The Computational Underpinnings of Learning from False Information

Hamid Razi

University of East Anglia

School of Psychology

October 2025

A thesis submitted in partial fulfilment of the requirements of the University of East Anglia for the degree of Doctor of Philosophy

Declaration

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

Abstract

Misinformation poses a significant challenge to modern society, yet our understanding of how people process false information remains limited. This thesis investigates the computational and neural mechanisms underlying learning from false information, with a focus on whether well-established learning biases persist even when information is debunked. Across three studies combining behavioural testing, computational modelling, and neuroimaging, I demonstrate that people continue to learn from information explicitly marked as false, and that this learning is biased. In the first study (two experiments), using a reinforcement learning task, I show that confirmation bias persists for false information. Participants exhibited higher learning rates for confirmatory versus disconfirmatory feedback no matter the veracity. The second study, using a belief-updating paradigm, reveals that optimistic update bias similarly persists for false information. Participants updated their beliefs more strongly in response to false good news than false bad news about adverse future life events.

Computational modelling across both paradigms identified a consistent pattern: a model with four learning rates, separating information desirability (confirmatory/good news versus disconfirmatory/bad news) and accuracy (true versus false), best explained participants' behaviour. Further, the strength of both confirmation bias and optimistic update bias was similar for true and false information. Albeit effective in reducing false information integration, debunking was less effective for desirable vs undesirable false information.

The third study used functional MRI to examine the neural basis of biased false information processing. Results revealed that activity in the ventromedial prefrontal cortex (vmPFC) was modulated by an interaction between accuracy and confirmation, showing higher activation when participants learned that confirming (vs disconfirming) evidence was true, or that disconfirming (vs confirming) evidence was false. These findings identify mechanisms that support learning from false information despite debunking, with implications for understanding vulnerability to misinformation and developing effective interventions.

Contents

List of Figures

List of Tables

Acknowledgments

The past three years were filled with all sorts of new experiences for me. I moved from my home country Iran to England and this, in and of itself, was a monumental change for a guy who had never set foot in another country in his life. Albeit challenging with many bouts of home sickness, several people made this journey fulfilling. My supervisor Neil made me feel so lucky to be his student and have the opportunity to learn from him and I couldn't thank him enough for all the scientific and non-scientific support he gave me. He's the best. My second supervisor Tom was always a fun company whose scientific insights and encouragements I always appreciated - too bad I didn't know poker to take part in his alluring poker nights! My friend Amir made the transition to UK life smooth, always there when I needed him, and I will never forget all the cinema trips we made together and the deep talks that ensued every single time. Would also like to thank Will and Mark who supported me in conducting the fMRI experiment within such a short period of time in the very last months of my PhD. Mark made the many hours-long data collection sessions fly by with his charming company – I support Everton now because of him!

Finally, I would like to thank my family who were my biggest fans and cheered me on every step of the way, without whom I don't think I could overcome the emotional turmoil of the first few months. My little sister Zahra and my nieces Elina and Avina were a beacon of joy. Special thanks goes to my younger brother Hossein who missed me the most and couldn't wait for me to come back and beat him in FIFA for the 1000th time.

Chapter 1: Introduction

1.1 Introduction Overview

In this introduction, I will lay the groundwork for my thesis by arguing for a computational approach to studying how people learn from false information. I will first outline the scope of the problem by establishing its scale and reviewing psychological accounts of why false beliefs emerge and persist. I will then posit that a key reason for this is biased information integration, reviewing studies that show we form false beliefs by selectively filtering *true* information (e.g., Lord, Ross, & Lepper, 1979; Kunda, 1990; Nickerson, 1998; Epley & Gilovich, 2016; Garrett & Sharot, 2016; Van Bavel & Pereira, 2018). Next, I propose that another reason why false beliefs persist is that a similar biased mechanism might also be at play when processing *false* information. Finally, I outline how the framework of reinforcement learning (RL) has been successful in formally modelling and quantifying these biases for true information; therefore, I decided to use a similar computational modelling approach to test whether the well-documented biases (e.g., confirmation bias) exist in the face of false information, generating false beliefs.

1.1 The Problem of False Beliefs & False Information

1.1.1 Defining False Beliefs and False Information

False beliefs occur when one's mental representation of the world does not correspond to its actual state (Wimmer & Perner, 1983). Such beliefs are often deeply held and can influence one's judgments, decisions, and behaviours (Ecker et al., 2022). For example, a person may hold the false belief that an unproven medication like ivermectin is an effective cure for a viral disease like COVID-19 (Van Scoy et al., 2022) or the belief that the 9/11 terrorist attacks were an inside job (Nyhan & Reifler, 2010) despite any proof. One of the reasons for the formation of false beliefs is exposure to false information. Put simply, false information is the external claim, while a false belief is the internal conviction that can be formed once you are exposed to false information.

False information refers to information that is verifiably false (Aïmeur et al., 2023) and represents an incorrect view of the state of the world (Pennycook & Rand, 2020). This definition, however, encompasses several related concepts such as fake news (Allcott & Gentzkow, 2017), misinformation (Ecker et al., 2022), and disinformation (Kapantai et al., 2021). One can distinguish between these concepts using two features: intent - whether the purpose is to mislead or cause harm - and authenticity - whether the content is verifiably false or genuine (Aïmeur et al., 2023). For instance, misinformation is a type of false information that does not *intend* to mislead, while disinformation does intend to mislead, so the only difference between the two is in the purpose behind them. In practice, however, it is difficult to understand the intention behind sharing false information (Allen et al., 2025; Pennycook & Rand, 2021), so following Pennycook & Rand (2022), I will use "False Information", "Misinformation", and "Disinformation" interchangeably as "any information that turns out to be false". This is in line with how a recent consensus report from the American Psychological Association (APA) defined misinformation: ***any information that is demonstrably false or otherwise misleading, regardless of intention or source*** (van der Linden et al., 2023).

1.1.2 Consequences of False Information

The proliferation of false information poses a danger to societies worldwide. A primary concern is its corrosive effect on civic and social cohesion. False narratives represent a threat to democratic processes, as they often spread along partisan lines, reinforcing existing political divisions and eroding public trust in institutions like science, government, and the media (Lazer et al., 2018). This erosion can have violent consequences, as demonstrated when online conspiracy theories spill over into real-world attacks, such as the "Pizzagate" incident (Fisher et al., 2016). The incident involved an online conspiracy theory accusing a Washington D.C. pizzeria of housing a child trafficking ring run by prominent Democratic Party officials. Motivated by these claims, a man travelled to the restaurant and fired an assault rifle inside to "self-investigate" and rescue children he falsely believed were captive (Fisher et al., 2016).

Beyond the civic sphere, false information contributes to vaccine hesitancy, obstructs disease containment measures, fosters divisive rhetoric, and leads to misallocation of vital health resources (Borges do Nascimento et al., 2022; Gabarron et al., 2021; Pierri et al., 2022; Roozenbeek et al., 2020; Zimmerman et al., 2023). The economic impact can also be severe, such as when a false tweet from a compromised account temporarily erased billions in stock market value (Rapoza, 2017). This widespread societal harm underscores the need to understand how false information is learned and integrated into an individual's system of beliefs. To do so, we should first recognise that while the problem of false information feels modern, its roots are ancient and extend to other animals.

1.1.3 The Pervasiveness of False Information

The use of false information as a tool of influence is a consistent feature of human history, dating back thousands of years (Soll, 2016). From the fabricated charges that led to the death of Socrates in ancient Athens to Roman emperors using propaganda on coins, leaders have long understood the power of manipulating public perception (Aïmeur et al., 2023). This practice continued through events like the "Great Moon Hoax" of 1835, which saw a newspaper publish fictional articles about life on the moon as fact, and the systematic state-sponsored propaganda of the 20th century's major conflicts (Allcott & Gentzkow, 2017; U. K. H. Ecker et al., 2022).

More recently, persistent falsehoods, such as the retracted and debunked link between vaccines and autism (Lancet, 2010), demonstrate how damaging misperceptions can endure (Lewandowsky et al., 2012). The manipulation of information has therefore been a constant force shaping human behaviour, but humans are not the only animals engaging in misinformation. Indeed, this behaviour is a well-documented phenomenon among other animals in the form of deception: the act of transmitting misinformation to mislead others (Drerup et al., 2025; Mitchell, 1986; Šekrst, 2022; Stuart-Fox, 2005; Whiten & Byrne, 1988). Examples include chimpanzees averting their gaze from food to avoid giving cues to others, cornered guenons emitting a false social alarm call to end an aggressive chase, chimpanzees feigning a limp to avoid a dominant individual, and low-ranking gorillas building "fake nests" to covertly approach a desirable infant (Whiten & Byrne, 1988).

Given how common transmitting false information is across our own history and throughout the animal kingdom, it is clear this is not a superficial or modern problem. To understand why false beliefs emerge and persist when exposed to such false information, I will now turn from this broad context to the specific, psychological accounts of what makes our species susceptible to it.

1.1.4 Psychological Accounts of Why False Beliefs Emerge and Persist

1.1.4.1 Inattention and Motivated Reasoning

It is useful to think of inattention and motivated reasoning as two interacting pathways that lead to the formation of false beliefs. The first pathway is a failure to engage in reasoning. This is pure inattention or "lazy thinking," where an individual's reliance on intuition in a fast-paced environment leads them to accept or share a falsehood without deliberation, regardless of its content (Bago et al., 2020; Pennycook & Rand, 2021). Simple prompts asking people to consider accuracy are effective at reducing belief in false headlines, indicating that this lack of scrutiny is a major driver of the problem (Brashier et al., 2020; Pennycook et al., 2021). For instance, Pennycook et al. (2021) had participants in the treatment group rate the accuracy of a single non-partisan news headline before performing the main news-sharing task. The results consistently showed that this intervention significantly increased sharing discernment, leading participants to be less likely to share false headlines while their willingness to share true headlines remained unchanged.

However, it is important to recognize that this reliance on intuition is not a shortcoming in and of itself. These heuristic responses are often adaptive strategies for efficient decision-making under uncertainty (Gigerenzer, 2008). In most daily contexts, reliance on simple cues - such as familiarity or social consensus - is an effective shortcut to the truth. The vulnerability to false information arises not because these heuristics are fundamentally broken, but because the modern information environment (e.g., social media feeds containing clickbait and bots) is designed to exploit them. Therefore, what appears as "lazy" thinking is often an adaptive trade-off that is not suited to a novel digital context.

The second pathway is a biased treatment of reasoning. This is motivated reasoning, where an individual's goal is not to find the truth, but to defend a prior belief or identity (Kunda, 1990). An individual will selectively use their attention and scrutiny only towards information that challenges their worldview, while uncritically accepting information that aligns with it. A classic study on capital punishment powerfully demonstrated this: when shown the exact same mixed body of evidence, both supporters and opponents of the policy became *more* convinced of their initial positions, a phenomenon known as attitude polarisation (Lord, Ross, & Lepper, 1979). Therefore, while pure inattention is a failure to scrutinize, motivated reasoning is a selective, biased way to scrutinize.

The broad concept of motivated reasoning is relevant to computationally defined learning biases in response to misinformation that are the focus of this thesis. Motivated reasoning is a descriptive theory proposing that human reasoning is often directed by goals other than accuracy, such as defending a prior belief or identity (Kunda, 1991). In contrast, biases like the Optimistic Update Bias (Sharot et al., 2011) or Confirmation Bias (Palminteri et al., 2017) are formal, mechanistic accounts of *how* this motivated reasoning can be implemented at the level of trial-by-trial learning. These learning biases can therefore be understood as the computational instantiation of the broader concept of motivated reasoning, quantifying the theory. Whilst motivated reasoning describes the 'why' - the drive to maintain a desired belief - the biased learning models used in this thesis will address the ‘how'.

1.1.4.2 The illusory truth effect

The illusory truth effect describes how simply being exposed to a false headline, even once, increases its perceived accuracy and the likelihood of believing it later (Dechêne et al., 2010). This effect can persist for weeks and occurs even when stories are labelled as contested or are inconsistent with a person's ideology, although implausibility can be a boundary condition (Pennycook et al., 2018; Unkelbach et al., 2019; Wang et al., 2016). This effect is also relevant to “errorful” learning, where learners are exposed to and then corrected on false information (Kornell et al., 2009). Whilst generating errors can sometimes strengthen later memory for the correct answer, initial exposure to misinformation can also increase its familiarity and thus perceived plausibility, amplifying the risk of illusory truth (Fazio et al., 2019).

An fMRI study has implicated the perirhinal cortex (PRC) in mediating the illusory truth effect (Wang et al., 2016). Twenty-four participants were scanned while they rated the truthfulness of unknown statements. In an initial exposure phase, participants rated 180 trivia statements, with each statement presented twice to enhance fluency. Later, in the MRI scanner, they rated the truthfulness of 60 repeated unknown, 60 new unknown, 30 repeated known (i.e., they had seen the statements before), and 30 new known statements on a 6-point confidence scale. The fMRI analyses focused on "maybe false," "maybe true," and "probably true" responses to unknown statements. The imaging analysis revealed that the PRC was the only brain region to show a significant interaction between repetition and perceived truth. Specifically, PRC activity increased with truth ratings for repeated statements (i.e., as statements were perceived as truer). However, PRC activity decreased with truth ratings for new statements. A trial-by-trial analysis further corroborated these findings, showing that increases in PRC activity predicted increases in the perceived truth of repeated statements, while decreases in PRC activity predicted increases in the perceived truth of new statements. This neurobiological evidence is important because it provides a mechanistic basis for the "fluency heuristic" account of the illusory truth effect. It suggests that our susceptibility to this bias is not a failure of high-level, critical reasoning, but rather the result of a low-level, automatic memory process. It appears that the brain is misinterpreting the signal for familiarity, processed by the PRC, as a signal for truth.

1.1.4.3 Continued Influence Effect

While the illusory truth effect describes how repetition can make a falsehood feel true, the Continued Influence Effect (CIE) explains a different but related challenge: why misinformation remains influential *even after* it has been explicitly corrected retracted (M. S. Chan & Albarracín, 2023; M.-P. S. Chan et al., 2017; Ecker et al., 2010; Johnson & Seifert, 1994; Lewandowsky et al., 2012). The key distinction is the presence of a retraction. The illusory truth effect is a phenomenon of belief formation driven by fluency, whereas the CIE is a phenomenon of belief updating failure. It describes how a known correction often fails to eliminate the influence of the original, now-debunked information, which continues to shape memory and reasoning.

A classic study (Johnson & Seifert, 1994) illustrating the CIE involved a story about a warehouse fire, which was initially blamed on negligently stored hazardous materials. Participants were later told that this detail was false, and the closet was empty. Despite remembering the correction, their subsequent inferences about the fire were still shaped by the debunked information (Johnson & Seifert, 1994). This occurs because the original information is not simply erased but coexists with the correction. This principle is well-established in research on associative learning, where the "extinction" of a conditioned response is understood not as the erasure of the original memory, but as the formation of a new, competing memory that inhibits the old one (Bouton, 2004). Similarly, a factual correction does not delete the original misinformation but instead creates a new belief that must actively compete with the original, often more compelling, narrative. People can then fail to retrieve the correction at the right moment or may struggle to update their understanding. This is particularly true if the correction creates a "causal gap" - that is, it removes the explanation for an event without offering an alternative (Ecker et al., 2010, 2022). This phenomenon is related to a broader tendency of some individuals to be less critical of weak or nonsensical claims, sometimes called "pseudo-profound bullshit", and to overestimate their own expertise, making them more vulnerable to falling for misinformation in the first place (Pennycook & Rand, 2020).

1.1.4.4 Emotional Content

Misinformation tends to elicit stronger emotional responses, particularly outrage, compared to reliable sources (McLoughlin et al., 2024; Rathje & Van Bavel, 2025). Misleading narratives are often crafted with emotionally charged content, using potent words like 'fight,' 'greed,' or 'evil' to provoke a reaction. This strategy is effective at capturing attention and prompting people to spread the content before they even know it is accurate. Consequently, efforts to combat misinformation that rely solely on factual corrections face a significant challenge, as they fail to address the underlying emotional drivers that make the content so compelling (Brady et al., 2020; Han et al., 2020).

From an evolutionary perspective, this susceptibility is a feature of a cognitive system designed for survival. In the ancestral environment, paying attention to emotional stimuli - particularly those signalling threat (fear) or moral violation (outrage) - was crucial for physical safety and group cohesion (Haselton & Nettle, 2006). Therefore, emotional information is prioritised, capturing attention rapidly and automatically to ensure that survival-relevant cues are not missed. Furthermore, this could suggest emotional misinformation is difficult to erase from memory. Under the framework of Error Management Theory, the cost of a false negative (failing to remember a valid threat, like a predator) is fatal whilst the cost of a false positive (believing a false alarm) is simply inconvenient (Haselton et al., 2015). Therefore, human memory is biased to retain emotionally arousing information as a protective mechanism.

1.1.4.5 Theory of Mind and Missing Social Cues

The evolutionary sensitivity to emotional content does not mean that we could always be fooled. Just as natural selection has shaped our attention to prioritize survival-relevant threats, it has also equipped us with counter-mechanisms to verify the source of that information. To understand the limits of our susceptibility - and the extent to which it is actually possible to fool people – I will take a look into a cognitive tool at our disposal to filter untrustworthy sources: Theory of Mind - the ability to attribute mental states, such as beliefs and intentions, to others (Sperber et al., 2010). When evaluating a claim from others, individuals simulate the mind of the source to assess two things: competence (do they know the truth?) and benevolence (do they intend to share it?) (Sperber et al., 2010). Consequently, people are difficult to fool thanks to such monitoring mechanisms whereby if someone detects cues of deceptive intent, they tend to discount the information. However, false beliefs can emerge when this monitoring mechanism fails, resulting in people missing the social cues. It is plausible that digital environments weaken this defence by stripping away the social cues (e.g., facial expressions) required for Theory of Mind to function effectively. In this view, misinformation succeeds by mimicking the signals of a benevolent source, bypassing the detectors that would otherwise prevent the false belief from taking root.

