi.org/10.1037/h0021721

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological Bulletin*, *102*(1), 122–138. http://doi.org/10.1037//0033-2909.102.1.122

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS ONE*, *9*(1). http://doi.org/10.1371/journal.pone.0084888

Rakow, T., & Newell, B. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, *14*, 1–14. http://doi.org/10.1002/bdm

Reekum, V., Marije, C., van den Berg, H., & Frijda, N. H. (1999). Cross-modal preference acquisition: evaluative conditioning of pictures by affective olfactory and auditory cues. *Cognition & Emotion*, *13*(6), 831–836. http://doi.org/10.1080/026999399379104

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory* (2nd ed., pp. 64–99).

Rodriguez, N., Bollen, J., & Ahn, Y.-Y. (2015). Collective dynamics of belief evolution under cognitive coherence and social conformity. *arXiv Preprint* *arXiv:1509.01502*, 23–25.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. http://doi.org/10.1016/j.paid.2010.09.004

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. http://doi.org/10.1037/0033-2909.86.3.638

Roswarski, T. E., & Proctor, R. W. (2003). The role of instructions, practice, and stimulus-hand correspondence on the Simon effect. *Psychological Research*, *67*(1), 43–55. http://doi.org/10.1007/s00426-002-0107-4

Rozin, P., & Millman, L. (1987). Family environment, not heredity, accounts for family resemblances in food preferences and attitudes: A twin study. *Appetite*, *8*(2), 125–134. http://doi.org/10.1016/S0195-6663(87)80005-3

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social …*. Retrieved from http://psycnet.apa.org/journals/psp/50/4/703/

Rudski, J. (2001). Competition, superstition and the illusion of control. *Current Psychology*, *20*(1), 68–84. http://doi.org/10.1007/s12144-001-1004-5

Rudski, J. (2004). The illusion of control, superstitious belief, and optimism. *Current Psychology*, *22*(4), 306–315. http://doi.org/10.1007/s12144-004-1036-8

Rudski, J., Lischner, M., & Albert, L. (1999). Superstitious rule generation is affected by probability and type of outcome. *The Psychological Record*, *49*, 245–260.

Rudski, J. M., & Edwards, A. (2007). Malinowski goes to college: factors influencing students' use of ritual and superstition. *The Journal of General Psychology*, *134*(4), 389–403. http://doi.org/10.3200/GENP.134.4.389-404

Schafer, S. M., Colloca, L., & Wager, T. D. (2015). Conditioned placebo analgesia persists when subjects know they are receiving a placebo. *Journal of Pain*, *16*(5), 412–420. http://doi.org/10.1016/j.jpain.2014.12.008

Schelling, T. C. (2006). *Micromotives and Macrobehaviour*. W.W. Norton & Company.

Schmeichel, B. J., Gailliot, M. T., Filardo, E.-A., McGregor, I., Gitter, S., & Baumeister, R. F. (2009). Terror management theory and self-esteem revisited: the roles of implicit and explicit self-esteem in mortality salience effects. *Journal of Personality and Social Psychology*, *96*(5), 1077–87. http://doi.org/10.1037/a0015091

Schöbel, M., Rieskamp, J., & Huber, R. (2016). Social Influences in Sequential Decision Making. *Plos One*, *11*(1), 1–23. http://doi.org/10.1371/journal.pone.0146536

Schul, Y., & Peri, N. (2015). Influences of Distrust (and Trust) on Decision Making. *Social Cognition*, *33*(5), 414–435. http://doi.org/http://dx.doi.org/101521soco2015335414

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions : BBF*, *6*, 24. http://doi.org/10.1186/1744-9081-6-24

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance*, *27*(2), 153–196. http://doi.org/10.1016/0030-5073(81)90045-3

Schwartz, M. (1982). Repetition and Rated Truth Value of Statements. *The American Journal of Psychology*, *95*(3), 393–407. http://doi.org/10.2307/1422132

Schwarz, K. A., & Büchel, C. (2015). Cognition and the Placebo Effect - Dissociating Subjective Perception and Actual Performance. *PloS One, 10*(7). http://doi.org/10.1371/journal.pone.0130492

Segovia-Juarez, J. L., Ganguli, S., & Kirschner, D. (2004). Identifying control mechanisms of granuloma formation during M. tuberculosis infection using an agent-based model. *Journal of Theoretical Biology*, *231*(3), 357–376. http://doi.org/10.1016/j.jtbi.2004.06.031

Seligman, M. (1972). Learned helplessness. *Annual Review of Medicine*, *23*, 407–412.

Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2013). A Pessimistic View of Optimistic Belief Updating. *Manuscript Submitted for Publication*.

Shanks, D. R. (1995). *The psychology of associative learning* (Vol 13). Cambridge University Press.

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. http://doi.org/10.1016/j.tics.2015.11.002

Shermer, M. (2008). Patternicity. *Scientific American*, 48.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–21. http://doi.org/10.3758/BF03210755

Siegrist, M., Cvetkovich, G., & Roth, C. (2000). Salient value similarity, social trust, and risk/benefit perception. *Risk Analysis*, *20*(3), 353–362. http://doi.org/10.1111/0272-4332.203034

Siegrist, M., Gutscher, H., & Earle, T. (2005). Perception of risk: the influence of general trust, and general confidence. *Journal of Risk Research*, *8*(2), 145–156. http://doi.org/10.1080/1366987032000105315

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Skinner, B. (1948). Superstition in the pigeon. *Journal of Experimental Psychology*, *38*, 168–172.

Sloman, S., & Fernbach, P. (2011). Human representation and reasoning about complex causal systems. *Information, Knowledge, Systems …*, *10*(2011), 1–15. http://doi.org/10.3233/IKS-2012-0187

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*(6), 649–744. http://doi.org/10.1016/0030-5073(71)90033-X

Smith, E. R. (2014). Evil acts and malicious gossip: a multiagent model of the effects of gossip in socially distributed person perception. *Personality and Social Psychology Review*, *18*(4), 311–325. http://doi.org/10.1177/1088868314530515

Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: distributed social cognition. *Psychological Review*, *116*(2), 343–364. http://doi.org/10.1037/a0015072

Sniezek, J. a., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. http://doi.org/10.1006/obhd.2000.2926

Snyder, M. (1982). When believing means doing: Creating links between attitudes and behaviour. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), *Variability in social behaviour: The Ontario Symposium* (pp. 105–130). Hillsdale, NJ: Erlbaum.

Staudinger, M. R., & Büchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *NeuroImage*, *76*, 125–133. http://doi.org/10.1016/j.neuroimage.2013.02.074

Stern, H., & Cover, T. (1989). Maximum entropy and the lottery. *Journal of the American Statistical Association*, *84*(408), 980–985.

Stewart-Williams, S., & Podd, J. (2004). The placebo effect: dissolving the expectancy versus conditioning debate. *Psychological Bulletin*, *130*(2), 324–40. http://doi.org/10.1037/0033-2909.130.2.324

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489. http://doi.org/10.1016/S0364-0213(03)00010-7

Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy*, *69*(3), 213–225.

Strojny, P., Kossowska, M., & Strojny, A. (2016). Search for Expectancy-Inconsistent Information Reduces Uncertainty Better : The Role of Cognitive Capacity. *Frontiers in Psychology*, *7*, 1–12. http://doi.org/10.3389/fpsyg.2016.00395

Swainson, R., Rogers, R. D., Sahakian, B. J., Summers, B. A., Polkey, C. E., & Robbins, T. W. (2000). Probabilistic learning and reversal deficits in patients with Parkinson’s disease or frontal or temporal lobe lesions: possible adverse effects of dopaminergic medication. *Neuropsychologia*, *38*, 596–612.