1.1.5 Socio-Technical Reasons for False Information Spread

1.1.5.1 Social and Habitual Drivers of Sharing False Information

The decision to share information is often driven by factors other than a concern for its accuracy. People may share information they know is inaccurate to fulfil other psychological needs, such as signalling group membership, expressing a moral stance, self-promotion, or even to incite chaos (McLoughlin et al., 2024). Social media platforms can amplify this behaviour through their design. These systems often prioritize content that is highly engaging, which frequently includes extremist, emotive, and polarizing misinformation (Rathje & Van Bavel, 2025). This design means that platforms can end up amplifying false content even if users are interacting with it to express outrage or disagreement (Budak et al., 2024). Therefore, users might develop social media habits of posting whatever is most likely to attract attention (Ceylan et al., 2023). Once these habits take hold, the act of sharing can become a thoughtless reflex, performed with little consideration for the truthfulness of the content or the real-world consequences of its spread (Pennycook et al., 2021).

1.1.5.2 Automated and Cross-Platform Spread

The spread of false information is significantly accelerated by automated non-human accounts, or "social bots." These bots are designed to mimic human behaviour-by liking, reposting, and commenting on content-to create an artificial sense of social consensus. By manufacturing the appearance that a piece of information is popular and widely endorsed, bots exploit human reliance on social proof, tricking users into perceiving the information as more credible and worthy of sharing (Ferrara et al., 2016; Le et al., 2019). This amplification is not confined to a single platform; coordinated bot networks are particularly effective at propagating the same false narratives across multiple online platforms simultaneously. This cross-platform contamination further increases a falsehood's reach and creates an illusion of ubiquity, making it seem more legitimate and pervasive than it actually is (Zannettou et al., 2019).

1.1.5.3 The Spread of True vs False Information

Finally, to get a sense of how problematic the spread of false information is, it is useful to compare it with how true information is disseminated. Large-scale analyses of social media platforms have revealed distinct patterns in how these two types of content travel. Vosoughi et al. (2018) analysed the spread of verifiable true and false news stories – around 126,000 news stories tweeted by over 3 million people between 2006 and 2017 - and found that falsehoods travelled six times faster than true information. While the truth rarely diffused to more than 1,000 people, the top 1% of misinformation reached between 1,000 and 100,000 individuals. As discussed in Section 1.1.4.4, this viral advantage is likely driven by the novelty and emotional intensity - particularly outrage and surprise – common in false narratives.

1.1.6 Interventions for Fighting False Information

1.1.6.1 Debunking

Debunking involves presenting a corrective message that explicitly identifies and refutes a prior piece of misinformation (Chan et al., 2017; Schwarz et al., 2007). Despite the persistence of falsehoods, research shows that corrections are effective in reducing false beliefs (Ecker et al., 2022; Porter & Wood, 2021). Its effects are often durable, lasting for weeks (Porter & Wood, 2021), and can lead to positive downstream effects by changing not just beliefs, but also behaviours like sharing and voting intentions (Ecker et al., 2022). To be effective, however, debunking must follow best practices. Corrections should provide detailed factual accounts and plausible alternative explanations that can fill the "coherence gap" left when a piece of misinformation is retracted (M. S. Chan & Albarracín, 2023; Johnson-Laird, 2012; Tenney et al., 2009). It is vital to lead with the accurate information rather than unnecessarily repeating the misinformation, which can inadvertently boost its familiarity. The correction should establish a factual frame from the outset and re-emphasize the truth at the conclusion (Ecker et al., 2022; Lewandowsky et al., 2012). Corrections are also ideally delivered by high-credibility sources, such as content experts (Vraga & Bode, 2018, 2020), though politicized voices can be effective in debunking rumours within their respective partisan groups (Berinsky, 2017). Given that resources are limited, efforts should be focused on correcting misinformation that is both widespread and has the greatest potential for harm (Ecker et al., 2022). Finally, all corrections should be delivered in a civil, careful, and thoughtful manner (Vraga & Bode, 2020), with the understanding that repeated interventions may be necessary as the effects of a single correction can wear off over time (Carnahan et al., 2021).

Despite its utility, debunking has significant limitations. The Continued Influence Effect ensures that misinformation can linger (Ecker et al., 2010). While early concerns focused on "backfire effects" where corrections strengthened misperceptions, these are now understood to be largely overstated and uncommon (Wood & Porter, 2019). More practical challenges include the fact that manual fact-checking is laborious, difficult to scale (Chuai et al., 2023a), and often arrives only after a claim has gone viral (Stein et al., 2023). Automated Natural Language Processing (NLP) fact-checking also faces hurdles due to the lack of readily available counter-evidence for novel, real-world misinformation (Glockner et al., 2022).

1.1.6.2 Prebunking and Inoculation

A proactive alternative to debunking is "prebunking," or psychological inoculation (Figure 1.1). This approach aims to prevent people from encoding misinformation in the first place by building "attitudinal resistance" (Compton, 2013; Van Der Linden, 2024). The core mechanism involves exposing people to weakened versions of persuasive messages ahead of time (Van Der Linden, 2024; van der Linden et al., 2017). By warning recipients about the threat of misleading information and identifying the manipulative technique, inoculation equips them with the cognitive tools to resist future attempts at persuasion (Christner et al., 2024; Ecker et al., 2022; Van Der Linden, 2024). Prebunking has been shown to be an effective counter to misinformation, with effects that can generalize across different topics, providing an "umbrella" of protection against various manipulation tactics (Schmid-Petri & Bürger, 2022; Traberg et al., 2022). Furthermore, successful approaches, like the interactive "Bad News Game," can spark conversations where people share their new skills with their peers. This "post-inoculation talk" helps spread the protective effects through social networks, amplifying the impact of the initial intervention (Roozenbeek & van der Linden, 2019). In Bad News Game, players take on the role of a fake news producer to attract as many followers as possible while maximizing credibility. Throughout the approximate 15-minute gameplay, players learn to master six documented techniques commonly used in misinformation: polarisation, invoking emotions, spreading conspiracy theories, trolling people online, deflecting blame (discrediting opponents), and impersonating fake accounts. Players are rewarded for using these strategies and penalized for ethical journalistic behaviour. The game simulates a social media environment, rendering text boxes, images, and Twitter posts. However, a key limitation of such approaches is requiring users to actively engage, which may exclude those most in need of inoculation, such as individuals with lower cognitive reflection (Pennycook & Rand, 2021).

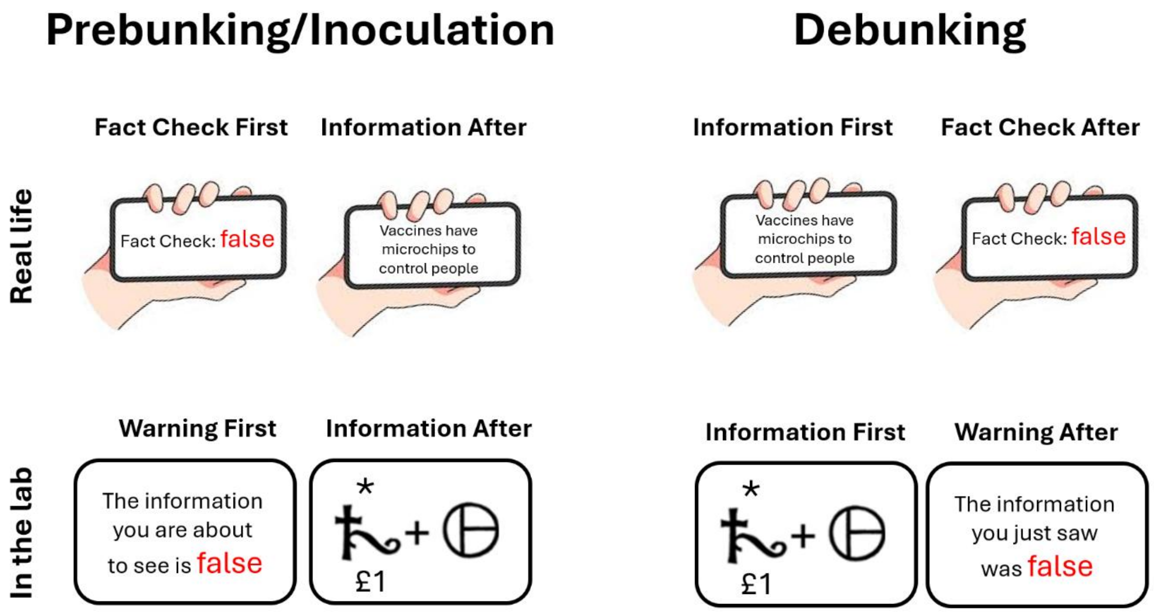


Figure 1.1: Prebunking vs Debunking. In prebunking (top left panel), people are given advance warning that the information they are about to see is false to help them prepare (inoculate themselves) against believing the false information when it arrives. This has been shown to be effective in reducing susceptibility to misinformation. In contrast, debunking (top right panel) presents the false information first, followed by a correction, which can be less effective. The lower panels show how these approaches are designed in laboratory settings, where participants receive either a warning before ("prebunking") or after ("debunking") exposure to false information. The "information" here refers to the outcome (£1) given to the chosen option in a two-arm bandit task.

1.1.6.3 Media Literacy and Critical Thinking

Improving media literacy and critical thinking skills is a long-term strategy for combating misinformation (Borges do Nascimento et al., 2022; Lee, 2018; Lutzke et al., 2019). Interventions include dedicated digital media literacy training, promoting civic online reasoning, and encouraging critical thinking through the questioning of logic, evidence, and sources (Apuke et al., 2023; Jones-Jang et al., 2021). For instance, one effective strategy is "lateral reading," which involves consulting external sources to examine the origins and credibility of a claim (Wineburg & McGrew, 2019).

1.2 How Biased Processing of True Information Creates False Beliefs

False information is not the only reason why false beliefs emerge. Biased processing of true information can also lead to false beliefs (Palminteri & Lebreton, 2022; Sharot et al., 2023; Sharot & Garrett, 2016). This process is observed when individuals receive objective, true feedback about themselves. For instance, in one study, participants took an IQ test and were then asked to estimate their rank relative to another participant. When they received 'good news' (true feedback that they had ranked higher), they updated their beliefs about their intelligence significantly. However, when they received 'bad news' (true feedback that they had ranked lower), they updated their beliefs to a lesser extent (Eil & Rao, 2011). This same asymmetric pattern is seen in financial decision-making, where students update their beliefs about their future earnings prospects far more in response to positive true information than to negative true information (Wiswall & Zafar, 2015). In each case, the resulting false belief is not caused by exposure to falsehoods, but by a motivated filtering of true information known as optimistic update bias (Sharot et al., 2011).

1.2.1 Biases Processing of True Information: Optimistic Update Bias

An optimistic update bias, also known as the "good news-bad news effect," is a pervasive tendency to integrate new information more readily when it is desirable or better than expected, compared to when it is undesirable or worse than expected, leading to false, optimistic beliefs (Garrett et al., 2014; Garrett & Sharot, 2014, 2017a; Kuzmanovic et al., 2019a; Kuzmanovic & Rigoux, 2017; Sharot & Garrett, 2016). The bias is relevant in myriad domains and contexts, including financial (Wiswall & Zafar, 2015b), health risks (Weinstein & Klein, 1995), personal attributes like intelligence (Eil & Rao, 2011b), social (Korn et al., 2012a) reinforcement learning (Lefebvre et al., 2017), and mental health (Garrett et al., 2014; Ossola et al., 2020). For instance, individuals discount negative feedback about their own attributes but incorporate positive ones (e.g., learning they are more attractive than previously thought) (Eil & Rao, 2011b).

One way to study this bias is through the update bias task (UBT) (Figure 1.2), which I have adapted for my own experiment in Chapter 4. In a typical UBT trial, participants are asked to consider a negative life event. They first estimate the probability of this event happening to them personally (E1) and to an average person like them (i.e., from the same location, age, and socioeconomic status as the participant) (eBR). After providing these estimates, they are shown the true statistical base rate (BR) for the event. Finally, they are prompted to give a second estimate of their personal risk (E2). The participant's belief change, or "update," is calculated by subtracting their first personal estimate from their second (E2 − E1). Trials are then categorized based on the information provided. When dealing with negative events, a trial is considered "good news" if the true risk (BR) is lower than the person's initial estimate (E1), suggesting the event is less likely than they thought. Conversely, a trial is labelled "bad news" if the true risk is higher than their estimate, indicating the event is more probable than they initially believed (Sharot et al., 2011; Sharot & Garrett, 2022). An optimistic update bias occurs when the update values for events classified as good news is higher than events classified as bad news, which is consistently shown and replicated in various studies (e.g., Garrett et al., 2014; Garrett & Sharot, 2017; Korn et al., 2012b; Kuzmanovic et al., 2019b; Oganian et al., 2019).

An alternative design for this task separates the first estimate (E1) and the second one (E2) into two distinct sessions held at different times (e.g., Sharot et al., 2011). Because of the delay between sessions, a control for memory is necessary. To address this, participants are often asked at the end of the second session to recall the actual probability they were shown for each event. Regardless of the task version, researchers often gather additional data to control for other confounds. These variables, collectively known as subjective ratings, include familiarity with the event, past experience with the events, how negatively or positively they found the event to be, how vividly they imagine the event, and how emotionally arousing they perceive the event. The bias survives after controlling for these ratings (Garrett & Sharot, 2017a).

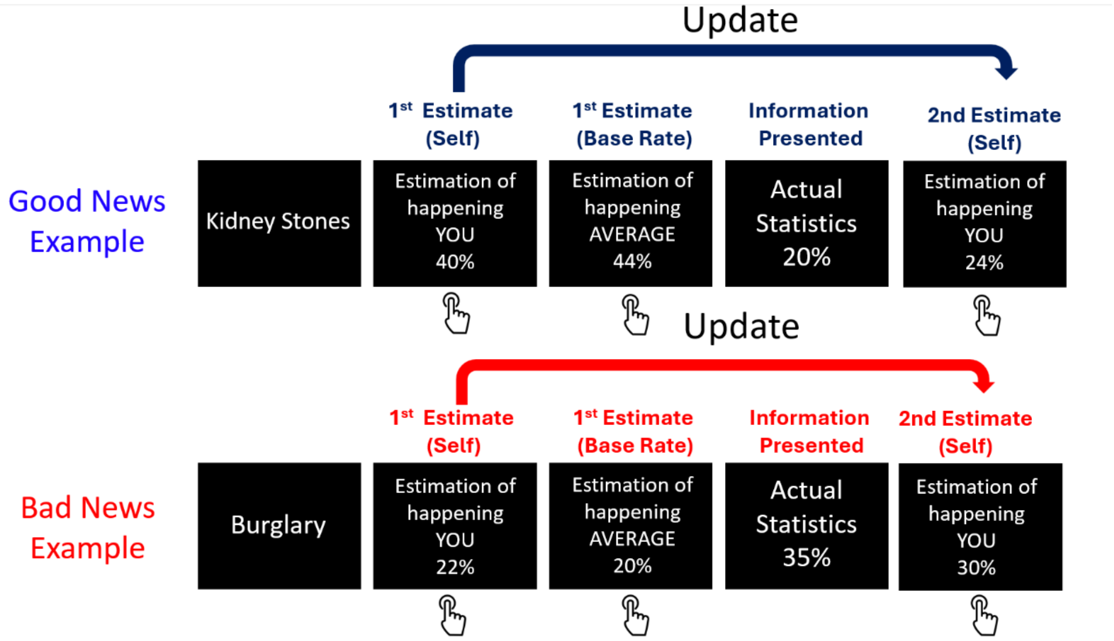


Figure 1.2: The Update Bias Task. In each trial, participants estimate their own future likelihood of experiencing an adverse life event (e.g., kidney stones, burglary) and that of someone like them (same age, location, and socioeconomic status). They are then presented with the actual probability of that event for someone like them. Finally, they are prompted to provide their estimates again. Displayed at top is an example of good news as the chances of kidney stones happening to the individual (20%) is *lower* than what they had thought (40%). Below is an example of bad news where the chances of burglary happening to the individual (35%) is *higher* than what they had thought (22%).

Learning in the UBT is driven by the "Estimation Error": the difference between the participant's first estimate and the true base rate (Sharot et al., 2011). This is similar to prediction error (PE) in reinforcement learning (RL) - the difference between expected and observed outcomes. Because of this shared principle of error-driven updates, the RL framework has been successfully adapted to build computational models of learning in the UBT (Kuzmanovic & Rigoux, 2017).

One of these models assumes that people process desirable and undesirable information differently. It formalizes this idea by using two learning rates: one for positive estimation errors (good news) and one for negative estimation errors (bad news). This model consistently provides the best fit for data, and the estimates from the model show a higher learning rate for good news than for bad news, which provides model-based evidence for an optimistically biased pattern of information integration (Kuzmanovic et al., 2018, 2019b; Kuzmanovic & Rigoux, 2017). The models help formally control for trial-wise fluctuations and provide insights into the mechanistic components of belief updating by including estimation error size and the relative personal knowledge (rP) in the updating process. rP captures how much an individual sees themselves as different from the average person and is calculated based on differences between their initial personal estimate (E_1) and their estimate of the base rate (eBR). Therefore, the fact that the optimistic update bias persists even after the model accounts for these components is a testament to its robustness. This computational approach provides a formal, mechanistic account of how the selective integration of true information leads to the formation and maintenance of false beliefs. In what follows, I will describe the neural findings on this bias to better understand how false beliefs are formed.