Tausch, N., Kenworthy, J., & Hewstone, M. (2007). The confirmability and disconfirmability of trait concepts revisited: Does content matter? *Journal of Personality and Social Psychology*, *92*, 554–556.

Tetlock, P. E. (1983a). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, *45*(1), 74–83. http://doi.org/10.1037/0022-

3514.45.1.74

Tetlock, P. E. (1983b). Accountability and the Perseverance of First Impressions. *Social Psychology Quarterly*, *46*(4), 285–292.

Tetlock, P. E. (1985a). Accountability : A Social Check on the Fundamental Attribution Error. *Social Psychology Quarterly*, *48*(3), 227–236.

Tetlock, P. E. (1985b). Accountability: The Neglected Social Context of Judgement And Choice. *Research In Organizational Behavior*.

Thorndike, E. L. (1931). *Human Learning*. New York and London: The Century Co.

Tobacyk, J., & Milford, G. (1983). Belief in paranormal phenomena: Assessment instrument development and implications for personality functioning. *Journal of Personality and Social Psychology*, *44*(5), 1029–1037. http://doi.org/10.1037//0022-3514.44.5.1029

Tsang, E. W. K. (2004). Toward a Scientific Inquiry into Superstitious Business Decision-Making. *Organization Studies*, *25*(6), 923–946. http://doi.org/10.1177/0170840604042405

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Tversky, A., & Kahneman, D. (1975). Judgement under Uncertainty: Heuristics and Biases. In *Utility, probability, and human decision making* (pp. 141–162). Springer Netherlands. http://doi.org/10.1126/science.185.4157.1124

Twyman, M., Harvey, N., & Harries, C. (2008). Trust in motives , trust in competence :

Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making*, *3*(1), 111–120.

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach-avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology*, *62*(3), 161–169. http://doi.org/10.1027/1618-3169/a000282

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology*, *63*, 1–9. http://doi.org/10.1016/j.jesp.2015.11.002

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2016). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*. http://doi.org/10.1016/j.jesp.2016.10.004

van Erkel, P. F. A., & Thijssen, P. (2016). The first one wins: Distilling the primacy effect. *Electoral Studies*, *44*, 245–254. http://doi.org/10.1016/j.electstud.2016.09.002

Vyse, S. A. (2013). *Believing in Magic: The Psychology of Supersition-Updated Edition*. Oxford University Press.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wager, T. D., & Atlas, L. Y. (2015). The neuroscience of placebo effects : connecting context , learning and health. *Nature Publishing Group*, *16*(7), 403–418. http://doi.org/10.1038/nrn3976

Wagner, M. (2016). Selective Exposure , Information Utility , and the Decision to Watch Televised Debates. *International Journal of Public Opinion Research*, 1–21.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58. http://doi.org/10.1016/S0010-0277(01)00141-X

Walton, D. (1997). *Appeal to Expert Opinion: Arguments from Authority*. Penn State University Press.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. http://doi.org/10.1080/17470216008416717

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. http://doi.org/10.1080/14640746808400161

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. http://doi.org/10.1037/0022-3514.67.6.1049

Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes*, *135*, 55–69. http://doi.org/10.1016/j.obhdp.2016.05.005

Whitman, J. C., Takane, Y., Cheung, T. P. L., Moiseev, A., Ribary, U., Ward, L. M., & Woodward, T. S. (2015). Acceptance of evidence-supported hypotheses generates a stronger signal from an underlying functionally-connected network. *NeuroImage*, *127*, 215–226. http://doi.org/10.1016/j.neuroimage.2015.12.011

Wickramasekera, I. (1980). A conditioned response model of the placebo effect. *Biofeedback and Self-Regulation*, *5*(1), 5–18.

Winkler, R. L., & Murphy, A. H. (1973). Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance*, *10*(2), 252–270. http://doi.org/10.1016/0030-5073(73)90017-2

Woo, C.-K., & Kwok, R. H. F. (1994). Vanity , superstition and auction price. *Economic Letters*, *44*, 389–395.

Wood, W. (2000). Attitude Change: Persuasion and Social Influence. *Annual Review of Psychology*, *51*(1), 539–570. http://doi.org/10.1146/annurev.psych.51.1.539

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. http://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. http://doi.org/10.1006/obhd.2000.2909

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. http://doi.org/10.1016/j.obhdp.2006.05.006

Yarritu, I., & Matute, H. (2015). Previous knowledge can induce an illusion of causality through actively biasing behavior. *Frontiers in Psychology*, *6*, 389. http://doi.org/10.3389/fpsyg.2015.00389

Yarritu, I., Matute, H., & Luque, D. (2015). The dark side of cognitive illusions: When an illusory belief interferes with the acquisition of evidence-based knowledge. *British Journal of Psychology (London, England : 1953)*, 1–12.

http://doi.org/10.1111/bjop.12119

Yarritu, I., Matute, H., & Vadillo, M. a. (2013). Illusion of control. *Experimental Psychology*, *32*(2), 38–47. http://doi.org/10.1027/1618-3169/a000225

Yarritu, I., Matute, H., & Vadillo, M. a. (2014). Illusion of control: the role of personal involvement. *Experimental Psychology*, *61*(1), 38–47. http://doi.org/10.1027/1618-3169/a000225

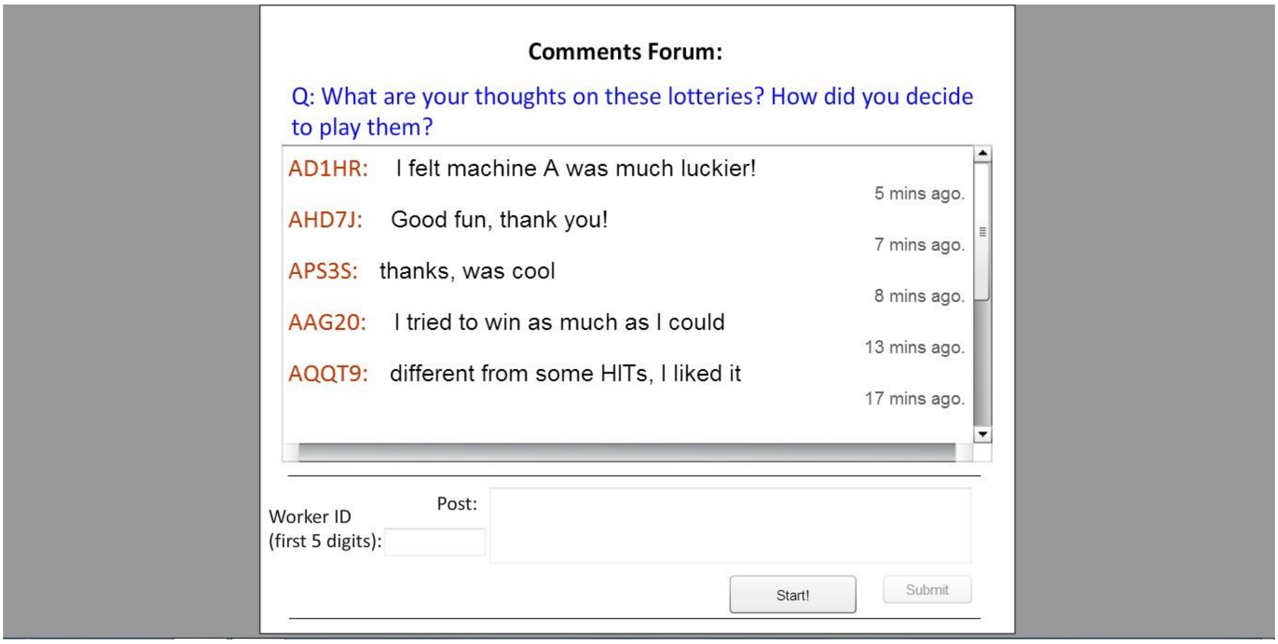
Yu, E. C., & Lagnado, D. a. (2012). The influence of initial beliefs on judgments of probability. *Frontiers in Psychology*, *3*(October), 381. http://doi.org/10.3389/fpsyg.2012.00381

Zhou, S., & Guo, B. (2016). The order effect on online review helpfulness: A social influence perspective. *Decision Support Systems*. http://doi.org/10.1016/j.dss.2016.09.016

APPENDIX

Appendix A: Experiment Screens

A.1.1 Lottery Experiment (3.2) Sample Comment Screen



Comments Forum:

Q: What are your thoughts on these lotteries? How did you decide to play them?