Distinct neural processes appear to mediate the integration of good and bad news. Kuzmanovic et al. (2018) implicated the ventromedial prefrontal cortex (vmPFC) in encoding the valence of belief updates in UBT. Specifically, the fMRI analyses revealed neural correlates for different stages of belief updating. During the presentation of the actual BR, regions like the anterior cingulate cortex (ACC), inferior frontal gyrus (IFG), and dorsolateral prefrontal cortex (dlPFC) tracked errors, with their activity increasing as error size decreased. Notably, the magnitude of this error tracking in the dlPFC correlated with individual learning rates, suggesting its role in adjusting initial beliefs based on errors. The vmPFC activity increased with larger updates towards lower risks (good news) and with smaller updates after bad news. This vmPFC activity specifically tracked the *improvement or worsening of final beliefs relative to initial ones*, not merely the valence of the new information or final beliefs themselves. Furthermore, the magnitude of this vmPFC valence-tracking effect correlated with the individual's optimism bias, indicating that vmPFC is sensitive to the subjective value of favourable belief updates. Importantly, this valence encoding in vmPFC occurred during the period of update consideration, not earlier during the reception of the new information. Dynamic Causal Modeling (DCM) was used to infer the causal interactions within the update circuit, consisting of the dlPFC (an "update processing" node receiving exogenous input), the vmPFC (a "valuation" node), and the dorsomedial prefrontal cortex (dmPFC, a "cognitive" node). The winning DCM model revealed a cyclic information flow from the dlPFC via vmPFC to the dmPFC, with a valence-dependent modulation of the coupling from dlPFC to vmPFC. This indicated that the vmPFC actively filtered incoming information based on its valence. Both the strength of this valence-dependent modulation of the dlPFC-vmPFC coupling and the strength of the vmPFC-dmPFC connection correlated with the individual's optimism bias. This suggests that a stronger optimism bias is associated with a greater valence-dependent filtering by the vmPFC and a stronger influence of this valuation system on the cognitive processing occurring in the dmPFC. In another study (Garrett et al., 2014) BOLD signal correlated positively with good news estimation errors in the left inferior frontal gyrus (left IFG) and bilateral superior frontal gyrus (bilateral SFG), but negatively with bad news estimation errors in the right inferior parietal lobule (right IPL) and positively in the Superior Temporal Gyrus and Superior Frontal Gyrus. Further, the study found that BOLD response in the right IPL of depressed participants tracked bad news errors with greater fidelity than in healthy controls. Additionally, a stronger negative correlation between BOLD activity in the right Inferior Frontal Gyrus (rIFG) and bad news estimation errors was observed in depressed patients compared to healthy controls. These findings indicate that the unbiased updating observed in MDD is mediated by stronger neural coding of estimation errors in response to both good news (left IFG, bilateral SFG) and bad news (right IPL, right IFG), particularly the adequate neural tracking of negative estimation errors.

This bias is flexible as it is modulated by psychological and environmental context. It is absent in individuals with clinical depression, who show more balanced belief updating (Garrett et al., 2014). Similarly, in situations of perceived threat, the bias is significantly reduced or eliminated, allowing for a more accurate risk assessment. This has been shown in studies of firefighters on duty (Garrett et al., 2018) and in the general population during the early, high-uncertainty phase of the COVID-19 pandemic (Beron et al., 2024). This flexibility could become costly when faced with false information. If the goal of the bias is to regulate mood and motivate action by filtering true information, it stands to reason that its operation - and its potential 'hijacking' by false information - will similarly depend on these contextual factors. Understanding the conditions under which the bias is weakened or strengthened for true information provides a theoretical roadmap for hypothesizing about when individuals will be most vulnerable to accepting desirable falsehoods.

1.3 Reinforcement Learning for Modelling Bias in Creating False Beliefs

1.3.1 The Rationale for a Reinforcement Learning Approach

Having established that humans filter true information to form false beliefs, I turn to computational modelling to formalize the learning process using the reinforcement learning (RL) framework. RL describes how living beings and artificial systems learn through experience to make better decisions, maximizing rewards and minimizing punishments (Rescorla & Wagner, 1972; Sutton & Barto, 2018). Here, I argue that RL is suited for building a mechanistic model of learning from false information and showing how it has previously been used to model confirmation and positivity biases in response to true information, leading to false beliefs. Further, RL provides a formal, quantitative model of the learning process itself. Instead of merely describing a phenomenon like "confirmation bias," a tractable RL model can operationalize it through specific, measurable parameters, such as distinct learning rates for confirmatory versus disconfirmatory evidence (Palminteri, 2023; Palminteri & Lebreton, 2022). This transforms a descriptive label into a testable, mechanistic hypothesis. It provides a benchmark of rational learning against which one can measure and characterise the systematic biases that could make individuals vulnerable to false information. This computational framework then provides specific, trial-by-trial latent variables - such as Prediction Errors (PEs) - that have a well-established neural basis. Numerous studies have linked these model-based estimates to activity in the dopaminergic midbrain and its targets in the striatum and ventromedial prefrontal cortex (vmPFC) (Collins et al., 2017; Daw et al., 2005, 2011; Jocham et al., 2014; Lefebvre et al., 2017; McDougle et al., 2019; Palminteri et al., 2012). This allows me to bridge the underlying computations to their neural bases. Finally, while my thesis uses controlled laboratory tasks to isolate these mechanisms, the core principles of RL are relevant in (mis)information-rich environments (da Silva Pinho et al., 2024).

1.3.2 The Reinforcement Learning Framework

Imagine an agent, whether an animal, a human, or a computer program, trying to achieve a goal. It doesn't receive explicit instructions on what to do; instead, it must discover which actions lead to desirable outcomes (rewards) and which lead to undesirable ones (punishments) through trial and error (Yoo & Collins, 2022). This trial-and-error process is at the heart of RL, an example of which is the two-armed bandit task (Figure 1.3). In this task, participants repeatedly choose between two options, each offering probabilistic outcomes. For example, one option gives 10 points with 80% probability, while another with 20% probability. Through accumulated feedback, participants learn the value of each option (Lefebvre et al., 2017; Palminteri, 2023; Sutton & Barto, 2018).

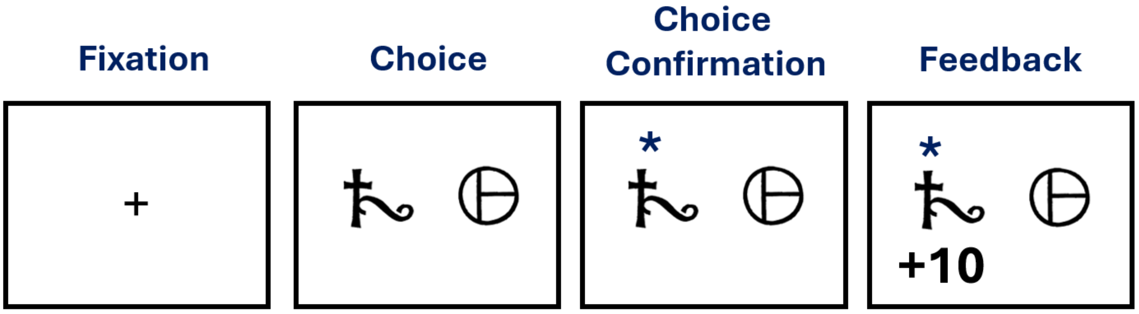


Figure 1.3: An example of a two-arm bandit task. Participants choose between two cues to maximize rewards. The goal is to learn which option is better (i.e., offers reward most of the time).

Learning happens through the continuous updating of expectations. Each option is assigned an "expected value" or "Q value," representing what the agent believes that option is worth. These expectations are then updated based on observed outcomes, which can be rewarding (e.g. +10) or punishing (e.g. -10). The difference between the expected outcome and the observed outcome is called Prediction Error (PE or δ) (Sutton & Barto, 2018). Formally, for a given trial (t), PE for Option A is written as:

$$\delta(t) = r_A(t) - Q_A(t)$$

In which r_A is the *observed* outcome for option A and Q_A is the *expected* outcome of the option. Positive PEs occur when outcomes are better than expected ($r_A(t) > Q_A(t)$) while negative PEs occur when they are worse than expected ($r_A(t) < Q_A(t)$).

Therefore, PEs serve as a learning signal to update future expectations. The extent to which PEs are used as a learning signal can be determined by a learning rate (α) parameter (Sutton & Barto, 2018). A high learning rate means the agent quickly adjusts its beliefs based on new information, while a low learning rate produces more gradual changes.

The value update follows this formula:

$$Q_A(t+1) = Q_A(t) + \alpha * \delta(t)$$

Since the two-armed bandit task involves comparing two options (A and B), their learned values must be converted into choice probabilities. One way to do the conversion is through the SoftMax function, which transforms the learned values into the probability of selecting each option: The probability of choosing option A, for example, is calculated as follows:

$$\frac{1}{1 + \exp(\beta(Q_B - Q_A)}$$

In which, β is the inverse temperature parameter that controls how sensitive the choice probabilities are to the differences in the learned values of the options (Sutton & Barto, 2018). The higher the value of β the greater the tendency to exploit (i.e., the agent is very sensitive to the learned values and will almost always choose the option with the higher value, even if the difference is tiny); and the lower the value of β the greater the tendency to explore (i.e., the agent is less sensitive to the value differences, and its choices become more random). This parameter thus controls the exploration-exploitation trade-off in action selection (Sazhin et al., 2025; Wilson et al., 2014).

The above model with two free parameters, a learning rate and an inverse temperature, is a simple computational model that produces trial-by-trial latent estimates of PE and Q values. The latent estimates can then be used in neural studies, allowing researchers to link computational processes to brain activity (Collins et al., 2017; Daw et al., 2005, 2011; Hare et al., 2008; Huys et al., 2011; Jocham et al., 2014; Lefebvre et al., 2017; O'Doherty et al., 2007; Zhang et al., 2020).

1.3.3 Neural Correlates of Reinforcement Learning Computations

A key strength of the RL framework is that its core computational variables have well-established neural correlates. Foundational work has demonstrated that the firing of midbrain dopamine neurons closely tracks the Reward Prediction Error (RPE), increasing for better-than-expected outcomes and decreasing for worse-than-expected outcomes (Schultz et al., 1997). This dopaminergic signal is broadcast to the striatum, where it is thought to drive the synaptic plasticity necessary for updating the values of actions and states (Wickens, 2009). However, this classic model has been refined, with recent work showing that dopamine's role is more complex, also encoding the motivational value of an action and acting as many local, "partial" teaching signals within specific corticostriatal circuits rather than as a single, global broadcast (Berke, 2018; Lak et al., 2020) .

While the striatum and its dopaminergic inputs are crucial for computing the error signal, a distinct but overlapping network, centred on the ventromedial prefrontal cortex (vmPFC), is responsible for representing the value of different choices to guide decisions (Boorman et al., 2009; Kable & Glimcher, 2007). Activity in the vmPFC has been shown to track the subjective value of chosen options, integrating different attributes of a choice into a "common currency" signal that can be used for comparison (Levy & Glimcher, 2012). This valuation system is not static; for example, under time pressure, the vmPFC represents the overall value of all available options, but with more time for deliberation, its activity shifts to represent the specific value difference between the chosen and unchosen option (Jocham et al., 2014).

This neural architecture for value-based learning is the fundamental system through which beliefs, including false beliefs, are likely formed and updated. Critically, its role extends beyond processing external rewards like money or food. A body of research now demonstrates that these same neural circuits - particularly the dopaminergic midbrain, striatum, and orbitofrontal/ventromedial prefrontal cortex - are also responsible for encoding the intrinsic value of information itself (Bromberg-Martin & Monosov, 2020). Monkeys, for instance, will “pay" by forgoing a guaranteed juice reward for the mere opportunity to gain advance information about future rewards, and their dopamine neurons track this information-seeking preference just as they would a primary reward (Bromberg-Martin & Hikosaka, 2009). This indicates that the brain's valuation circuitry treats the resolution of uncertainty as inherently rewarding.

This neurobiological finding - that the brain treats information as a form of currency processed by its core reward system - provides a mechanistic rationale for applying a value-based learning framework to the problem of false information. If the brain is wired to seek out and assign value to information, the unaddressed neurocomputational question becomes: what happens to this valuation process when the currency is counterfeit and false? How does the brain's valuation and learning circuitry respond when the information it receives is known to be false? Understanding the baseline neural mechanisms for processing true information is therefore important for investigating how those mechanisms are altered and potentially hijacked in the face of falsehoods.

1.3.4 Reinforcement Learning and Working Memory

The reinforcement learning system does not operate in isolation. Its interaction with working memory (WM) is indispensable, even in simple instrumental tasks (Collins, 2018; Yoo & Collins, 2022). This relationship is potentially important when encountering unreliable or false information. Efficient learning in such a context requires actively filtering out false feedback before it can erroneously one's beliefs. This mirrors the cognitively demanding nature of modern online environments, where users must constantly ignore or discount information from bots and unreliable accounts (Chuai et al., 2023b; Pittman & Haley, 2023). One possibility is that a reason why people learn from false information is a failure of this WM-dependent filtering mechanism. When cognitive control is taxed or when a piece of false information is particularly compelling (e.g., it confirms a prior belief), this filter may fail, allowing the falsehood to be processed by the RL system as if it were true. To understand the nature of this potential failure, it is first necessary to detail the relationship between WM and RL.

WM is a capacity-limited process for temporarily holding and manipulating information to guide behaviour, particularly when that information is no longer present (Yoo & Collins, 2022). The limitations of WM include a finite capacity and a temporal limit, meaning information is only accurately remembered for a short duration (Collins, 2018). These limitations are well-documented by classic findings. For example, as the number of items held in memory grows, accuracy and reaction time decrease, a phenomenon known as the set size effect (Sternberg, 1966). Further, these mental representations are inherently tenuous, susceptible to fading over time or being disrupted by distracting stimuli (Peterson & Peterson, 1959).

Although often conceptualized as a separate cognitive module, WM operates in concert with RL, an interaction that is evident across behavioural, computational, and neural levels of analysis (Yoo & Collins, 2022). This relationship is bidirectional. On one hand, WM supports RL by maintaining stimulus information (e.g., in partially observable Markov decision processes) or even reward information itself, feeding these inputs into RL computations. This assistance allows the RL system to form more abstract, task-relevant representations, which in turn promotes generalization to new contexts and simplifies the learning problem by effectively ignoring irrelevant aspects of the environment (Yoo & Collins, 2022). Similarly, RL can influence WM. It has been demonstrated, for example, that people can acquire more efficient strategies for using their WM through feedback-based learning. This finding implies that a trial-and-error process, guided by reinforcement, helps to optimize how WM is recruited (Yoo & Collins, 2022).

This dynamic leads to a surprising trade-off, often called the "tortoise and hare" effect (Collins, 2018). In low WM load problems (e.g., fewer stimuli to learn) where WM capacity is sufficient (the fast "hare"), individuals learn quickly. However, this comes at the cost of poorer long-term retention because the slower, more robust RL system (the "tortoise") is less engaged (Collins, 2018; Collins et al., 2017). Collins (2018) showed this in an experimental protocol including a learning phase, an unrelated n-back task serving as a delay, and a surprise testing phase. During the learning phase, participants learned stimulus-action associations by receiving feedback, with varying set sizes (ns=3 for low load and ns=6 for high load) across 14 blocks. The set size manipulation was crucial because WM is capacity-limited, unlike RL, allowing researchers to disentangle their contributions. The surprise testing phase, conducted without feedback after a 10-minute delay and involving 54 different stimulus-action associations, was designed to assess the retention of learned associations, specifically probing RL function as WM was assumed to play no direct role due to its temporal and capacity limits. Behavioural results replicated previous findings for the learning phase: participants learned in both set sizes, but learning was slower and performance was lower in blocks with a high set size (ns=6) compared to low set size (ns=3). This demonstrated WM's contribution, characterized by negative effects of set size and delay on performance, while RL was evident through sensitivity to reward history. However, the associations learned under high set sizes were better retained than those learned under low set sizes, indicating greater robustness of learning under high cognitive load. This meant that using working memory to learn quickly came at the cost of long-term retention. Further analysis confirmed this was not due to differences in reward history or error avoidance. To understand these interactions, computational modelling was employed, testing three families of models: pure RL models (RLs), mixture models with independent WM and RL (RLWM), and mixture models with interacting WM and RL (RLWMi). The RLWMi model, which assumed WM influences RL computations by contributing to PE calculations, was strongly favoured as the best fit for the behavioural data. This model posits that when WM learns faster than RL in low-set-size scenarios, it effectively decreases positive PEs, thereby impeding learning within the RL system. This is bolstered by a neural observation that brain signals related to RL encoding are weaker at lower set sizes (Collins et al., 2017). Ultimately, RL and WM can be viewed as partially redundant systems that learn with different dynamics: WM is fast but fleeting, while RL is slower but more robust (Collins, 2018). Therefore, cognitive load is a key factor that modulates the balance between these two learning systems such that under high load imposed by environments containing false information, such as social media, the slow-but-steady RL system may become more dominant, potentially making it more vulnerable to integrating false information.