AD1HR: I felt machine A was much luckier! 5 mins ago.

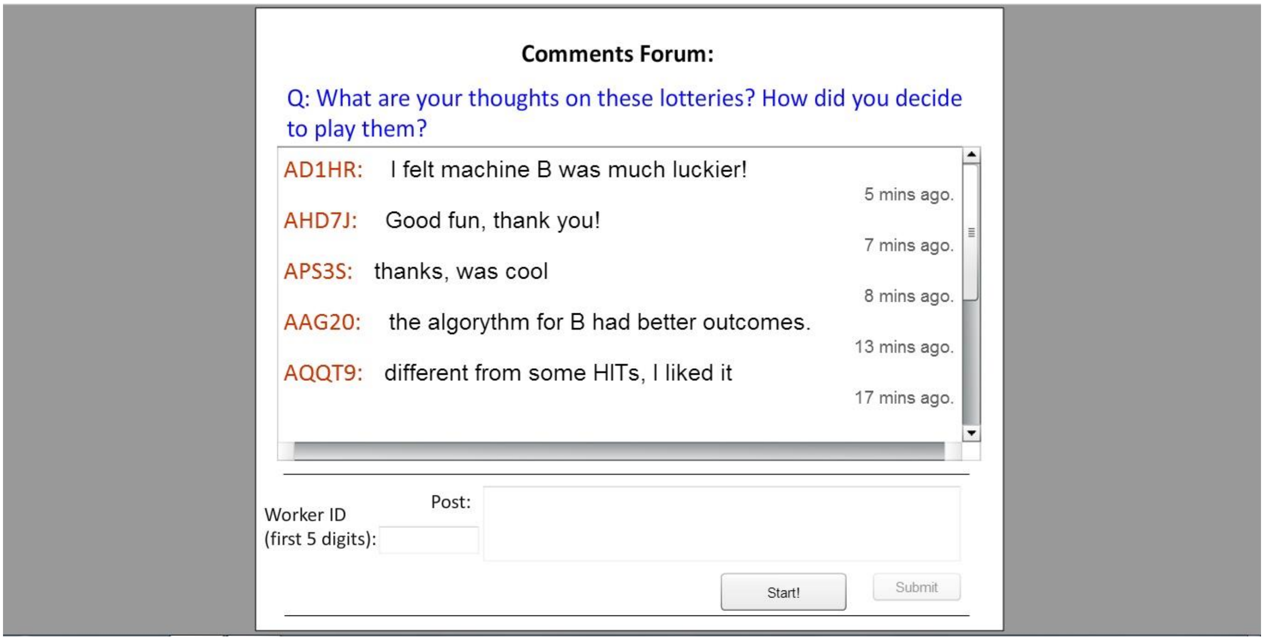
AHD7J: Good fun, thank you! 7 mins ago.

APS3S: thanks, was cool 8 mins ago.

AAG20: I tried to win as much as I could 13 mins ago.

AQQT9: different from some HITs, I liked it 17 mins ago.

A.1.2 Lottery Experiment (3.3) Sample Comment Screen



Comments Forum:

Q: What are your thoughts on these lotteries? How did you decide to play them?

AD1HR: I felt machine B was much luckier! 5 mins ago.

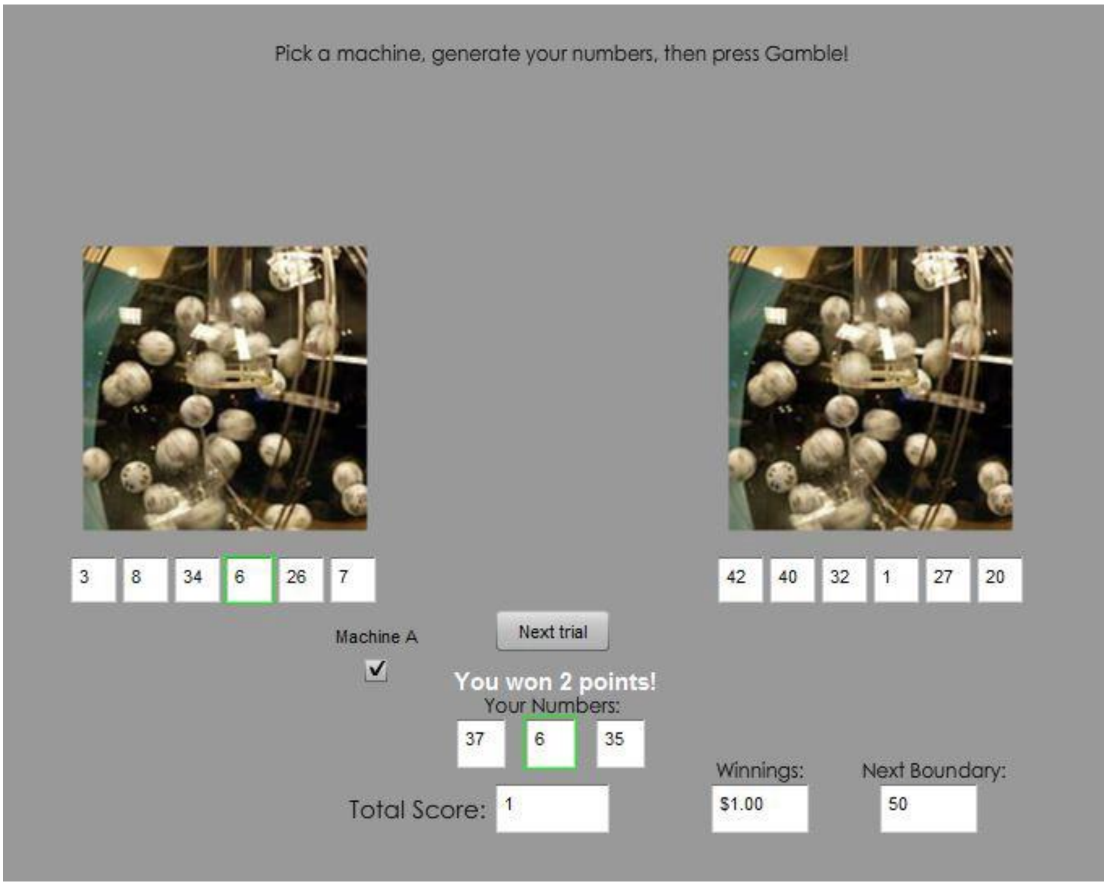
AHD7J: Good fun, thank you! 7 mins ago.

APS3S: thanks, was cool 8 mins ago.

AAG20: the algorythm for B had better outcomes. 13 mins ago.

AQQT9: different from some HITs, I liked it 17 mins ago.

A.1.3 Lottery Experiment (3.2 & 3.3) Sample Feedback Screen



Pick a machine, generate your numbers, then press Gamble!

3 8 34 6 26 7

42 40 32 1 27 20

Machine A ☑

You won 2 points!

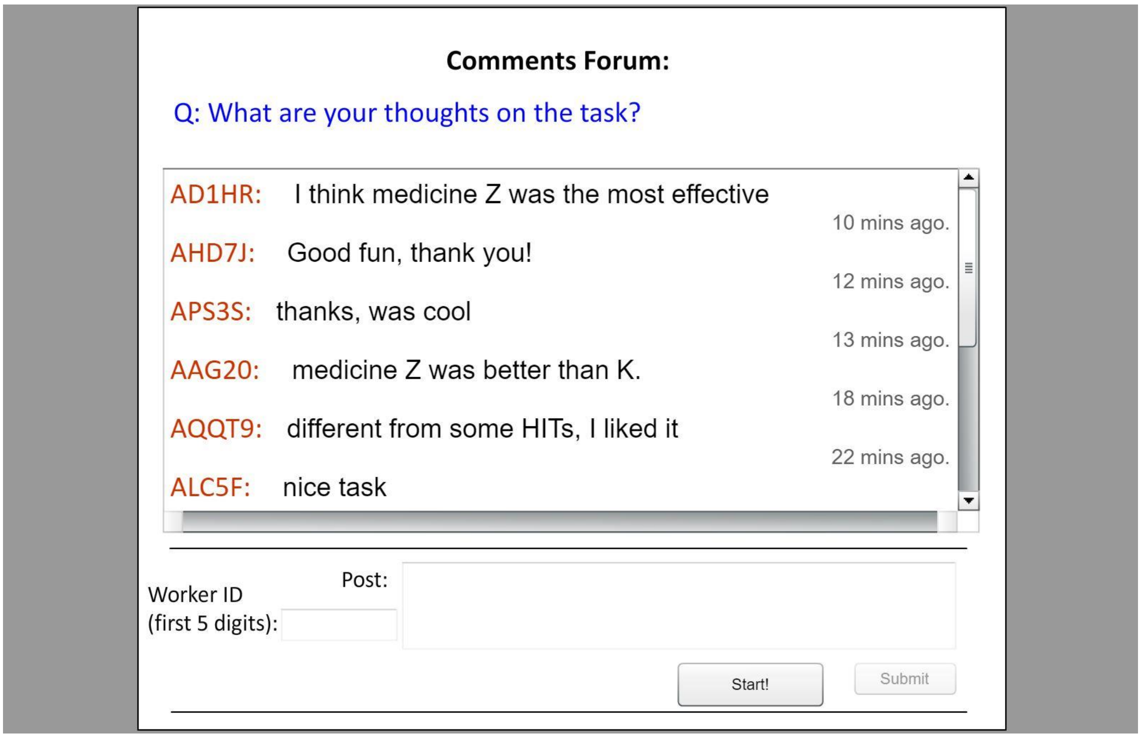
Your Numbers: 37 6 35

Total Score: 1

Winnings: $1.00

Next Boundary: 50

A.1.4 Health Experiment (3.4) Sample Comment Screen



Comments Forum:

Q: What are your thoughts on the task?

AD1HR: I think medicine Z was the most effective
10 mins ago.

AHD7J: Good fun, thank you!
12 mins ago.

APS3S: thanks, was cool
13 mins ago.

AAG20: medicine Z was better than K.
18 mins ago.

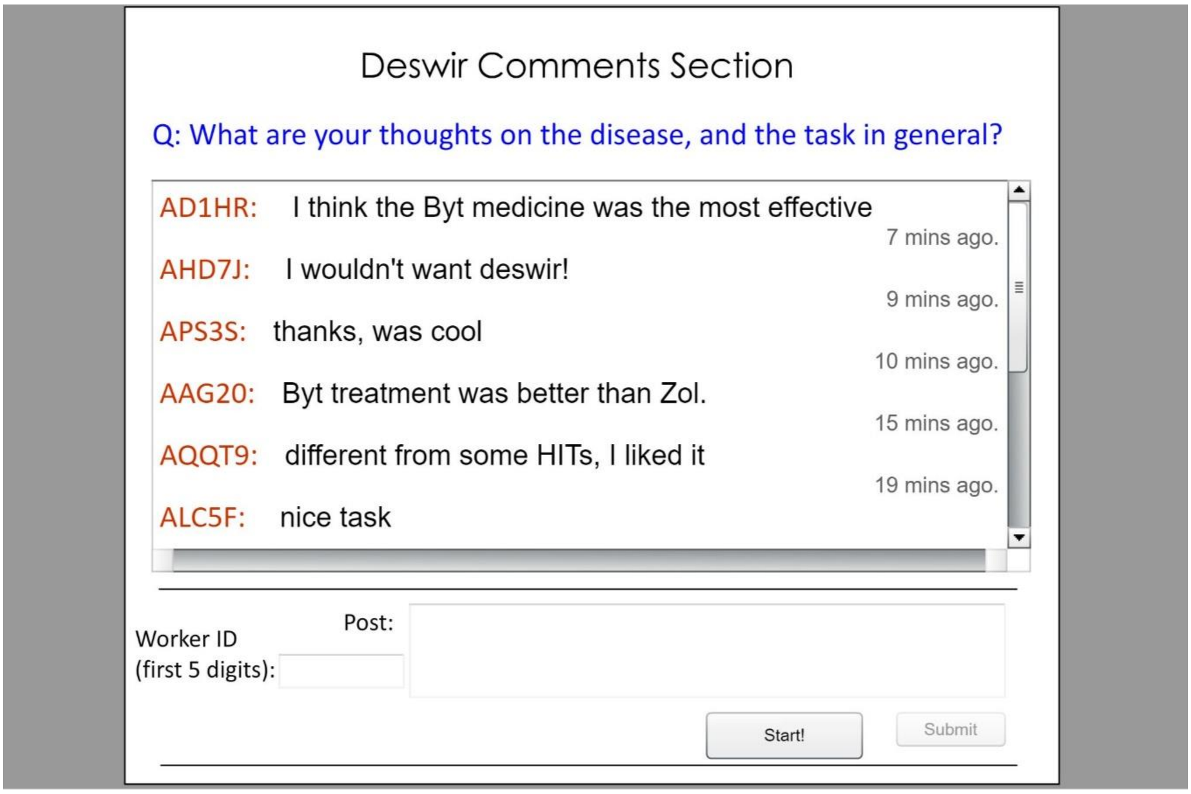
AQQT9: different from some HITs, I liked it
22 mins ago.

ALC5F: nice task

Worker ID (first 5 digits):

Post:

A.2.1 Medical Experiment (4.2, & 4.3) Sample Comment Screen (Deswir Version)



Deswir Comments Section

Q: What are your thoughts on the disease, and the task in general?

AD1HR: I think the Byt medicine was the most effective
7 mins ago.

AHD7J: I wouldn't want deswir!
9 mins ago.