The link between RL and WM is mirrored by a significant overlap in their underlying neural circuits (Yoo & Collins, 2022). While the prefrontal cortex (PFC) is associated with WM and the basal ganglia with RL, these networks are not entirely distinct. Instead, they are connected through multiple parallel loops, with the frontal cortex and basal ganglia projecting directly onto each other (Haber, 2011). This shared architecture has functional consequences. The PFC, for example, is implicated in many goal-directed RL tasks (Daw et al., 2005, 2011). Furthermore, dopamine levels in the PFC, a neuromodulator central to RL, are also related to WM performance (Fallon et al., 2015). This is evident in how damage to the basal ganglia can produce cognitive impairments similar to those caused by frontal cortex damage (Middleton & Strick, 2000). This neural evidence confirms that RL and WM are not separate brain modules but deeply integrated systems that work together to produce intelligent, adaptive behaviour.

1.3.5 Using Reinforcement Learning to Model Positivity Bias

RL has also been applied to formalize two learning biases: positivity bias and confirmation bias. Positivity bias is when an agent learns more from better-than-expected outcomes (positive prediction errors) than from worse-than-expected outcomes (negative prediction errors) (Palminteri & Lebreton, 2022). For instance, in one RL study (Lefebvre et al., 2017), where participants played a 2-arm bandit task (Figure 1.1) involving abstract cues that gave rewards and punishments for the chosen option, higher learning rates were observed for positive PEs versus negative ones. To model behaviour, the researchers compared a standard Rescorla-Wagner (RW) model with a modified version (RW±) that allowed for different learning rates for positive and negative PEs. The findings supported the hypothesis of a general learning asymmetry: the RW± model provided a better fit for subjects' behaviour. Further, the learning rate for positive PEs was significantly higher than for negative PEs, indicating that participants preferentially updated values following better-than-expected outcomes for the item they chose. This asymmetry was primarily driven by subjects categorized as "RW± subjects," who exhibited a significantly reduced negative learning rate compared to "RW subjects" who displayed unbiased learning. The positivity bias ($\alpha+ > \alpha-$) was replicated in the second behavioural experiment, where punishment replaced reward omission (i.e., loss pairs instead of gain pairs), indicating that the learning asymmetry is driven by the valence of the prediction error itself, not solely by the outcome valence (i.e., in a loss pair of -10 and -1, a -1 outcome would incur a positive PE despite having a negative outcome valence). The fMRI results from the same study (gain pairs only) revealed that PE encoding in the brain's reward circuitry, specifically the striatum and vmPFC, was enhanced in optimistic (RW±) subjects. This neural activity positively correlated with the learning rate asymmetry, establishing an association between the positivity bias and brain activity when outcomes are revealed. The individual differences in positivity bias were also linked to pupil dilation – a physiological proxy of neuromodulator activity - such that positive PEs increase the dilation and the negative ones increase the constriction (Van Slooten et al., 2018). Further, higher dopamine is associated with a stronger positivity bias in Parkinson's disease patients ON medication vs OFF in the dorsal striatum (McCoy et al., 2019), which is in line with the fMRI results of the study, as dopamine is sensitive to positive and negative PEs (Frank et al., 2004; Palminteri et al., 2009). Other studies show that the bias is robust to different outcome ranges (Ting et al., 2022) and outcomes of a different nature (electric shocks) (Gagne et al., 2020). It has also been observed in rhesus monkeys (Farashahi et al., 2019), rodents (Harris et al., 2021; Ohta et al., 2021), foraging (Garrett & Daw, 2020) , multi-attribute RL (Steinke et al., 2020), and large language models (LLMs) such as Claude, ChatGPT, and Llama (Hayes et al., 2025). The models' positivity bias comes from their training on human language, suggesting this bias is important in how we communicate (Palminteri, 2025a).

1.3.5.1 The Role of Positivity Bias in Learning from False Information

The positivity bias could also offer a computational mechanism for belief formation when faced with false information. A recent study has already provided compelling evidence for this (Vidal-Perez et al., 2025). The study employed a novel "disinformation" version of the two-armed bandit, where individuals repeatedly chose between two options and received the outcome for the chosen option from computer-programmed "feedback agents" who varied in their truthfulness. Participants were explicitly informed about the credibility of each agent via a "star system": a 3-star agent was always truthful, a 2-star agent lied 25% of the time (75% truthful), and a 1-star agent lied 50% of the time (50% truthful), making its feedback statistically random. On each trial, participants were first told which agent will give the feedback and then made their decision. Participants were incentivized based on true bandit outcomes, not agent-provided feedback. The results indicated an exacerbated "positivity bias", where individuals boosted their learning from positive feedback relative to negative feedback. This bias was found to be amplified for information of low and intermediate credibility. When measured in relative terms, the positivity bias was significantly higher for the 1-star and 2-star agents compared to the 3-star agent. This bias could not be accounted for by Bayesian strategies, which instead predicted a negativity bias, nor could it be fully explained by choice perseveration. The researchers posited that feedback of ambiguous veracity might enable individuals to interpret positive feedback as true (as it confers desirable outcomes) and explain away negative feedback as false. This provides a formal, mechanistic account of how individuals might maintain an overly optimistic view of their choices by systematically overweighting desirable falsehoods and underweighting the undesirable ones.

In the two-arm bandit studies mentioned so far, partial feedback is given, meaning that when you choose an option, you get an outcome only for that option; therefore, it is not clear whether people are learning more from positive vs negative prediction errors or confirmatory vs disconfirmatory outcomes (Lefebvre et al., 2017). In other words, is the learning bias driven purely by the outcome's valence (i.e., all positive prediction errors are overweighed) or a confirmation bias (i.e., only positive PEs following obtained outcomes are overweighed)? To answer this question, we need a design that offers outcomes for the unchosen option as well (Palminteri & Lebreton, 2022), which the studies in the next section have offered.

1.3.6 Using Reinforcement Learning to Model Confirmation Bias

Confirmation bias differs from optimistic update bias or positivity bias in that valence is not the only determining factor in integrating a piece of information; rather, whether it aligns with one’s prior beliefs, choices, judgements, and decisions or not is as important (Nickerson, 1998; Palminteri & Lebreton, 2022). While distinct in their definitions - positivity bias relates to the valence of new evidence, and confirmation bias to alignment with prior beliefs - these two biases frequently co-occur in real-world scenarios. This is typically because people hold opinions and make choices that they anticipate will lead to positive subjective outcomes. Consequently, an outcome that is better than expected often simultaneously provides positive news and confirms a prior decision or belief (Palminteri & Lebreton, 2022). However, the two can be dissociated in a controlled experiment. Consider a two-arm bandit task with complete feedback, where one sees the outcome of both the chosen and the unchosen option. An outcome can be "confirmatory" - meaning it provides evidence that your choice was correct - in two ways. First, if your chosen option yields a reward (a positive PE). Second, and more subtly, if the option you *did not choose* would have resulted in a loss (a negative PE). While this second case is technically "bad news" about the unchosen option, it is motivationally "good news" for the decision-maker as it validates their choice. The key finding is that people learn more from both types of confirmatory evidence, demonstrating a learning process that is biased towards validating prior decisions, not just towards seeking positive outcomes (Palminteri et al., 2017; Palminteri & Lebreton, 2022).

Within the RL framework, several studies have shown the existence of confirmation bias, where people integrated confirmatory information to a greater extent than disconfirmatory one (e.g., Chierchia et al., 2023; Palminteri et al., 2017). The key manipulation in these studies was offering complete feedback (i.e., outcomes shown for chosen and unchosen cues), specifically aiming to distinguish between a general "positivity bias" and a "confirmation bias". To model the behaviour, first, the researchers used a modified Rescorla-Wagner computational model that allowed for different learning rates for positive and negative PEs for both chosen (“factual learning”) and unchosen (“counterfactual learning”) outcomes. Replicating previous findings, they found a positivity bias in factual learning, where participants preferentially learned from outcomes that were better than expected. Specifically, the learning rate for positive PEs from chosen options (α_{c+}) was significantly higher than for negative PEs (α_{c-}). However, the results for counterfactual learning revealed an opposite valence-induced bias: unchosen negative PEs drove stronger learning than unchosen positive PEs, as the learning rate for negative unchosen PEs (α_{u-}) was higher than for positive unchosen PEs (α_{u+}) (Figure 1.4 (a)). The pattern of results supported a confirmation bias in learning, suggesting that, on one hand, people are more sensitive to information that validates the decision; on the other hand, individuals tend to discount evidence that suggests their choice was wrong, such as a negative PE from their chosen option or a positive PE from the one they rejected. Then, they created a more parsimonious model called the Confirmation Model by collapsing Positive-Chosen and Negative-Unchosen learning rates into one learning rate called Confirmatory and Negative-Chosen and Positive-Unchosen into one learning rate called Disconfirmatory. Model comparison showed this model provided the best fit for the data, surpassing the four-learning-rate model with separate learning rates for the chosen and chosen options. Further, its estimates revealed that the Confirmatory learning rate was significantly higher than the Disconfirmatory one (Figure 1.3 (b)). This also suggested that factual and counterfactual outcomes might be processed by the same underlying learning systems. The superiority of this model and the asymmetrical pattern of the learning rates have been replicated in other studies (Lebreton et al., 2019; Palminteri, 2023; Schüller et al., 2020).

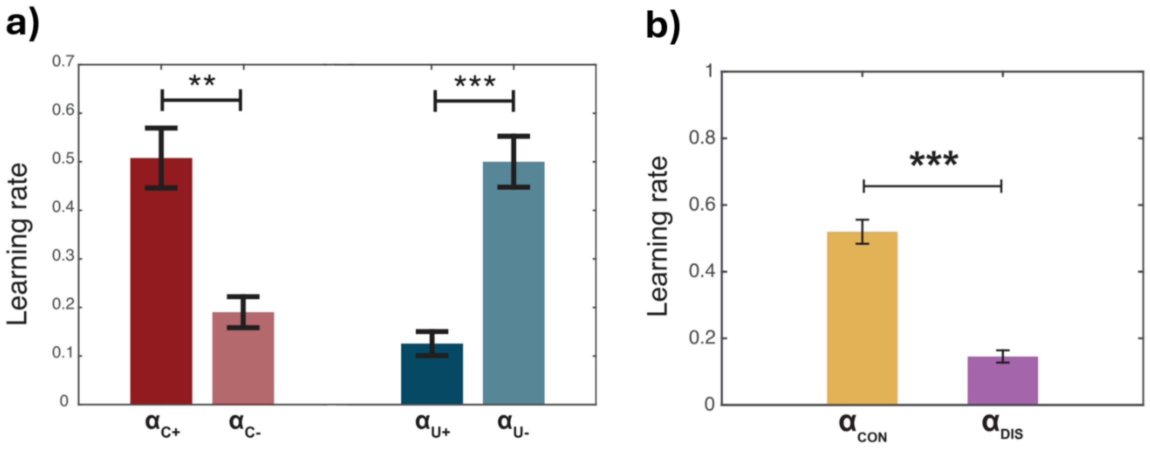


Figure 1.4: Positivity vs Confirmation. The pattern of learning rates for positivity and confirmation biases (Palminteri, Lefebvre, et al., 2017). a) Although there is positivity bias for the chosen options ($\alpha_{c+} > \alpha_{c-}$), this pattern is reversed for the unchosen options ($\alpha_{U+} < \alpha_{U-}$), consistent with confirmation bias. b) The estimates of the Confirmation Model where α_{CON} (learning rate for positive obtained and negative forgone outcomes) is significantly higher than α_{DIS} (learning rate for negative obtained and positive forgone outcomes).***P<0.001 and *P<0.05, two-tailed paired t-test. *Reproduced from Palminteri et al. (2017), "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing," PLOS Computational Biology 13(8): e1005684, licensed under CC BY 4.0.*

The robustness of this bias has been questioned by researchers who point out that choice perseverance can create a statistical artifact that mimics confirmation bias (Katahira, 2018; Sugawara & Katahira, 2021). To resolve this, they developed a "Hybrid model" that accounts for both asymmetric learning rates and gradual perseverance (where multiple previous choices influence the current one). When they applied this model to their own data, they found that perseverance appeared to be the dominant factor driving behaviour. In response to these claims, Palminteri (2023) re-examined data from nine separate experiments, encompassing over 126,000 trials from 363 individuals. This re-analysis compared a model that only included asymmetric learning with a "full" model that also incorporated the gradual perseverance term. The reanalysis revealed two key findings. First, the inclusion of the gradual perseveration significantly reduced the estimated confirmation bias. Second, even after accounting for gradual perseveration, the confirmation bias remained present at the meta-analytical level and was significantly different from zero in most experiments. This robust presence indicates that confirmation bias is a reliable feature of human reinforcement learning, not simply a byproduct of the tendency to repeat choices.

1.4 The Adaptive Value of The Biases

If these biases are suboptimal, why would evolution allow them to persist? One could argue that they were favoured by evolution because they are adaptive and have ecological rationality, meaning that they confer real-world advantages that outweigh the costs of deviating from perfect logic (Palminteri, 2025b). Supporting this argument are studies on optimism that show maintaining an optimistic outlook can be inherently rewarding, fostering a sense of self-competence, personal growth, positive emotions, reduced stress, and a sense of control over outcomes (Chang, 2001; S. E. Taylor & Brown, 1988). Further, the absence of optimistic update bias in belief updating is observed in individuals with clinical depression (Garrett et al., 2014) and precede clinical manifestations of relapse in bipolar disorder (Ossola et al., 2020), suggesting its importance for mental health.

The adaptive side of the optimistic update bias has been shown within the RL framework as well. In a simulation study, for instance, it has been shown that an agent learning in a biased manner – positivity bias - can objectively outperform an "unbiased" agent in certain probabilistic learning tasks, particularly in low-rewarding environments or when payouts are rare (Cazé & van der Meer, 2013). Another study using evolutionary simulations showed that positivity bias is evolutionary stable (Hoxha et al., 2024). The core of its methodology was an evolutionary algorithm designed to identify the optimal set of parameters for each environment. A population of 1000 agents, each with a unique set of parameters, performed the task for 200 "generations". In each generation, an agent's fitness was determined by its performance. The bottom 5% of performers were eliminated, while the top 5% were duplicated, ensuring the population size remained constant and that successful "genotypes" propagated over time. Agents with positivity bias evolved across different two-armed bandit scenarios, including volatile environments where reward probabilities change.

Similar findings have been reported for the simulation studies of confirmation bias. Studies conducted in a range of learning environments – e.g., stable, volatile, rich, and poor - have indicated that confirmation bias optimizes reward learning, with biased agents outperforming their unbiased counterparts (Kandroodi et al., 2021; Lefebvre et al., 2022; Tarantola et al., 2021). This counterintuitive result is explained by the bias mechanistically neglecting uninformative, stochastic negative PEs associated with the best response, leading to more efficient resource accumulation and reward collection (Palminteri & Lebreton, 2022). Furthermore, confirmation bias can improve decision-making in the presence of noise (Lefebvre et al., 2022, 2024). By making subjective action values more extreme (overvaluing good options and undervaluing bad ones), it increases the relative distance between options, thereby making decisions more robust to random fluctuations and increasing the probability of selecting the better option in subsequent trials. It has also been shown to enhance collective decision-making in reinforcement learning agents (Bergerot et al., 2024). Finally, this bias can be advantageous when paired with efficient metacognition, as it allows for the neglect of probabilistic negative feedback that sometimes inevitably follows correct choices, creating a normative basis for positivity and confirmation biases (Rollwage & Fleming, 2021).

But how can these computationally advantageous biases also contribute to vulnerability to false information? The answer could lie in the dramatic shift in the information environment. Biases like confirmation and positivity are adaptive when filtering a world that is noisy but reliable - they help an agent maintain a stable course and avoid overreacting to random negative outcomes. However, the modern information ecosystem is not just noisy; it contains deliberately crafted false information (disinformation) that is built to exploit these very mechanisms (Aïmeur et al., 2023). In this new context, the adaptive machinery could be a liability. The confirmation bias, which is useful for ignoring a single bad outcome from a generally reliable food source, becomes a liability when it causes an individual to ignore a well-sourced factual correction that contradicts a desirable political falsehood.

1.5 The Gap: The Need for Computational Models of Learning from False Information

So far, I have established that the brain selectively filters *true* information, creating false beliefs, such as an overly optimistic view of the future. The framework of Reinforcement Learning provides a mechanistic account of *how* these biases operate, formalizing hypotheses with different computational models that give insight into the reasons behind the formation of false beliefs, such as the asymmetric treatment of prediction errors (or estimation errors in the UBT). I propose using the same computational approach to build models of learning from false information because our understanding of the underlying computations involved in learning from false information and in the success or failure of interventions like debunking is lacking. For instance, the very existence of the Continued Influence Effect (Ecker et al., 2010) - where a debunked falsehood continues to shape reasoning - reveals that the process of belief updating is not a simple matter of replacing one fact with another. The computations involved in a debunking event are often treated as a "black box." We can measure the input (a correction) and the output (a change in belief), but we lack a formal model of how that correction is processed, how it competes with the original falsehood in memory, and what factors determine the degree of learning or belief change. To open this black box and move toward developing more informed interventions, we should first develop mechanistic models of learning from false information. Also, this would allow me to answer a key question: do the biases detailed so far, such as the confirmation bias, persist when faced with false information? I hypothesize that one of the reasons for vulnerability to learning from *false* information is the very same biases used to process *true* information.

1.6 Aim and Outline of the Thesis

The first goal is to characterise the computational underpinnings of learning from false information in two different environments. I will start with a chapter on computational modelling (**Chapter 2**), detailing and justifying the modelling approaches I have used in the thesis. **Chapter 3** will modify the classic two-armed bandit task to assess if people learn from false information and whether they do so in a biased manner – be it positivity or confirmation bias – just as they do in learning from true information (Palminteri & Lebreton, 2022). Building on previous work (Garrett et al., 2014; Garrett & Sharot, 2014, 2017a; Kuzmanovic et al., 2019a; Kuzmanovic & Rigoux, 2017; Sharot & Garrett, 2016), **Chapter 4** will adapt the UBT to include explicitly given statements about information accuracy as a variable, testing whether people learn from information explicitly described as false, and if so, whether the well-documented optimistic update bias persists when people encounter false information. Similarly,

Following these behavioural and computational investigations, the second goal is to assess the neural underpinnings of this process, which represents a significant gap in the current literature. To my knowledge, no study has examined how the brain processes false information within an RL framework using fMRI. Given that the striatum and vmPFC are consistently implicated in RL studies (Daw & Tobler, 2013; Fouragnan et al., 2018; Lefebvre et al., 2017) and the fact that information itself can be rewarding (Bromberg-Martin & Monosov, 2020) It may be that the striatum and vmPFC are also involved in processing false information. Therefore, **Chapter 5** will detail an fMRI study that uses the modified two-armed bandit task from **Chapter 3** to elucidate the neural mechanisms involved in learning from both true and false information.

Chapter 2: Computational Modelling

Computational modelling serves as a bridge between qualitative psychological theory and quantitative behavioural, simulated, and physiological data (e.g., neural recordings, pupil dilation, etc). By requiring theories to be instantiated as mathematical formulations, modelling forces theoretical assumptions to be made explicit which in turn can reveal theoretical ambiguities in the process (Guest & Martin, 2021).

The process of Parameter Estimation (which I outline in greater detail later in this chapter) - fitting a specific model to data such as participants choices in a decision-making task - results in a set of parameters for each participant that best describe their data. These “best fit” estimated parameters can provide insights into individual differences (Montague et al., 2012). For example, variations in a learning rates (which characterise the rate at which beliefs change following new evidence) have been associated with mood and anxiety disorders (Pike & Robinson, 2022). Specifically, higher and more volatile learning rates from negative outcomes have been observed in individuals with anxiety, potentially reflecting a cognitive mechanism of over-weighting recent, adverse events. Similarly, differences in a decision-temperature parameter (which characterises sensitivity to differences in value) can map onto traits like impulsivity; a lower temperature parameter, reflecting more stochastic or 'noisy' decision-making, is often observed in individuals with higher trait impulsivity (Maia & Frank, 2011). These associations move beyond simply correlating a symptom with behaviour, providing a falsifiable hypothesis about the underlying computational mechanism that may generate that symptom, which could be a target for therapeutic interventions.

The process of model comparison, when implemented correctly and verified with simulations (which I also outline in greater detail later in this chapter) allows the researcher to quantitively compare competing mathematical accounts of *how* a process might occur. For instance, a standard experiment might find that participants perform better in high-reward contexts - a simple directional effect. However, modelling allows one to ask how this arises and compare different possibilities in terms of how well they each explain the observed data. Is it because the learning process itself is amplified by reward magnitude, with larger rewards increasing the learning rate? Or is the learning process constant, while the expression of that learning in choices becomes more precise, with the prospect of high rewards leading to less noisy decisions (a higher choice-temperature parameter)? These competing accounts can be instantiated as distinct models. By formally comparing them, one can move from asking if a phenomenon occurs to understanding how it occurs, providing a more rigorous testing ground for theories (Farrell & Lewandowsky, 2018).

In the current thesis, the specific models vary by chapter but the methodological approaches for model fitting, comparison, recovery, and validation are consistent throughout (Figure 2.1). In this chapter, I detail these approaches, abiding by the best practices (Wilson & Collins, 2019).

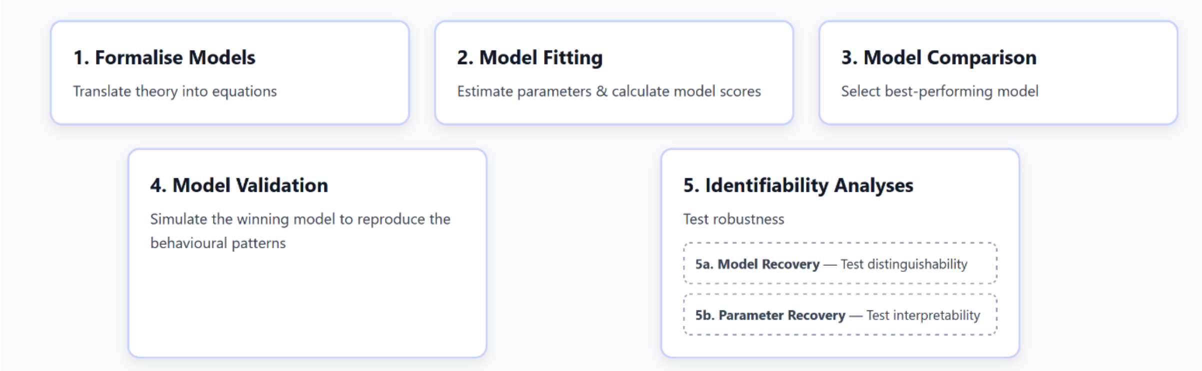

1. Formalise Models
Translate theory into equations
2. Model Fitting
Estimate parameters & calculate model scores
3. Model Comparison
Select best-performing model
4. Model Validation
Simulate the winning model to reproduce the behavioural patterns
5. Identifiability Analyses
Test robustness
5a. Model Recovery — Test distinguishability
5b. Parameter Recovery — Test interpretability

Figure 2.1: The Computational Modelling Workflow.

2.1 Model Fitting

The goal of model fitting is to find the parameter values for a given model that maximize the likelihood of the observed behavioural data. For this, I used a hierarchical approach using the Expectation-Maximization (EM) algorithm (Huys et al., 2011). I used a publicly available package (https://github.com/ndawlab/em/tree/master) written by Nathaniel Daw to implement this algorithm in the Julia programming language (version 1.9.4) (Bezanson et al., 2017) which has been successfully used in a number of computational modelling studies previously (e.g., Garrett & Daw, 2020; Nussenbaum et al., 2025).

EM is an iterative method for finding maximum a posteriori (MAP) estimates in models with latent variables - in this case, the individual-subject parameters. A more traditional approach would be to fit each participant's data independently, finding the best-fitting parameter set for each person in isolation. However, this method is problematic for two reasons. First, it treats all individual estimates as equally reliable, giving the same weight to a parameter derived from noisy, sparse data as to one derived from clean, consistent data (Huys et al., 2011). Second, because it fails to account for the uncertainty in each point estimate, it can produce extreme and psychologically implausible parameter outliers, especially for noisy participants (Ahn et al., 2011). These unreliable estimates can in turn inflate the overall variance at the group level and obscure true effects.

The hierarchical approach I employed using EM addresses these issues through the advantage of regularization (or "shrinkage"). Instead of fitting in isolation, the model assumes that while each participant has a unique set of parameters, these are drawn from a common group-level distribution - a Gaussian distribution defined by a group mean and variance. By doing so, the model "borrows statistical strength" from the entire group to inform each individual's parameter estimate, down weighting the influence of unreliable participants (Morris, 1977). The unreliable estimates from these participants are gently pulled toward the more stable group mean. This process effectively manages the bias-variance trade-off, reducing the variance of individual estimates at the cost of a small amount of bias. This bias is introduced at the individual level - an estimate for a participant whose true parameter value is far from the population average will be pulled, or biased, toward that mean - but it allows for a more stable and accurate estimation of the group-level distribution as a whole. This approach has been shown to yield superior predictive performance on unobserved data compared to fitting each participant independently (Scheibehenne & Pachur, 2015).

The EM algorithm finds these hierarchical estimates by alternating between two steps until convergence: Expectation and Maximization. The process begins with an initialization step, where I provided the algorithm with plausible starting values for the group-level parameters (the population mean, β, and covariance, Σ). These initial values represent the prior beliefs about the population before observing the data. The algorithm then iterates between the two main steps. The Expectation (E) step essentially asks: "Given our current belief about the population, what are the likely parameters for each individual?". It uses the current group-level parameters (β, Σ) as a prior to find the Maximum A Posteriori (MAP) estimate of each subject's individual parameters, x_i, by maximizing the log posterior probability:

$$x_i^{MAP} = \arg\max_{x_i}\left(\log P\,(y_i|x_i) + \log P\,(x_i|\beta, \Sigma)\right)$$

where y_i is the data for subject *i*. It then computes a Gaussian approximation to the full posterior distribution at that MAP estimate, characterized by its mean (x_i^{MAP}) and its variance (the inverse of the Hessian matrix, h_i, at the peak). The Maximization (M) Step then asks: "Given these individual parameter distributions, what is the most likely population distribution?". It uses the sufficient statistics from the E-step - the individual MAP estimates and their posterior variance - to update the group-level parameters. The group means, β, are updated based on the subject-level estimates, while the group covariance, Σ, is updated based on the squared errors and the average posterior variance from the E-step. These steps are repeated iteratively until the parameter estimates stabilize, indicating convergence on the most likely set of hierarchical parameters.

2.2 Model Comparison

Following fitting, different competing models were formally compared to determine which provided the most parsimonious and generalizable account of the data. This step attempts to balance goodness-of-fit and model complexity. Without penalizing for complexity, a more complex model will always fit the data better, but this can lead to overfitting - a scenario where the model captures idiosyncratic noise in the current dataset rather than its underlying structure, resulting in poor predictions for new data (Pitt & Myung, 2002). To guard against this, the model comparison process is guided by the principle of parsimony (also known as Ockham's razor), which favours the simplest model that can adequately explain the data (Burnham & Anderson, 2002).

A common method for this in non-hierarchical contexts is the Bayesian Information Criterion (BIC), which provides an approximation to the model evidence (Schwarz, 1978). It is calculated as:

$$BIC = k \ln(n) - 2 \ln(L)$$

where L is the maximized likelihood of the model, k is the number of free parameters, and n is the number of data points.

The penalty for model complexity is the explicit kln(n) term. For a more complex model to be favoured, its improvement in log-likelihood must be large enough to outweigh the penalty incurred by its additional parameters.

While influential and widely used (Lefebvre et al., 2017; Pitt & Myung, 2002), standard information criteria like BIC are ill-suited for the hierarchical models. Their calculation requires a single, unambiguous value for the number of free parameters and the number of data points, both of which are difficult to define in a hierarchical context where parameters exist at both the individual and group levels (Vehtari et al., 2017). It should be noted that specialized variants, such as the integrated Bayesian Information Criterion (iBIC), have been developed to approximate the model evidence in a hierarchical context (Stephan et al., 2009). This approach works by first integrating out the individual-subject parameters using a Laplace approximation to get a single marginal likelihood for the entire dataset and then applying a BIC penalty based on the number of *group-level* parameters. However, this method is still an approximation.

Given these limitations, I chose leave-one-out cross-validation (LOOcv) as my model comparison metric. This is a method for estimating a model's out-of-sample predictive accuracy (i.e. a model's ability to generalize to new, unseen individuals). The process begins by temporarily holding out a single subject from the dataset. The hierarchical model is then re-fitted (using EM) using the data from all *other* subjects except the held out subject which generates a set of cross-validated group-level parameters that are not influenced by the held-out subject's data. These cross-validated group parameters are subsequently used as a Bayesian prior to compute the marginal likelihood of the held-out subject's data, a score that quantifies how well the model, trained on the rest of the population, predicts the behaviour of a novel individual. To compute this marginal likelihood, the subject-level parameters are integrated out using the Laplace approximation, a standard technique that approximates the integral by finding the MAP estimate for the held-out subject's parameters under the new prior and using the curvature of the posterior at that point (the Hessian) to estimate the total probability.

This entire process is then repeated iteratively for every subject (each time holding out that subject). This yields a predictive likelihood score for each one. Because each subject's score is computed based on a model that was not trained on their own data, the resulting set of scores across the group can be treated as independent, which makes them suitable for subsequent classical statistical tests at the group level (e.g., ttests) and use in approaches like the Variational Bayesian Approach (VBA).

Compared with simpler information criteria like BIC, this method's penalty for model complexity is implicit and often more reliable (Vehtari et al., 2017). Rather than using an explicit penalty term based on the number of parameters, it penalizes complexity through the process of cross-validation itself. An overly complex model with too many free parameters will tend to overfit the training subjects by capturing their specific behavioural noise. When the group-level parameters from this overfitted model are used as a prior to predict the held-out subject’s data, the predictions will be poor because the noise it has learned is not present in the new subject. This failure to generalize results in a lower predictive likelihood score, which is the mechanism of the penalty. Conversely, an overly simple model that underfits by failing to capture key patterns in the training data will also generate an inaccurate prior and predict the held-out subject's behaviour poorly. The procedure thus favours models that are just complex enough to capture the true, generalizable patterns in the data (Browne, 2000).

Subject-level LOOcv scores were then submitted to the mbb-vb-toolbox in MATLAB for group-level Bayesian model selection (BMS) (Daunizeau et al., 2014). The toolbox implements a random-effects VBA. A random-effects analysis is conceptually superior to a fixed-effects analysis as a fixed-effects approach assumes the same model is best for all subjects, whereas a random-effects approach has a more plausible assumption: different models may best describe different subjects (Stephan, et al., 2009). The core challenge this approach addresses is that for many nonlinear models, the integrals required to compute the exact model evidence or posterior densities are analytically intractable. VBA provides a solution by using an iterative scheme to optimize an approximation to both the model evidence and the posterior density. It does this by maximizing a tractable lower bound on the log model evidence, known as the free energy, This process simultaneously minimizes the Kullback-Leibler (KL) divergence between an approximate posterior density, q(θ), and the true posterior, p(θ|y,m). The specific implementation in the toolbox, known as a variational-Laplace scheme (K. Friston et al., 2007), uses a mean-field assumption to partition the parameters and a Laplace (Gaussian) approximation for the resulting marginal distributions to make the optimization computationally efficient. For random-effects BMS, this approach treats the model frequencies in the population as the unknown parameters to be estimated. It assumes a Dirichlet distribution prior over the vector of model frequencies, r, which is updated based on the log-evidence provided by the model scores – in my case, this is LOOcv - for each model and subject. The output of this inversion process is a posterior Dirichlet distribution over r, from which the key metrics are derived (Stephan et al., 2009).

The VBA provides two key metrics for inference:

1. **Model Frequency:** This is the estimated posterior probability that a given model generated the data for a randomly chosen subject from the population. The frequencies for all models under consideration sum to 1, which should be compared to the chance level (1 divided by the total number of models). This metric is particularly useful for understanding population heterogeneity, as it may reveal that multiple competing models are prevalent, rather than one single winner. For a given model k, its expected frequency is computed from the parameters (α) of the posterior Dirichlet distribution as:

$$E[r_k] = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$$

2. **Exceedance Probability (XP):** This represents the posterior probability that a specific model is more frequent than all other competing models combined. Whereas model frequency gives the expected prevalence, XP quantifies the belief that a given model is the *most* prevalent. For instance, a model could have the highest model frequency (e.g., 0.4) but still have a low XP if other models have similar frequencies (e.g., 0.35 and 0.25), reflecting uncertainty about which is truly the most common. An XP near 1, however, provides strong evidence for a single "winning" model, indicating a high degree of confidence that it is the most common data-generating process in the population (Rigoux et al., 2014). It is calculated by integrating the posterior distribution over the simplex region where its frequency is the largest:

$$XP_k = P!\left(r_k > r_j \ \ \forall j \neq k \;\middle|\; \text{data}\right)$$

As an additional check, I also compared the LOOcv scores of the winning model to the other models using paired sample t-tests, correcting for multiple comparisons by adjusting the p value according to the number of models being compared.

2.3 Model Validation using Simulations

Relative model comparison criteria (like LOOcv) focus on evidence for the best model given the set of models being compared. But they say little about how well a model fits the data in an absolute sense. A winning model is just comparatively better, even if all models poorly describe important data features (Palminteri, Wyart, et al., 2017). Consequently, a model deemed "winning" based on fitting alone may still fail at reproducing key behavioural signatures of the data. Further, the likelihood maximization procedure used in model fitting can inadvertently inflate a model's performance. It may favour parameters that, by chance, maximize the probability of observing an effect already present in the data, even if the model's intrinsic computational process cannot generate that effect. This differs from cross-validation by shifting the focus from descriptive accuracy to generative capability. For example, a simple model might achieve a better descriptive fit (lower LOOcv) than a more complex one, yet be unable to recreate a key behavioural pattern in simulations. A simulation is the process of creating a synthetic dataset by having a model "perform" the experimental task. It is the analysis of model simulations that provides insight into the behavioural bases for accepting or rejecting a model, elucidating *why* a particular model is effective, rather than merely *which* model fits best (Palminteri, Wyart, et al., 2017).

I validated winning models by seeing the degree to which it was able to reproduce the pattern of behavioural data through simulations. For each simulated participant, I began by drawing a set of parameter values from uniform distributions that spanned the plausible ranges observed in the parameters estimated from the real data. This synthetic agent, equipped with these parameters, then progressed through the task trial by trial. On each trial, the model's equations were used to update its internal states (e.g., Q-values) based on the outcomes and to generate a probabilistic choice based on those states, mirroring the decision-making process hypothesized for human participants. I repeated this for the same number of trials and conditions as in the actual experiment, resulting in a synthetic behavioural dataset for each model. To evaluate the simulations, I treated the generated data as if it were from real participants. I applied the exact same model-free statistical analyses (e.g., t-tests or ANOVAs or regressions) to the synthetic data as I did to the human data. I was interested to see whether the key experimental effects observed in the human participants were also present in the simulated data. The key experimental effects were defined as the statistically significant findings from the model-free analyses, which tested the study's main hypotheses. A successful validation required that the winning model not only fit the data well quantitatively but could also qualitatively reproduce these effects (e.g., show a statistically significant difference in performance between Condition A and B that was comparable to the human effect). This confirms that the model's internal mechanisms provide a sufficient explanation for the observed behaviour.