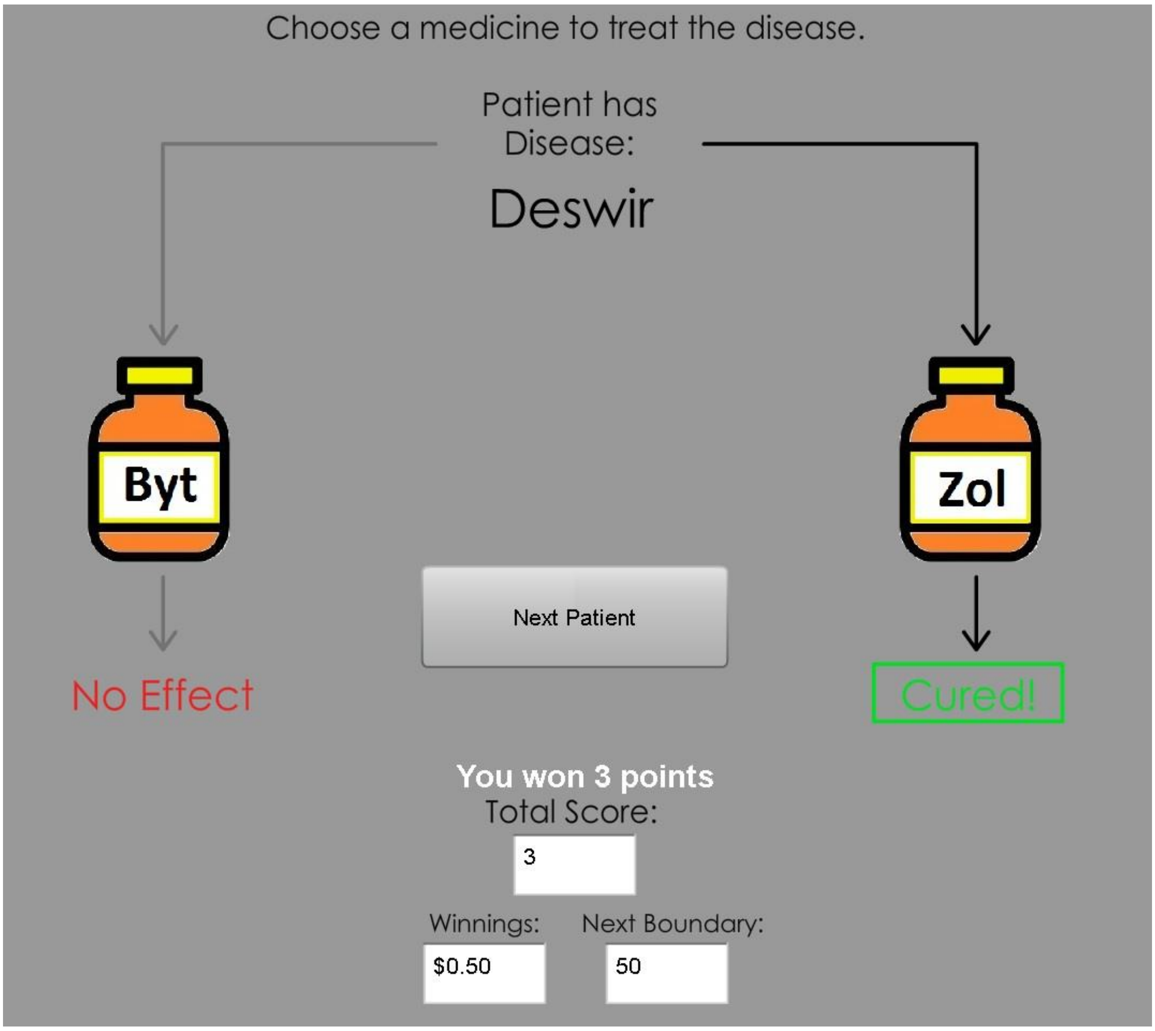
APS3S: thanks, was cool
10 mins ago.

AAG20: Byt treatment was better than Zol.
15 mins ago.

AQQT9: different from some HITs, I liked it
19 mins ago.

ALC5F: nice task

A.2.2 Medical Experiment (4.2, & 4.3) Sample Feedback Screen (Counterfactuals Present)



Choose a medicine to treat the disease.

Patient has Disease:

Deswir

No Effect

Cured!

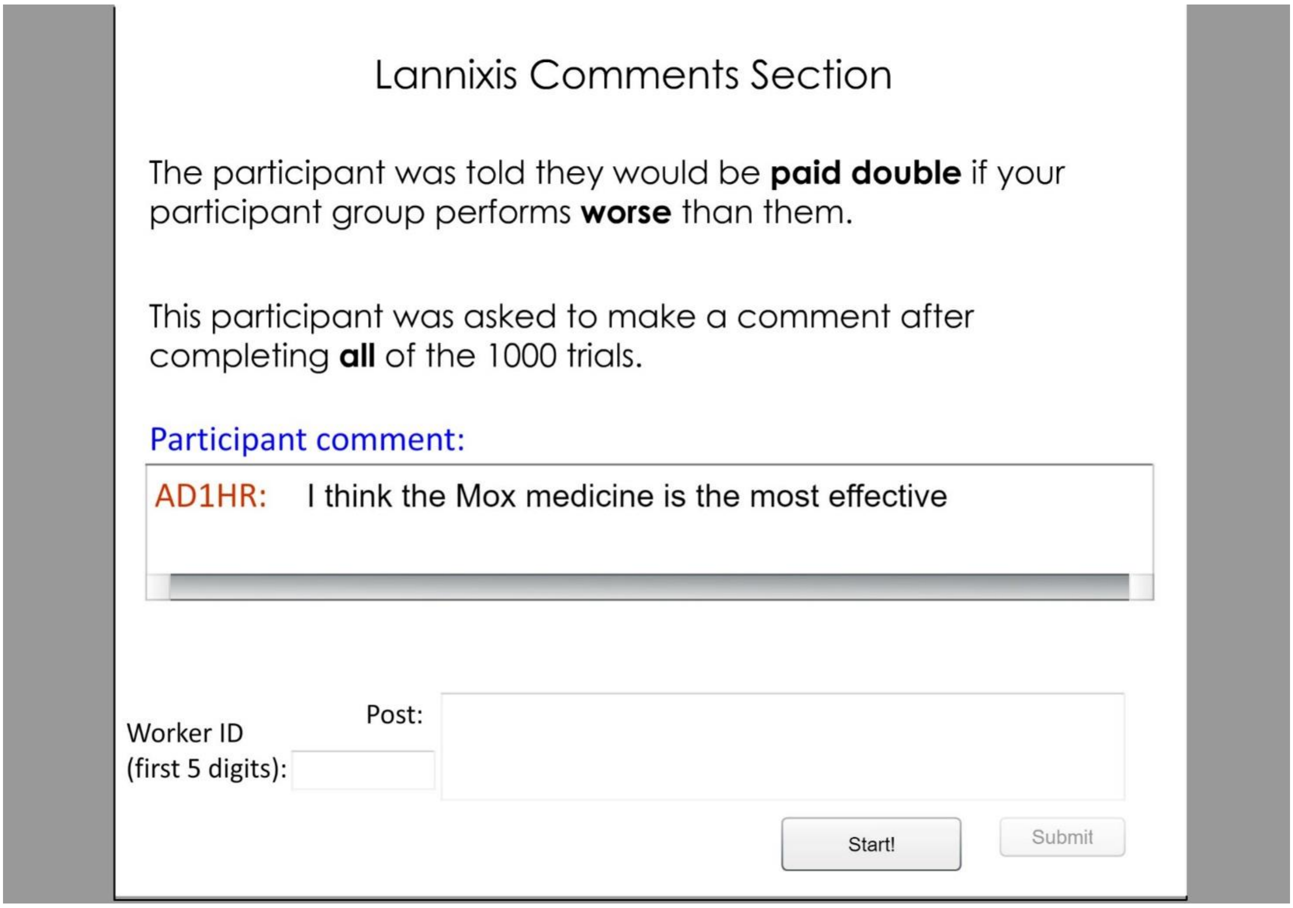
You won 3 points

Total Score: 3

Winnings: $0.50

Next Boundary: 50

A.3.1 Medical Experiment (5.3, & 5.4) Sample Comment (with low Trust and high Expertise example manipulations) Screen



Lannixis Comments Section

The participant was told they would be **paid double** if your participant group performs **worse** than them.

This participant was asked to make a comment after completing **all** of the 1000 trials.