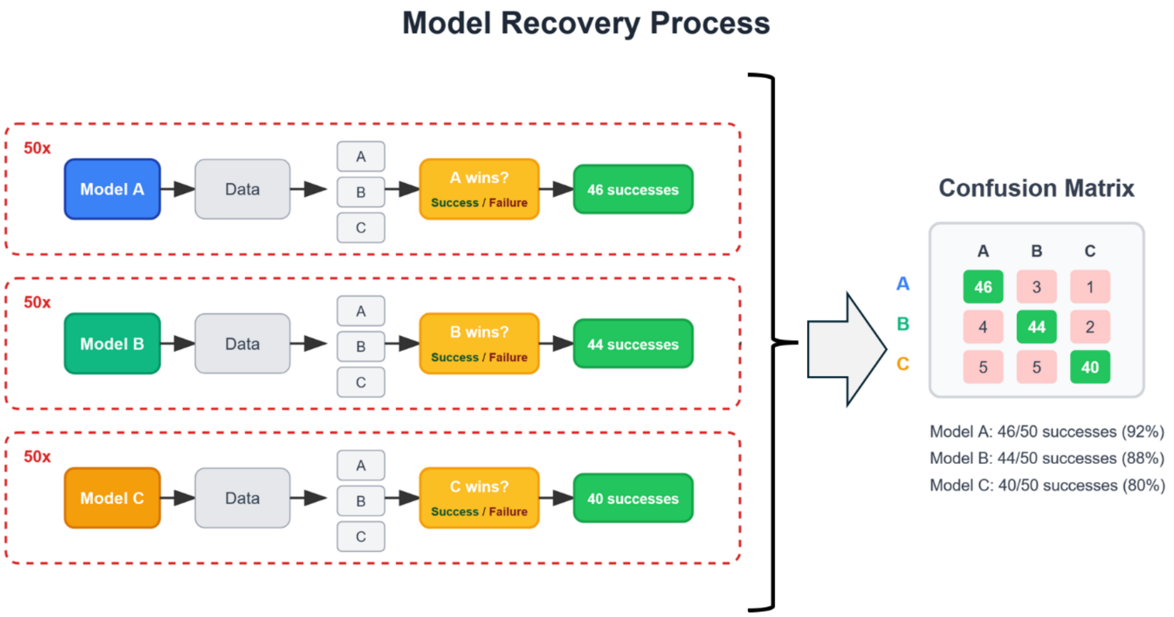
2.4 Identifiability Analyses

The validation confirms that the winning model can reproduce the observed behaviour. However, for this conclusion to be robust, two further methodological conditions must be met. First, the model comparison procedure itself must be sensitive enough to distinguish between the candidate models – model recovery. Second, the parameters of the winning model must be uniquely identifiable and meaningful – parameter recovery (Wilson & Collins, 2019).

2.4.1 Model Recovery

Model recovery asks "if one of our models were the true data-generating process, could our model comparison procedure reliably detect it? "This is essential for ruling out the possibility that a "winning" model is simply more flexible and can mimic data from other models - a problem known as model confusion. A failure of model recovery would imply that, within the context of the current experimental design, two or more models are "conceptually unidentifiable" as they produce nearly indistinguishable patterns of behaviour (Heathcote et al., 2015).

To conduct model recovery (Figure 2.2), I began with a simulation step. For each candidate model, I generated a synthetic dataset for a simulated group of subjects, with the number of subjects and trials mirroring the real experiment. The parameter values I used for this simulation were the ranges observed in the actual data. This process resulted in a collection of synthetic datasets, one for each model under consideration. Next, I subjected each of these synthetic datasets to the full model comparison pipeline. Specifically, I fitted every candidate model to each synthetic dataset and fed the resulting LOOcv scores into the VBA toolbox to determine a winning model based on the exceedance probability. I repeated this entire simulation-and-comparison process for 50 iterations and recorded how often the model used to simulate the data was correctly identified as the best-fitting model (e.g., if data were simulated using model M1, M1 should have the highest XP value). Next, I aggregated the outcomes into a confusion matrix, which visualizes the proportion of iterations in which data generated from a specific model (rows) was correctly identified as best fit by that same model (columns). An ideal confusion matrix has high values on the diagonal (indicating successful recovery) and low values on the off-diagonals (indicating low confusion between models). This result validates that the models make sufficiently different predictions and that the comparison method is sensitive enough to identify the true underlying model. Additionally, I calculated the mean LOOcv score for each model during each iteration as another model identifiability metric.



	A	B	C
A	46	3	1
B	4	44	2
C	5	5	40

Model A: 46/50 successes (92%)
Model B: 44/50 successes (88%)
Model C: 40/50 successes (80%)

Figure 2.2: The model recovery process for a hypothetical model space of three models (A, B, and C). The workflow is iterated for all candidate models. In each iteration, one of the models is designated as the 'ground truth', meaning it is used to simulate the data. For example, the top row shows the process where Model A is the ground truth. A synthetic dataset is generated using Model A and then all models are fitted to this dataset and compared. If Model A comes out as the

winner, then this iteration is a success. This entire simulation-and-comparison loop is repeated for 50 iterations, and the number of successes and failures are counted. In the example shown, out of 50 iterations 46 were a success and 4 a failure, corresponding to a 92% recovery rate. The results from all iterations are aggregated into a Confusion Matrix (right panel). In this matrix, the rows represent the true, data-generating model, and the columns represent the model that was selected as the winner by model comparison. The diagonal elements show the proportion of successful recoveries (e.g., the cell A,A shows that Model A was correctly recovered 46 times). The off-diagonal elements show instances of "model confusion," where the model that generated the data did not win (e.g., the cell A,B shows that Model B emerged as the winner 3 times when Model A was the true model). An ideal confusion matrix has high values on the diagonal and low values on the off-diagonals, providing confidence that the models are distinguishable.

2.4.2 Parameter Recovery

Parameter recovery assesses whether the individual parameters of a given model are identifiable (Wilson & Collins, 2019). A lack of identifiability can arise from model misspecification or poor experimental design, leading to parameter trade-offs where similar behaviour can be produced by opposing changes in two parameters (e.g., a low learning rate might be compensated by higher decision noise). This would render the interpretation of fitted parameter values at best difficult, at worst meaningless (Kruschke, 2015).

To check for this, I simulated data from the winning model using "true" parameter values drawn randomly from gaussian distributions – to focus on the most common values - spanning the empirically observed range. The standard hierarchical fitting procedure was then used to "recover" the parameters from this simulated data. Recovery success was assessed in two ways. First, by examining the relationship between the true and recovered parameter values. This is typically visualized using scatter plots, where strong recovery is indicated by the points clustering tightly around the identity line (y=x), and a correlation matrix where the values are given by the Pearson correlation between the true and recovered values. I used a correlation value of 0.80 and higher as the evidence that a parameter is well-constrained by the data and can be estimated reliably (Figure 2.3 (A)) while values lower than that would raise concerns (Figure 2.3 (B)). Second, to diagnose potential trade-offs, the correlation matrix of the *recovered* parameters was examined. Strong correlations between different recovered parameters which were generated independently, would indicate an identifiability issue, suggesting that the model is unable to disentangle the unique contribution of each parameter to the behaviour (Wilson & Collins, 2019). Successful recovery, marked by high true-to-recovered correlations and low correlations between different recovered parameters, provides confidence that the model's parameters can be validly interpreted.

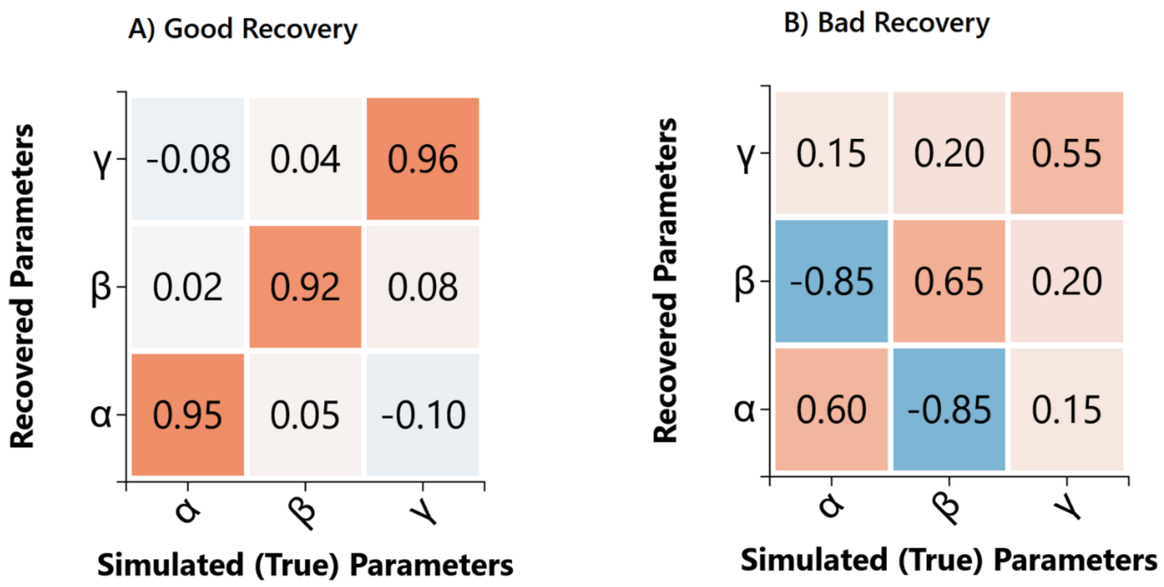


Figure 2.3: Examples of good and bad parameter recovery. A) High correlations on the diagonal indicate that true parameters were successfully recovered. B) Low correlations on the diagonal and strong off-diagonal correlation (here, between α and β) indicate parameter recovery failure. The numbers are fictitious and just for illustrative purposes.

2.5 Testing for differences between parameters

A key question I address using computational models in this thesis is whether learning (i.e. the degree to which beliefs are updated) varies between different conditions. This rate of learning is captured in the models by learning rates, with different learning rates used to characterise learning in different conditions (e.g., one learning rate for confirmatory feedback and one for disconfirmatory). To assess whether learning is indeed statistically different between conditions, I tested for differences using a hierarchical ttest. This test was used as a result of the EM model fitting process used; it was important to use a method that accounts for the statistical properties of the hierarchically estimated parameters. Applying standard frequentist tests (e.g., t-tests) to individual parameter point-estimates from a hierarchical model is statistically invalid. The regularization or "shrinkage" inherent in the fitting process violates the assumption of independence required by such tests, as each individual's estimate is influenced by the group distribution. This shrinkage artificially reduces inter-subject variance, leading to a substantial inflation of the Type I error rate (i.e., false positives) (Piray et al., 2019). To overcome this, following Piray and colleagues' (2019) approach, I used hierarchical t-tests. This method operates on the posterior distribution of the *group-level* parameters (e.g., the group mean), which correctly reflects the uncertainty of the estimate at the population level. The test evaluates whether a credible interval for the group mean effect includes zero, based on the estimated mean and its hierarchical standard error.

Chapter 3: Confirmation Bias Exists in the Face of False Information

3.1 Introduction

When a financial trader pursues the markets and ignores an increase in the price of shares they recently sold whilst boosting their ego from an increase in the price of shares they recently bought, they are applying a well-known learning bias prevalent in decision making. Across a wide range of domains, information consistent with past choices and judgments is integrated into beliefs to a greater extent than information that challenges them. This phenomenon, known as *confirmation bias* ((Bronfman et al., 2015; Klayman & Ha, 1987; Nickerson, 1998; Talluri et al., 2018), impacts a range of domains, ranging from finance (Park et al., 2010) to science (C. X. Cheng, 2018; Darley & Gross, 1983) to politics (McClung Lee, 1949).

The computational principles that enable biased beliefs to persist in the face of new evidence are thought to arise from a key feature of how we learn: the differential use of prediction errors, which quantify the difference between expected and received outcomes (Sutton & Barto, 2018). By enabling prediction errors to selectively have a greater impact when these confirm versus disconfirm our expectations, information that confirms past choices is amplified, whilst information that undermines them is ignored. This mechanism – a form of *asymmetric learning* - goes against classic normative theories from economics (Neumann & Morgenstern, 1944), machine learning (Russell & Norvig, 1995), and psychology (Körding & Wolpert, 2004; Maslow, 1950). However, a raft of neurobiological (Dabney et al., 2020; Garrett et al., 2014; Lefebvre et al., 2017; Sharot et al., 2011) and computational (Garrett & Daw, 2020; Lefebvre et al., 2017; Palminteri, Lefebvre, et al., 2017) evidence converge to suggest that the process of updating beliefs in the face of new evidence involves prediction errors changing beliefs to differing degrees, depending on both their sign (whether the prediction error is positive – greater than expected or negative – less than expected) and whether this sign signals one has made the right (a positive prediction error for a chosen option or a negative prediction error for an unchosen option) or wrong (the converse) decision.

Whilst the computational principles that give rise to confirmation bias have been established, much of the theory and empirical work has been confined to cases in which information is accurate (Chierchia et al., 2023; Lefebvre et al., 2022; Palminteri, 2023; Palminteri, Lefebvre, et al., 2017; Rollwage et al., 2020; Rollwage & Fleming, 2021) (but see recent work from (Vidal-Perez et al., 2025). Yet much of our everyday experience involves gathering and processing information, which often transpires to be either intentionally or unintentionally false. Understanding whether confirmation bias also exists in the face of such cases and, if it does, establishing if it arises from similar computational mechanisms, is an increasingly prescient question in an era where platforms that are regularly used to receive and share information prioritise engagement over accuracy, which can lead to the proliferation of misleading content (Lewandowsky et al., 2017).

Here, I combined behavioural testing with a novel learning paradigm in conjunction with computational modelling in two separate studies. In the task (Figure 3.1(a)), participants (study 1, online: N=47; study 2, in the lab: N=57) made choices between pairs of options (abstract symbols). Following the outcome (gain/loss), cues indicated whether the outcome was genuine or false. The task dissociated outcome from true and false information and used learning to both avoid losses (where getting -1 is the better outcome compared to -10) and accrue gains (where getting +1 is the worse outcome compared to +10) in order to disassociate effects driven solely by outcome valence. I also provided counterfactual (outcome shown for the unchosen option) as well as factual (outcome shown for the chosen option) outcomes to be able to disassociate confirmation bias from positivity bias.

3.2 Methods

3.2.1 Participants

A total of 70 participants were recruited online via Prolific for the first study, 23 of whom were excluded; therefore, the final sample was 47 participants (mean [standard deviation] age: 30.45 [7.2]; 28 female). A total of 91 participants were recruited from the university pool for the second study, 34 of whom were excluded, leaving the final sample of 57 participants (mean [standard deviation] age: 20.45[3.8]. While these exclusion rates (roughly 30–37%) might seem high, they actually align with current best practices that emphasize data quality over sheer numbers, where exclusion rates like my experiments are often necessary to weed out careless responding by inattentive participants (Nadler et al., 2021; Peer et al., 2022). Further, Zorowitz et al., (2023) found that keeping such participants in the dataset can create spurious correlations between behavioural tasks and self-reported measures.

Three exclusion criteria were applied to ensure data quality. First, participants who incorrectly answered more than one of the ten catch trials were excluded (n=1 in Study 2). Second, trials with reaction times below 100ms or above 4 seconds were removed from analysis. Third, participants showing subpar learning performance were excluded, defined as choosing the better option less than 55% of the time in solvable conditions (n=23 in Study 1, n=33 in Study 2). Study 1 participants received £3 plus a performance-based bonus ranging from £3-£6. Study 2 participants received course credits plus a performance bonus up to £3. All participants provided informed consent prior to participation. The research protocol received approval from the University of East Anglia's ethics committee and complied with all relevant ethical guidelines.

3.2.2 Behavioural task

Participants completed an instrumental learning task where, on each trial, they chose between two options, received an outcome, and were told whether the outcome was true (a tick symbol) or false (a cross symbol) – Information Cues (Figure 3.1(a)). Participants were told that when they received a cue saying the preceding outcome was false this represented a "glitch" showing an outcome from an unrelated game, and they should ignore this information when learning which options were better.

The experiment consisted of 160 trials organised into gain and loss contexts. In gain trials, participants could receive +1 or +10 points, while loss trials involved -1 or -10 points. Information cues were equally split between true and false trials (50% each), with assignments made randomly. Four consistent option pairs were used throughout the experiment, where a given option was always paired with the same counterpart. These option pairs were presented in random order with randomised left/right positioning.

Two experimental conditions defined the underlying reward structure (Figure 1(b), and (c)). In the solvable condition, when the outcome was to be subsequently designated true, one option had an 80% chance of giving the superior outcome of the two outcomes available and the other 20%. When the outcome was to be subsequently designated false it was 50% likely to be favourable for both options. In the unsolvable condition, when the outcome was to be subsequently designated true, it was 50% likely to be favourable for both options (hence unsolvable). When the outcome was to be subsequently designated false, one option had an 80% chance of giving the superior outcome of the two outcomes available and the other 20%.

Participants' goal was to identify which option more frequently provided rewards (in gain pairs) or less frequently provided punishments (in loss pairs). They learned these preferences through trial and error, as they were not privy to probability contingencies, but had to pay attention to true and false information. Although participants were instructed to ignore false information, these trials still contributed to their final bonus payment.

As an attention check, ten separate catch trials were randomly shown to participants, and they had to answer whether the previous trial they just saw was true or false. The task was programmed in JavaScript using the toolbox JSPSYCH version 6.3 (Leeuw et al., 2023).