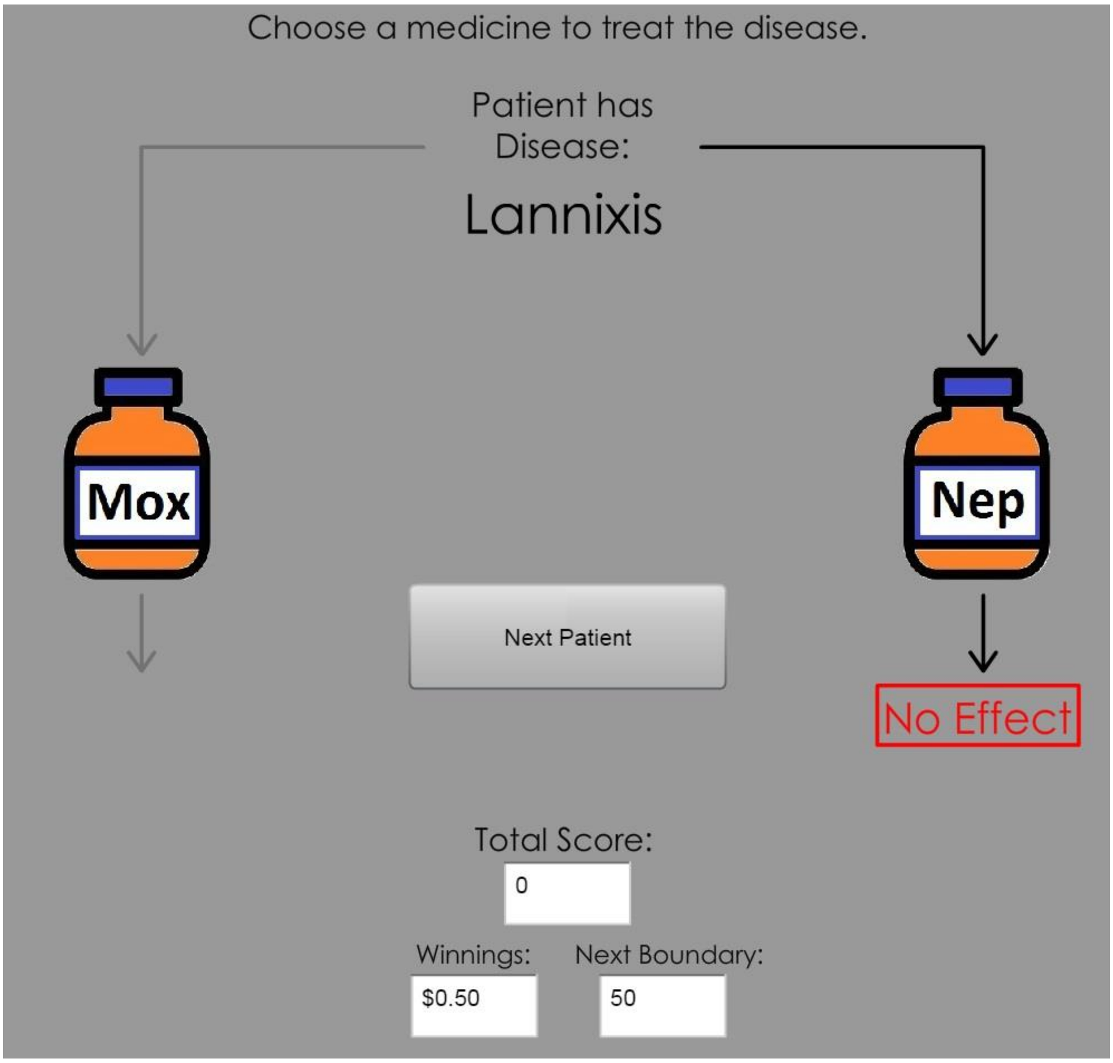
Participant comment:

AD1HR: I think the Mox medicine is the most effective

Worker ID (first 5 digits):

Post:

A.3.2 Medical Experiment (5.3, & 5.4) Sample Feedback Screen



Choose a medicine to treat the disease.

Patient has Disease:

Lannixis

Total Score: 0

Winnings: $0.50

Next Boundary: 50

Appendix B: Source Credibility Statement Pre-testing

B.1 Method

Following the premise set out in 5.2, the Pre-tests were designed to assess how statements regarding the trust and expertise of a belief's source were interpreted, when seen in conjunction with the belief itself. It also allowed for a sampling of the conditional probabilities of the eight-hypothetical trust by belief by expertise combinations from the participant population. Further, pre-testing also allowed for base-line measures of source trust and expertise ratings, as such ratings were taken without any first-hand experience from the participant, along with baseline measures of the "posterior" judgements used in the full task (binary preference, confidence in that preference, and a probability estimate). The context provided to participants was the same as that used in the previous chapter; a medical decision-making task with two diseases, each with their own pair of medicines.

Participants. Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 500 prior HITs. Participants completed the experiment under the assumption the purpose of investigation was improving medical decision-making. Participants were Native English speakers between ages 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments. The task was advertised to last approximately 10 minutes.

Design. The instructions given to participants explained that they would be assessing a comment made by a previous participant, making judgements and answering some associated questions. As such, during pre-testing, participants did not complete any trials themselves, only hearing about the diseases and their medicines in the instructions. The purpose of the design was instead to extract ratings of the trust and expertise statements regarding the source, which appeared along with the communicated belief. These statement types were manipulated between-subjects in a 2 (high or low trust) by 2 (high or low expertise) design. Further, the conditional probabilities for the 8 possible combinations of source credibility elements (trustworthy or not; expert or not; belief being true or not) were also elicited in line with the Bayesian model of source credibility outlined in 5.1.2.

Procedure. Before making any judgements or answering any questions, participants were shown the "comment section" in which a previous participant had written their thoughts regarding one of the diseases. This comment was rigged to appear to be from another MTurk participant whom had taken a full version of the health task (complete with fake MTurk ID numbers). This comment indicated a directional hypothesis regarding one of the medicines for the disease ("I think the Zol medicine was the most effective"). This comment was accompanied by trust and expertise statements regarding the source of the belief, which were randomized to be either high (“*The participant below was told they would be* ***paid double*** *if the next participant group performed* ***better*** *than them.*”, for trustworthiness; “*This participant was asked to make a comment after completing* ***all*** *of the 1000 trials.*”, for expertise) or low (“*The participant below was told they would be* ***paid double*** *if the next participant group performed* ***worse*** *than them.*”, for trustworthiness and “*This participant was asked to make a comment after completing* ***only 1*** *of the 1000 trials.*”, for expertise).

After viewing the comment section (with trust and expertise statements), participants then had to rate the trust and expertise of the previous participant (source), based on the statements they had just read. A reminder of each statement was provided for each rating. The rating question for trust was “*The participant is* ***completely trustworthy*** *(i.e. the participant will be truthful* ***to the best of their abilities****)?*”, which was rated from 0 (certain this is false) to 100 (certain this is true). Meanwhile, the rating question for expertise was "*The participant has* ***accurate knowledge*** *about the* ***effectiveness*** *of the medicines?*", again rated from 0 (certain this is false) to 100 (certain this is true).

Once trust and expertise ratings were completed, participants then completed posterior measures for the disease they had seen a comment about. This consisted of a binary preference for that disease's pair of medicines, the confidence in that preference, and a probability estimate for the distribution of cures between those medicines. Subsequently, participants then completed the 8 conditional probability ratings based on idealized scenarios, each rated from 0 (certainly false) to 100 (certainly true), and example for the trustworthy expert telling the truth (P(H|T,E)) is below:

"*Imagine that a* ***completely reliable*** *person has access to* ***complete and accurate knowledge*** *about the effectiveness of a medical product. How likely is it that this person will state the medicine was effective is the medicine was in fact* ***effective****?*"

Parts in bold changed based on the particular conditional being elicited. Upon completing the conditional section, participants then completed a demographics questionnaire and the Need for Closure measure (Roets & Van Hiel, 2011; Webster & Kruglanski, 1994). Following completion of the task, participants were debriefed and given an email to contact if they had any further questions. Those who failed the attention check were removed (i.e. those not reading carefully).

B.2 Pre-testing Outline

Over the course of the pre-testing period, 3 pre-tests were run. Each pre-test consisted of 200 US based MTurk participants, with 50 participants in each of the 4 groups (high/low trust by high/low expertise manipulations). Across all pre-tests the methodology and stimuli remained the same (i.e. conditional statements, posterior measures, belief statement, etc.), with the following changes made between pre-test versions, in the process of developing suitable manipulation statements for the full experiments:

Moving from Pre-Test 1 to Pre-test 2, a change was made to the expertise statements to provide a frame of reference for the expertise (i.e. the low expertise statement was changed from "*The participant was asked to make their comment after experiencing **only 1** trial.*" to "*The participant was asked to make their comment after experiencing **only 1 trial of 1000**.*"). This change was made due to low expertise statements being rated optimistically expert (high expertise had a mean rating of 62.85 (+/- 23.55), whilst low expertise had a mean rating of 61.5 (+/- 23.46)).

Moving from Pre-test 2 to Pre-test 3, a change was made to strengthen the polarity of the trust statements, in light of the stronger expertise statements efficacy on trust ratings (i.e. the high trust statement was rated lower if had been accompanied by the low expertise statement, rather than the high expertise statement, and the reverse for low trust statements). To resolve this, Pre-test 3 strengthened the wording of the trust statements and trust rating procedure to highlight its independence from expertise level. Based on the success of this change, the manipulations used in Pre-test 3 were carried forward to the Experiments.

B.3 Results

Where appropriate, pre-test data were combined into joint analyses. In such instances, the variables in question (either independent or dependent) must have methodological coherence between pre-test versions and have been assessed with Bayesian T-tests / ANOVAs using pre-test version as a factor. If pre-test version as a factor yields a Bayes Factor of less than 1/3rd for the DV in question, then a joint analysis is eligible (taken as evidence for the null, in accordance with recommendations by Dienes, 2014).

Conditionals. The conditional probability assessments remained unchanged across all three versions of the pre-tests, and were therefore eligible for Bayesian assessment. Assessing all 8 of the DV conditionals individually, strong support was found for the null hypothesis that pre-test version did not affect conditionals (highest BF = 0.121). As such, the means for the 8 conditionals are shown below in Table B.1, taken as an average across 588 valid cases:

Table B.1: Pre-test: Conditional Probabilities

Conditional					Probability
Probability of belief being **true** given:	**Expertise**	Expert	**Trustworthiness**	High	86.54
				Low	48.89
		Novice	**Trustworthiness**	High	47.08
				Low	43.58
Probability of belief being **false** given:	**Expertise**	Expert	**Trustworthiness**	High	30.53
				Low	50
		Novice	**Trustworthiness**	High	35.36
				Low	50

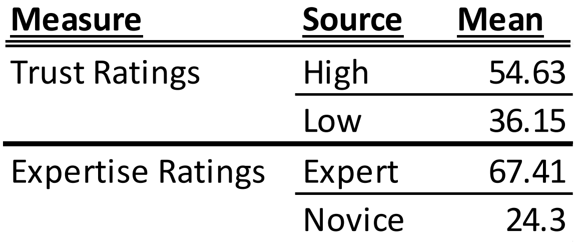
Expertise. Given the methodological change to expertise statements was made between first and second pre-test versions, the second and third pre-tests were eligible for Bayesian assessment. Once again using Pre-test version as the between-subjects factor in a Bayesian independent samples T-Test, strong support was found for the null, BF_{10} = 0.125. Accordingly, using the conjoined sample of 388 (190 high expertise, 198 low expertise) from Pre-tests 2 and 3, the mean ratings of expertise for the high and low expertise statements were 67.41 (+/- 27.04) and 24.3 (+/- 29.5) respectively. This difference was found to be highly significant, $BF_{10} = 7.137 * 10^{36}$.

Effect of Expertise on "Posterior" Judgements. Expertise as a manipulated factor was found to have a significant effect on the rated confidence in a binary preference (i.e. without having seen any evidence, how confident are you in choosing one of the two medicines), $BF_{10} = 2.015 * 10^{12}$, with higher expertise resulting in higher levels of confidence. This effect was further corroborated by a Bayesian Pearson correlation of expertise ratings (irrespective of condition) on confidence, finding a strong positive relationship, r = .510, N = 388, BF_{10} = 1.717 * 10^{29}, indicating that those who rated the expertise of the source as higher were also more confident in their medicine preference.

Trust. Given the methodological change to trust statements was made in the final pre-test version, only the sample from Pre-test 3 was eligible for subsequent analysis. Accordingly, using the sample of 186 valid cases (87 high trust, 99 low trust), the mean ratings of trust for the high and low trust statements were 54.63 (+/- 33.26) and 36.15 (+/- 28.18) respectively. This difference was found to be highly significant, BF_{10} = 319.4.

Effect of Trust on "Posterior" Judgements. Trust as a manipulated factor was found to have an effect on the binary preferences using a non-parametric Chi-squared analysis, χ^2 (1, N = 188) = 6.427, p = .011, with those in high trust conditions choosing the belief indicated medicine proportionately more than those in the low trust conditions. Similarly, the trust manipulation had an effect on probability estimates, with a Bayesian independent samples T-test finding significantly higher probability ratings for the belief-indicated medicine in the high trust condition, BF_{10} = 18.43. Both of the above effects were further ratified by Bayesian Pearson correlations of trust ratings (irrespective of condition) on binary preferences, r = .453, N = 186, BF_{10} = 1.105 * 10^8, and probability estimates, r = .52, N = 186, BF_{10} = 1.196 * 10^{14}, with higher trust ratings indicating greater preference for the belief-indicated medicine in both measures.

Table B.2: Pre-test: Summary Table of Ratings



Measure	Source	Mean
Trust Ratings	High	54.63
	Low	36.15
Expertise Ratings	Expert	67.41
	Novice	24.3

B.4 Discussion

The three rounds of pre-testing allowed for the development of both the high and low expertise and trust statements (see Table B.2 for summary values), as well as ratification of the conditional probabilities for the task context. In the development of the manipulation statements, one incidental finding of interest was the need for a frame of reference when determining level of expertise. In particular, judgements regarding level of expertise tended towards the optimistic (i.e. assuming a greater degree of expertise than warranted) until a more concrete context was provided.

Interestingly, through the use of the judgement measures typically reserved for the end of evidence integration (binary preference, confidence in that preference, and probability estimates), the two source credibility factors were found to have selective impact. Specifically, trust was found to predict choices and judgements (i.e. high trust lead to higher proportions of binary preferences for the belief-indicated medicine, and greater allocation of probability of optimal outcomes), whilst expertise was found to predict confidence in those judgements (i.e. beliefs from experts lead to greater confidence in participant binary preferences). Such effects are preliminary, but not too surprising given the relationship between trust and motives (Cuddy et al., 2011; Schul & Peri, 2015), and expertise with competence (Goodwin, 2011; Sniezek & Van Swol, 2001). Such findings were taken into account when forming the hypotheses for Experiment 7 (5.3).