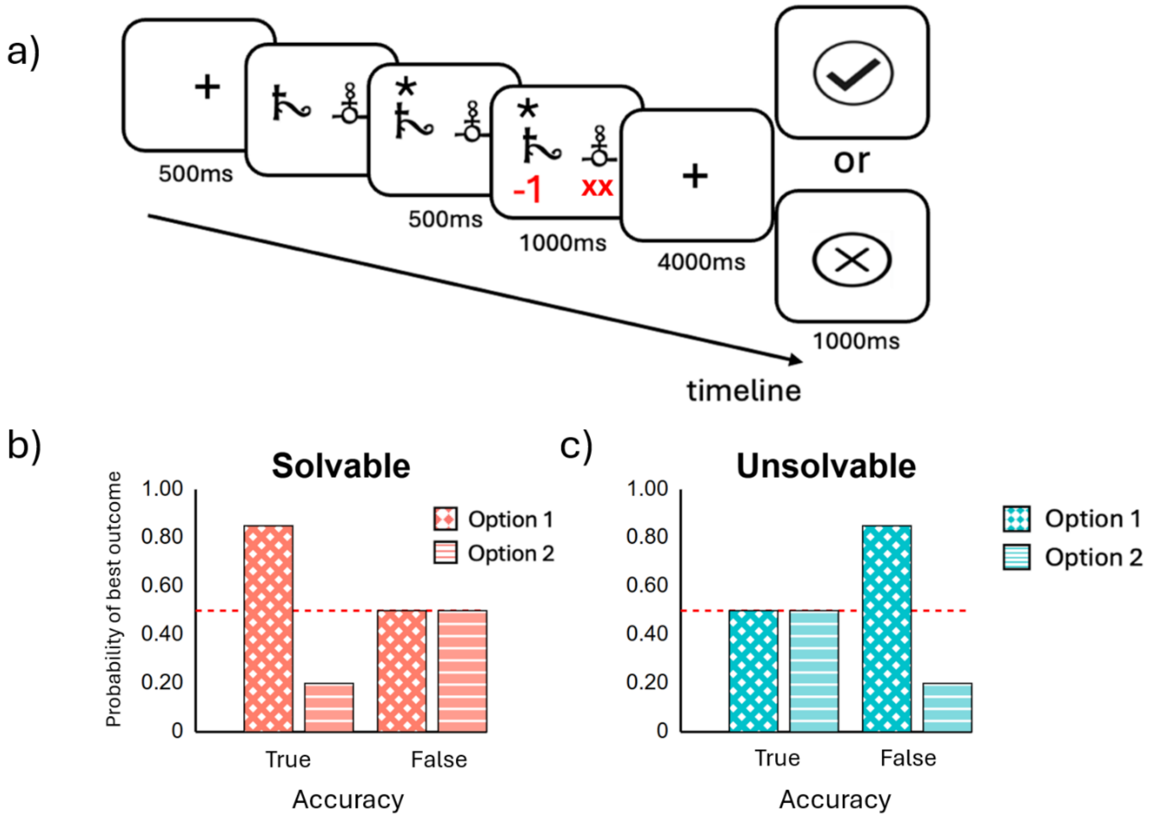


Figure 3.1: The Task. (a) Participants made choices between four pairs of options (abstract symbols) and received an outcome (win/lose money) about one of the two options (the option chosen or the unchosen option). A cue (tick or cross) then indicated whether the outcome was genuine (tick) or false (cross). In 6.25% of trials, we then asked participants to report what the accuracy cue had been as an attention check. **(b)** For two of the pairs, in true trials, in most trials one of the two options provided the best outcome (either winning money or avoiding losing money), but outcomes were random for each option in false trials. **(c)** For the other two pairs, this pattern was reversed, such that in true trials, outcomes were random for each option; hence, there was no better option. In false trials, however, one option gave favourable outcomes 80% of the time, potentially misleading the participants into thinking it was the better option.

3.2.3 Behavioural Analysis

To assess whether participants had used true or false information, I averaged the number of times each participant had selected the correct option in the solvable condition and the misleading option in the unsolvable condition (i.e. the option that gives them a false favourable outcome 80% of the time) and then averaged this average to obtain the mean choice rate for them. Next, I ran a one-sample t-test on this mean against 0.50. A significant result would indicate that they have learned about the correct and misleading options.

3.2.4 Computational Models

I fitted several models that were variations of the standard Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972). In each model, in each trial (t), either the chosen or unchosen option is updated depending on whether the outcome is shown for one or the other. The formula is the same. For instance, for the chosen option, we have:

$$Q_c(t+1) = Q_c(t) + \alpha * \delta_c(t)$$

In which δ_c is the prediction error (δ) for the chosen option, defined as the difference between the expected outcome and the observed outcome:

$$\delta_c(t) = R_c(t) – Q_c(t)$$

And α is the learning rate parameter determining the extent to which PEs are used as a learning signal.

Then, the learned values of the two A and B options are converted to choice probabilities using the SoftMax function. For instance, the probability of choosing option A is given as:

$$\frac{1}{1 + \exp(\beta(Q_B - Q_A)}$$

Where β is the inverse temperature parameter that controls the degree of stochasticity in choice behaviour. Larger values of β yield more deterministic choices, while smaller values reflect more exploratory behaviour.

The contribution of each trial to the likelihood was given by the log probability of the observed choice. For a choice between the chosen (i_c) and unchosen (i_u) options:

$$l_t = \log\left(\frac{1}{1 + \exp\left(-\beta\left[Q_{i_c}(t) - Q_{i_u}(t)\right]\right)}\right)$$

The log-likelihood for a subject was then:

$$\mathcal{L} = \sum_t l_t$$

and the model minimised the negative log-likelihood, –L, during estimation.

To test the effect of Feedback (Confirmatory vs Disconfirmatory) and Accuracy (True vs False) on decision-making, I tested four different models with different numbers of α (α = 2, 3 or 4). The number of learning rates was varied according to Feedback and Accuracy dimensions. The confirmatory feedback is defined as a positive PE (+10 in the gain context and -1 in the loss) for the chosen option or a negative PE (+1 in the gain context and -10 in the loss) for the unchosen option. Conversely, the disconfirmatory feedback is when a negative PE outcome occurs for the chosen option or a positive PE for the unchosen option.

The four models were formulated as follows:

Model 1 (M1)

$Q(t+1) = Q(t) + \alpha_{true} * \delta(t)$ if information accuracy cue = True

$Q(t+1) = Q(t) + \alpha_{false} * \delta(t)$ if information accuracy cue = False

Free parameters (n=3): $\alpha_{true,}$ $\alpha_{false,}$ β

Model 2 (M2)

$Q(t+1) = Q(t) + \alpha_{true} * \delta(t)$ if accuracy = True

$Q(t+1) = Q(t) + \alpha_{Conf, false} * \delta(t)$ if accuracy = False and feedback = Confirmatory

$Q(t+1) = Q(t) + \alpha_{Disconf, false} * \delta(t)$ if accuracy = False and feedback = Disconfirmatory

Free parameters (n=4): α_{true}, $\alpha_{Conf, false}$, $\alpha_{Disconf, false}$, β

Model 3 (M3)

$Q(t+1) = Q(t) + \alpha_{false} * \delta(t)$ if accuracy = False

$Q(t+1) = Q(t) + \alpha_{Conf, true} * \delta(t)$ if accuracy = True and feedback = Confirmatory

$Q(t+1) = Q(t) + \alpha_{Disconf, true} * \delta(t)$ if accuracy = True and feedback = Disconfirmatory

Free parameters (n=4): α_{false}, $\alpha_{Conf, true}$, $\alpha_{Disconf, true}$, β

Model 4 (M4)

$Q(t+1) = Q(t) + \alpha_{Conf, false} * \delta(t)$ if accuracy = False and feedback = Confirmatory

$Q(t+1) = Q(t) + \alpha_{Disconf, false} * \delta(t)$ if accuracy = False and feedback = Disconfirmatory

$Q(t+1) = Q(t) + \alpha_{Conf, true} * \delta(t)$ if accuracy = True and feedback = Confirmatory

$Q(t+1) = Q(t) + \alpha_{Disconf, true} * \delta(t)$ if accuracy = True and feedback = Disconfirmatory

Free parameters (n=5): $\alpha_{Conf, true}$, $\alpha_{Disconf, true,}$ $\alpha_{Conf, false}$, $\alpha_{Disconf, false}$, β

Next, I created two 8-learning-rate models. In the first, to see if the observed pattern of learning rates was indeed confirmation bias and not positivity bias, Model 4 was expanded to include separate learning rates for factual and counterfactual outcomes (see Appendix 3.1 for full details). Second, to see if the confirmation bias for true and false information exists for *both* Gain and Loss contexts, Model 4 was expanded to include separate learning rates for Gain and Loss contexts (see Appendix 3.2). An additional supplementary model introduced gradual perseveration parameters to Model 4. The goal here was to see if the confirmation bias is robust once the gradual perseveration is considered, as some argue the confirmation bias is a pseudo-bias bias emerging from perseveration (Katahira, 2018; Sugawara & Katahira, 2021), while others defend the validity of the bias (Palminteri, 2023) .

The core idea behind the gradual perseveration model is to maintain a "choice trace" - like a memory of how often an option has been selected - for both chosen and unchosen options:

$$C_c(t+1) = C_c(t) + \tau(CPE(c))$$

$$C_u(t+1) = C_u(t) + \tau(CPE(u))$$

$$CPE(c) = 1 – C_c(t)$$

$$CPE(u) = 0 – C_u(t)$$

Where C_c and C_u are the choice traces for the chosen and unchosen options, respectively. When an option is chosen, its trace is increased towards 1; when an option is not chosen, its trace is decreased towards 0. This update process is driven by a “choice prediction error" (CPE) and choice trace accumulation rate (τ) akin to a learning rate, which controls how quickly the trace adapts. For instance, if it is set to 1, then only the previous choice affects the current choice, while lower values mean more of the past choices are influential.

This choice trace then biases future decisions. The probability of picking one option (A) over another (B) is determined not just by its expected value (Q) but also by its choice trace (C) as given by the following:

$$\frac{1}{1 + \exp(\beta(Q_B - Q_A) + \varphi(C_B - C_A))}$$

In which, β determines how much the learned value of an option influences the choice, while φ (phi) determines how much the history of past choices sways the decision. If φ is positive, it encourages repeating past choices (perseveration), while if negative, it encourages switching to different options (alternation).

Therefore, the gradual perseveration model has seven free parameters: $\alpha_{\text{Conf, true}}$, $\alpha_{\text{Disconf, true}}$, $\alpha_{\text{Conf, false}}$, $\alpha_{\text{Disconf, false}}$, β, φ, τ.

It should be noted that I focus on the four main models in model comparison and model recovery for two reasons. First, these models test my hypotheses of interest, while the supplementary models serve as robustness checks. Second, I employ the same modelling approach in subsequent chapters, ensuring a consistent and generalisable framework throughout the thesis.

3.2.5 Model Fitting Procedure

Models were fitted hierarchically by maximising the likelihood of observed choices using an Expected Maximisation (EM) algorithm (Huys et al., 2011) in Julia (v1.9.4) (Bezanson et al., 2012). This hierarchical approach was chosen for its superior performance in predicting unobserved data (Scheibehenne & Pachur, 2015). A full description of the fitting procedure is available in **Chapter 2.**

3.2.6 Model Comparison

To compare model performance, I calculated subject-level leave-one-out cross-validation (LOOcv) scores. I analysed these scores using a Variational Bayesian Approach (VBA; Daunizeau et al., 2014) to determine model frequencies and the exceedance probability for each model. The exceedance probability indicates how likely it is that one model is a better fit than all others in the set (Daunizeau et al., 2009). The full model comparison strategy is detailed in **Chapter 2.**

3.2.7 Statistical Tests on the Learning Rates

After identifying the winning model, I reparametrized it to create parameters that directly quantified the magnitude of confirmation bias for true and false information, respectively (see Appendix S for equations). Because these parameters were estimated hierarchically, standard t-tests could yield biased results (Piray et al., 2019). Therefore, I used hierarchical t-tests, which are designed for this data structure, to assess the effects of interest. See **Chapter 2** for the full details.

3.2.8 Model Recovery

To ensure my models were distinguishable, I conducted a model recovery analysis. For each of the four models, I generated 50 synthetic datasets using parameters from the experimental data. I then fitted all four models to each synthetic dataset to verify that the data-generating model could be correctly identified via the VBA procedure. The full model recovery strategy is detailed in **Chapter 2.**

3.2.9 Parameter Recovery

I also conducted a parameter recovery analysis for the winning model to ensure its parameters could be reliably estimated. I generated data for 5000 synthetic participants using a range of known parameter values. After fitting the model to this data, I compared the original and recovered parameters using Pearson correlations to confirm a high degree of correlation. The full parameter recovery strategy is detailed in **Chapter 2.**

3.3 Results

Participants learn from true and false information. Analysing choice rates revealed that on average participants selected Option 1 (O1) over Option 2 (O2) in both the solvable (Study 1: $t(46) = 14.66$, $p < 0.01$; Study 2: $t(56) =12.50$, $p < 0.01$) and unsolvable conditions (Study 1: $t(46) = 2.64$, $p < 0.05$; Study 2: $t(56) = 3.47$, $p < 0.01$), Figure 3.2. This suggests that participants integrated feedback from both true and false information. However, this was not the result of ignoring accuracy cues since there was clear evidence that information integration was modulated by these cues; the propensity to choose O1 over O2 was greater in the solvable condition compared to the unsolvable one (Study 1: $t(46) = 5.42$, $p < 0.01$; Study 2: $t(56) = 5.48$, $p < 0.01$). As an additional check to verify that participants paid attention to this cue, on 6.25% of trials, participants were asked to report what the accuracy cue had been on the previous trial – participants correctly reported this 90% of the time.

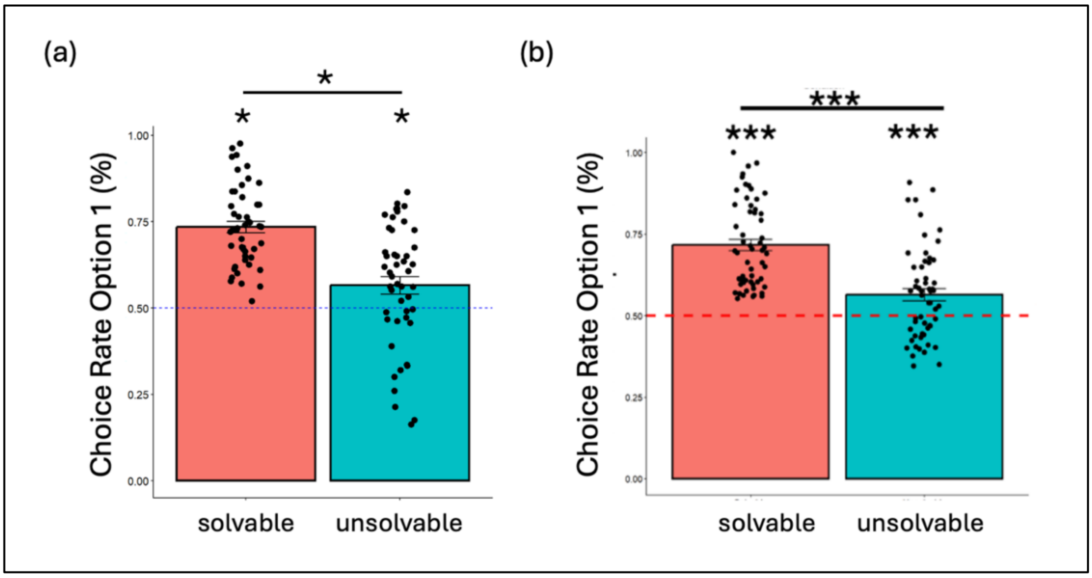


Figure 3.2: Choice Rates for (a) Study 1 and (b) Study 2. Participants opt to select option 1 - the option that provides the best outcome (+10 in the gain context or -1 in the loss context) more often relative to option 2 - to a greater degree in solvable compared to the unsolvable conditions(Experiment 1: t(46) = 5.42, p < 0.01; Experiment 2: t(56) = 5.48, p < 0.01). But, in both solvable (Experiment 1: t(46) = 14.66, p < 0.01; Experiment 2: t(56) =12.50, p < 0.01) and unsolvable (Experiment 1: t(46) = 2.64, p < 0.05; Experiment 2: t(56) = 3.47, p < 0.01) conditions, participants chose O1 to a greater degree than chance. Choice rates are averaged over gain and loss contexts. *p < 0.05, ***p < 0.001 (one-tailed test vs 0.5 or paired sample t-test as appropriate)

Next, I sought to assess if participants learned from false information in a biased manner. I tested four computational models that differed in their number of learning rates (see **Methods**). All models shared the same basic structure but varied in how they parsed confirmatory versus disconfirmatory feedback and true versus false information.

Model 1 used two learning rates: one for true and another for false information. Model 2 had three learning rates, maintaining a single rate for true information but splitting false information into separate rates for confirmatory and disconfirmatory false information. Model 3 also used three learning rates but took the opposite approach, using one rate for false information while distinguishing between confirmatory true and disconfirmatory true information. Finally, Model 4 incorporated four learning rates, providing separate rates for each combination: confirmatory true, disconfirmatory true, confirmatory false, and disconfirmatory false information. This modelling approach allowed me to examine whether participants processed information differently based on feedback (confirmatory vs disconfirmatory) and veracity (true vs. false). Full model specifications are provided in the **Methods** section.