Finally, it should be noted that ratings of trust and expertise show a degree of dependency upon one another (i.e. ratings of trust are slightly higher if accompanied by a high expertise statement). However, the final result of the pre-testing is a suitable set of both trust and expertise statements, along with some prior measures for tentative comparison to subsequent experiments in which evidence has then informed communicated beliefs, and therefore assessments of source credibility, as a consequence.

Appendix C: Experiment for Elicitation of Priors for Agent-Based Models

C.1 Method

Following the requirements of the Agent-Based Models of Chapter, with the intended purpose of eliciting priors for the impact of belief alone (i.e. prior to evidence exposure), the method was designed to assess solely this impact. The context and general method provided to participants was the same as that used in the Chapter 4; a medical decision-making task with two diseases, each with their own pair of medicines, with the notable exception of *no exposure to any evidence / trials* beyond the belief manipulation.

Participants. 100 Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 100 prior HITs. Participants completed the experiment under the assumption the purpose of investigation was improving medical decision-making. Participants were Native English speakers between ages 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments. The task was advertised to last approximately 10 minutes.

Design. The instructions given to participants explained that they see comments made by previous participants, and would have to make judgements and answering some associated questions. As such, participants did not complete any trials themselves, only hearing about the diseases and their medicines in the instructions and belief manipulation. The purpose of the design was instead to extract judgements regarding the medicines based solely on the belief (or not, for the control disease judgement). As such, there were no between-subject factors.

Procedure. Before making any judgements or answering any questions, participants were shown the "comment section" in which previous participants had written their thoughts regarding *one* of the diseases. These comments were rigged to appear to be from other MTurk participants whom had taken a full version of the health task (complete with fake MTurk ID numbers). These comments either indicated a directional hypothesis regarding one of the medicines for the disease ("I think the Zol medicine was the most effective") or neutral comments (“fun task, thanks”).

After viewing the comment section, participants then had to make judgements for both the belief and control diseases. These consisted of a binary preference for that disease’s pair of medicines, the confidence in that preference, and a probability estimate for the distribution of cures between those medicines.

Probability Estimate Elicitation. Given that the probability estimate forms the validation basis of the priors used in the Agent-Based Models of Chapter 6, it is worth highlighting how such an elicitation was phrased. Accordingly, participants were asked the following question:

“What is the distribution of cures between the two medicines?”

Participants then responded on a 100-point scale, using a slider from “100% Zol” to “100% Byt”, with a midpoint of 50/50 (the starting point for the slider).

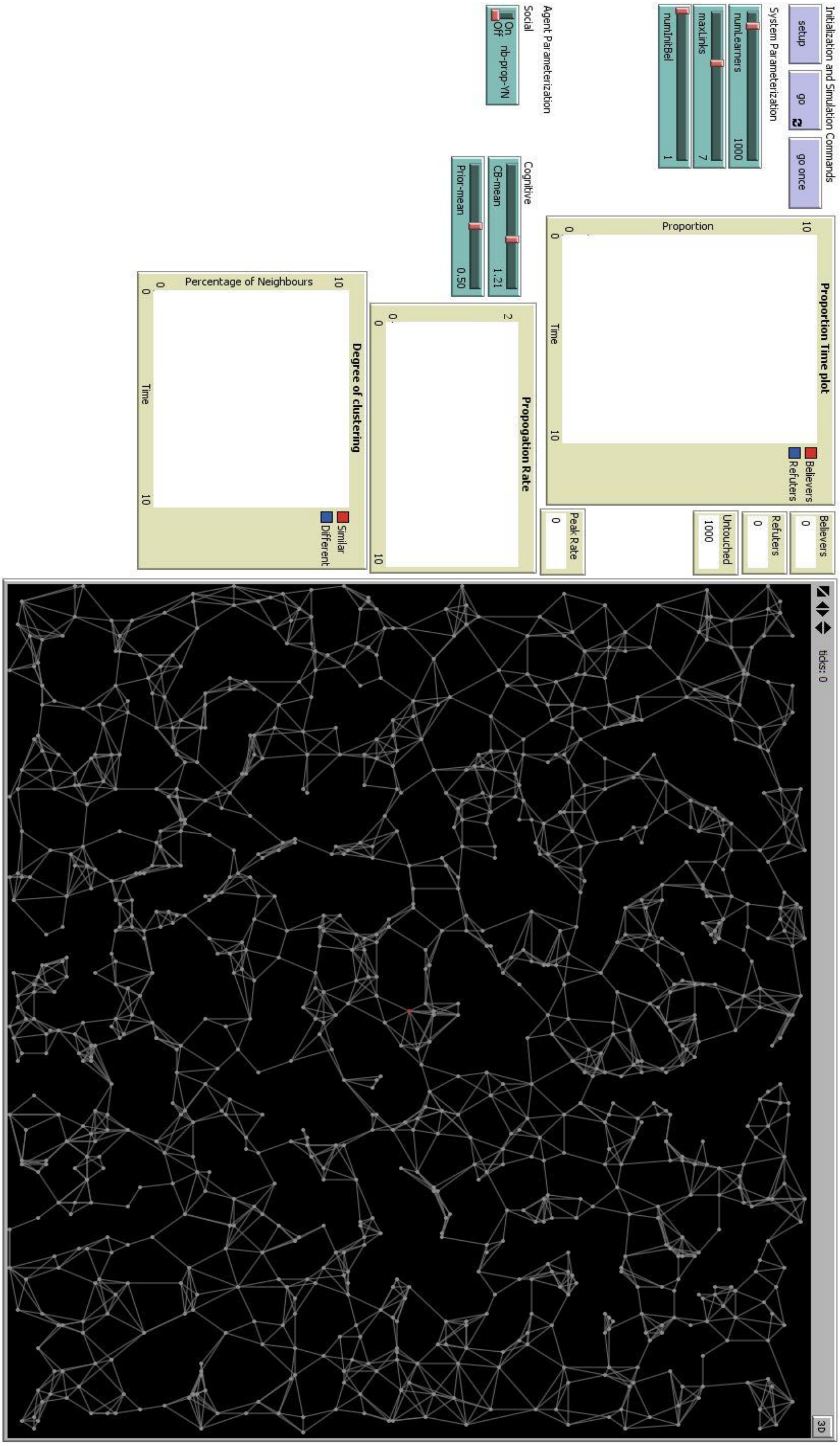
After completing the judgements, participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. Upon completing this, participants then completed a demographics questionnaire and the Need for Closure measure (Roets & Van Hiel, 2011; Webster & Kruglanski, 1994). Following completion of the task, participants were debriefed, paid for their time, and given an email to contact if they had any further questions.

C.2 Results

Descriptives and Processing. If any of the 100 participants had no recollection of the belief manipulation comments, they were removed from the subsequent analysis, leaving 77 participants. The decision to remove those who failed was taken following the same protocol and reasoning as in all previous experiments. The average age of remaining participants was 38.09 years (*SD* = 12.9) and they were 66.2% female.

Probability Estimate. Consequently, the variable of interest (for subsequent use in model validation as a prior) was the probability estimate for the belief disease. This yielded a mean value of 38.52 (i.e. a deviation from the neutral point of 50; no probability weighting towards either medicine), indicating a probability favouring the sub-optimal, belief-indicated option. In comparison, the control disease probability estimate yielded a mean of 50.18. Accordingly, a deviation of 11.48 (or .1148 in terms of P(H)) could be attributed as the impact of belief on the prior.

Appendix D: Example Agent-Based Model GUI set-up for Basic Model



numLearners 1000

maxLinks 7

numInitBel 1

○ On ● Off nb-prop-YN

CB-mean 1.21

Prior-mean 0.50

Believers 0

Refuters 0

Untouched 1000

Peak Rate 0