Four learning rate model the best fit to the data. Across both studies, Model 4 consistently provided the best fit to the data. I compared the four models using leave-one-out cross-validation (LOOcv) scores and a Variational Bayesian Approach, and in both studies, Model 4 achieved the highest model frequency (~83% in Study 1 - Figure 3.5(a) - and ~93% in Study 2 – Figure 3.5(c)) and an exceedance probability of 1.0 (Figure 3.3(b)). These frequencies were well above the 25% chance level, indicating strong evidence that the four-parameter learning structure of Model 4 was the best at capturing participants' behaviour. Model recovery (Figure 3.3 (b) and (c)) indicated all models are identifiable and parameter recovery showed all parameters of the winning model had a high recovery rate (Figure 3.4 (a)) and the correlation between the parameters in the real data was low for study 1 (Figure 3.4 (b)) and study 2 (Figure 3.4 (c)).

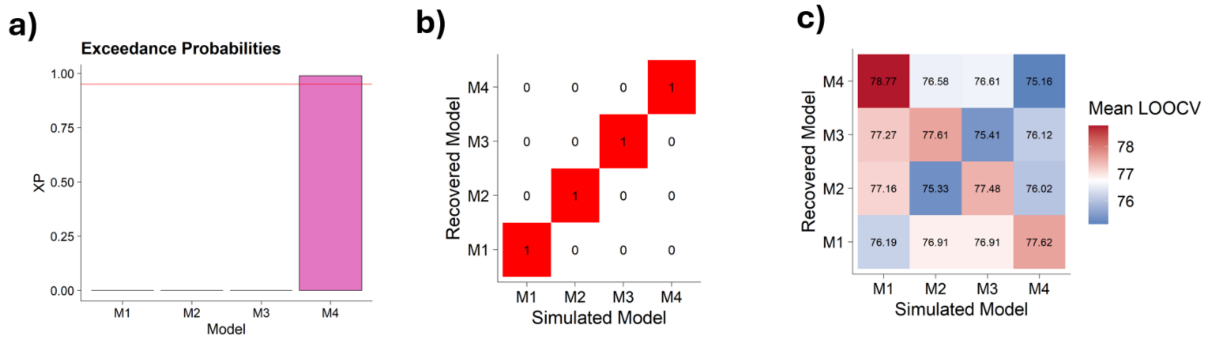


Figure 3.3: Model Fit and Recovery. (a) Shows the exceedance probabilities (XP), which quantify the confidence that each model is more likely than all other models in the set. M4 achieved an exceedance probability of nearly 1.0, indicating extremely high confidence that it outperforms the competing models. **(b)** Shows the confusion matrix representing model recovery accuracy. Each simulated model (x-axis) is correctly identified by the model comparison procedure as the best-fitting model (y-axis), with all values on the diagonal equal to 1 and off-diagonal values equal to 0. This indicates perfect recoverability and discriminability between the models, confirming that the model-fitting approach can reliably distinguish among the candidate models. **(c)** Displays the mean Leave-One-Out Cross-Validation (LOOcv) scores for each model averaged over 50 iterations, where lower values indicate better predictive performance. The simulation results demonstrate that when data were generated from a specific model (columns), the corresponding model generally achieved the lowest LOOcv score when fitted to that data, validating my model recovery procedure. Notably, M4 showed strong recovery performance.

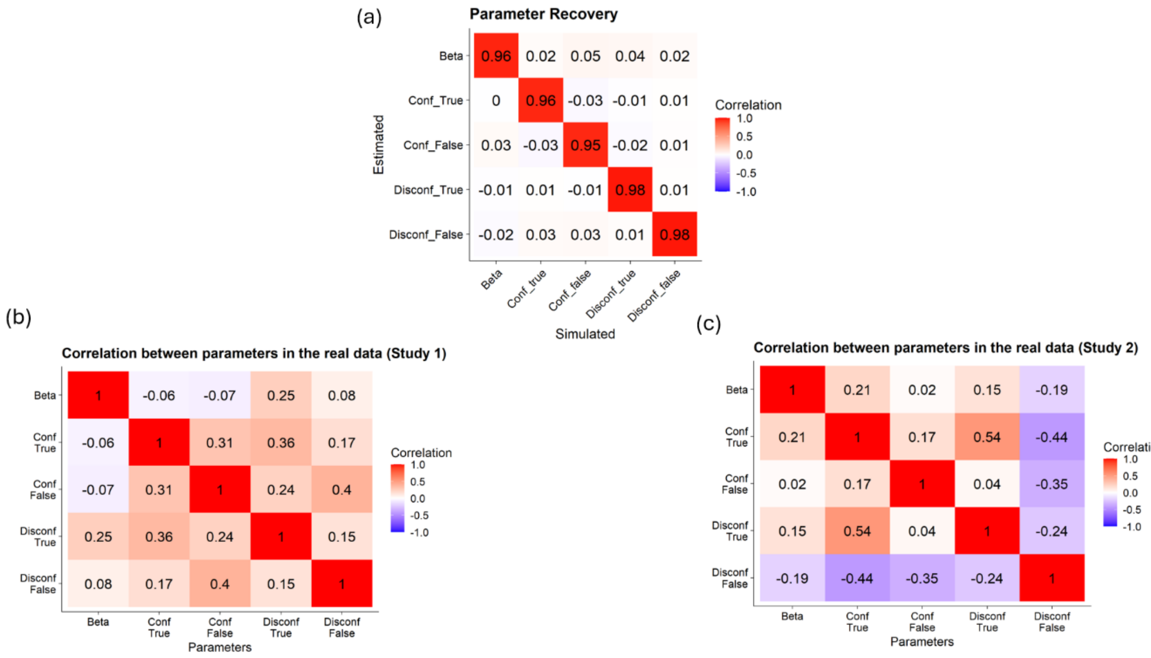


Figure 3.4: Parameter Recovery. (a) Successful parameter recovery of the winning model with high correlations between the simulated and estimated parameters. **(b)** The correlation between the parameters of the winning model in Study 1 and **(c)** in Study 2. The weak correlations demonstrate that parameters do not systematically trade off against each other during estimation, supporting the model's identifiability.

As an additional check, paired sample t-tests with FDR correction confirmed Model 4's superiority over all other models. In Study 1, Model 4 was significantly better than Model 1 (t(46) = −4.75, p_adj < 0.001), Model 2 (t(46) = −3.69, p_adj<0.001), and Model 3 (t(46) = −3.07, p_adj < 0.01). This was also true in Study 2, where Model 4 again outperformed Model 1 (t(56) = −4.94, p_adj < 0.001), Model 2 (t(56) = −3.82, p_adj < 0.001), and Model 3 (t(56) = −3.34, p_adj < 0.01).

This advantage was also apparent at the individual level. In Study 1, Model 4 provided the best fit for the largest portion of participants (44.7%), followed by Model 3 (25.5%), and then Models 2 and 1 (both 14.9%). Similarly, in Study 2, Model 4 accounted for the largest group of individuals (42.1%), compared to Model 3 (22.8%), Model 2 (19.3%), and Model 1 (15.8%). These results show differential treatment of PEs across accuracy (true vs false) and feedback (confirmatory vs disconfirmatory), with four learning rates governing the decision-making.

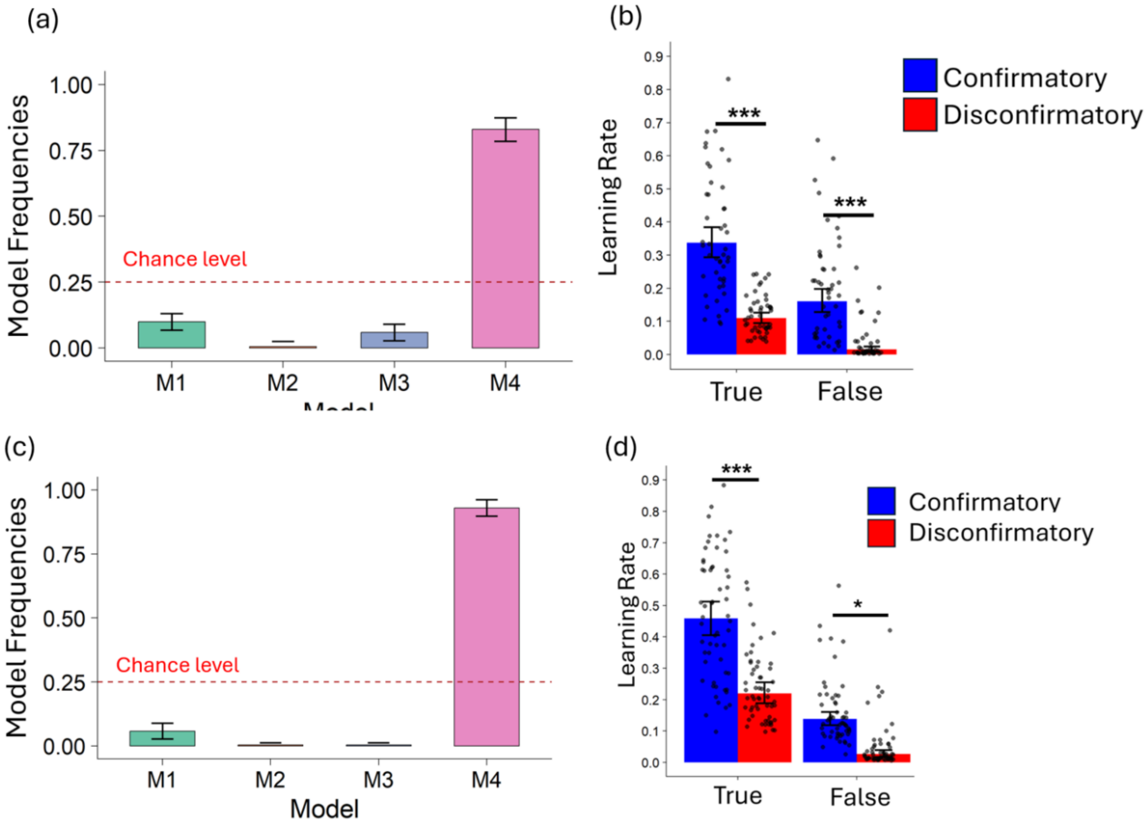


Figure 3.5: Modelling Results and Estimates. (a) Estimated model frequencies from the VBA model comparison in study 1. Model 4 (M4) had the highest frequency, selected for approximately 83% of participants, with an exceedance probability (XP) of 1. **(b)** The estimates from M4 showed a higher learning rate for confirmatory versus disconfirmatory feedback for true (t(46) = 6.50, p < 0.001, hierarchical t-test comparing α_{conf_true} with $\alpha_{disconf\_true}$) and false (t(46) = 5.19, p < 0.001, hierarchical t-test comparing α_{conf_false} with $\alpha_{disconf\_false}$) information, indicating confirmation bias. (c) Estimated model frequencies from the VBA model comparison in study 2. Similar to study1, Model 4 (M4) had the highest frequency, selected for approximately 93% of participants, with an exceedance probability (XP) of 1. **(d)** The estimates from M4 show the existence of confirmation bias for true (t(56) = 5.24, p < 0.001, hierarchical t-test comparing α_{conf_true} with $\alpha_{disconf\_true}$) and false (t(56) = 2.43, p = 0.01, hierarchical t-test comparing α_{conf_false} with $\alpha_{disconf\_false}$) in this study as well. The model frequency reflects the proportion of the population best accounted for by each model. ***p < 0.001, *p < 0.05, hierarchical t-test.

Confirmation bias exists for false information. I then probed the pattern of learning rates from the winning model (M4) (Figures 3.5 (b) and (d)). In both studies, participants exhibited a strong confirmation bias for false information (t(46) = 5.19, p < 0.001; t(56) = 2.43, p = 0.01) and true (t(46) = 6.50, p < 0.001; t(56) = 6.38, p < 0.001), learning more from confirmatory vs disconfirmatory feedback. There was no significant difference in the size of this bias between true and false information (t(46) = 1.81, p = 0.07; t(56) = 1.11, p = 0.26). I confirmed that this was indeed confirmation bias and not positivity bias (see Appendix 3.3) and that it is robust across Gain and Loss contexts (see Appendix 3.4). Then I controlled for perseveration by estimating the learning rates of the gradual perseveration model to see if the confirmation bias holds (Figure 3.6). The reason for this analysis was that within the context of such RL experiments, a group of researchers argues that the confirmation bias is a "pseudo-bias" that emerges from a simpler tendency to persevere with previous choices (Katahira, 2018; Sugawara & Katahira, 2021). However, another group defends the bias's validity that cannot be explained away by perseveration (Palminteri, 2023). My results from this model showed that the bias for false information survived in study 1 (t(46) = 3.13, p < 0.001) and study 2 (t(56) = 5.24, p < 0.001). The bias for true information, however, survived in study 1 (t(46) = 2.43, p = 0.01) but not in study 2 (t(56) = 1.15, p = 0.25). It should be noted that the choice

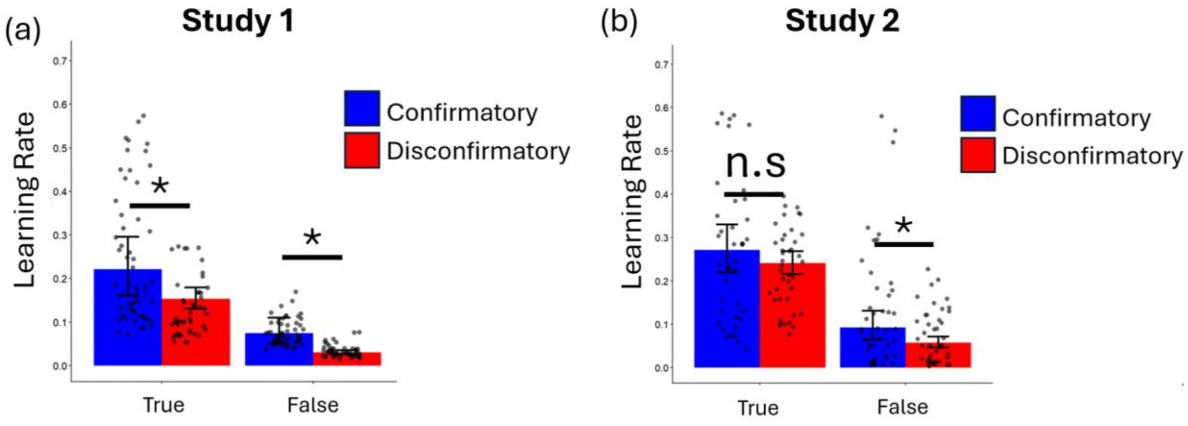


Figure 3.6: Gradual Perseveration Model Estimates. (a) The estimates from the gradual perseveration model showing confirmation bias for study 1 for both true (t(46) = 2.43, p = 0.01) and false (t(46) = 3.13, p < 0.001) information. **(b)** The estimates of the model in study 2 showed confirmation bias for false information (t(56) = 5.24, p < 0.001) but not for true (t(56) = 1.15, p = 0.25). n.s: not significant, *p < 0.05, hierarchical t-test.

3.4 Discussion

Debunking reduces learning from false information compared to true information, but it is less effective when false information confirms vs disconfirms one's beliefs, as demonstrated by a higher learning rate for confirmatory vs disconfirmatory false information – a confirmation bias. In two reinforcement learning (RL) studies, through computational models, I showed that confirmation bias exists when faced with false information, not only replicating the well-documented asymmetric treatment of prediction errors as the mechanism behind this bias for true information (Chierchia et al., 2023; Palminteri, 2023, 2025; Palminteri et al., 2017; Palminteri & Lebreton, 2022) but also extending it to the misinformation domain. The winning model had different learning rates across feedback (confirmatory vs disconfirmatory) and accuracy (true vs false) domains, surpassing other models with fewer learning rates in a formal model comparison. Therefore, both model estimation and model comparison indicated sensitivity to information based on its feedback and accuracy.

Confirmation bias for false information could explain the phenomenon of echo chambers or filter bubbles (Flaxman et al., 2016). The key feature of these bubbles is the preference for confirmatory information; and, as my results demonstrate, confirmatory false information is harder to debunk because the earlier information whose veracity was unknown told people what they wanted to hear, acting as a reward. On social media, features such as “Like” that confirm one’s beliefs can indeed act as a reward akin to my RL task. Turner et al. (2025) modelled social media using RL whereby they treated its features - such as receiving likes on posts - as rewards that updated action values. They showed that when users repeatedly receive this type of social validation through likes and shares they develop habits around content they have learned to be rewarding. The issue is that this reinforcement mechanism doesn't distinguish between true and false information and only responds to whether content aligns with existing beliefs. One cannot establish causality here, but the outcome of such interactions is a polarized environment with different echo-chambers that boost their own confirmatory content and dispense with disconfirmatory information. According to my results, debunking is bound to be less effective in such environments as the content that is being fact-checked is most likely in line with the bubble’s beliefs.

This problem gets worse because in real life, people actively choose what information they like to consume - unlike in my experiment where I presented information *to* them. The exposure effect (Sears & Freedman, 1967) explains that people tend to seek information that is in line with what they already believe while avoiding information that contradicts their views. Given that we are inundated with so many choices today, it is much easier to choose sources that we know will confirm our beliefs, downweighting accuracy (Iyengar & Hahn, 2009; Karlsen et al., 2020). Bromberg-Martin and Sharot (2020) argue that confirmatory information produces rewarding "internal outcomes" akin to positive emotional experiences, reframing motivated reasoning as the rational pursuit of information that is rewarding in and of itself. Hart et al.'s (2009). A meta-analysis supports this (Hart et al., 2009), demonstrating that while accuracy matters, the drive toward confirmatory information intensifies when people want to defend their beliefs. Therefore, information seeking becomes a tool to feel good, whereby people seek information that brings them positive emotions rather than accurate information. This keeps happening over and over, creating a loop whereby not only do we interpret information in a biased manner, but our choices also control what information we see in the first place. This creates a system that favours information matching our beliefs, no matter the veracity. Therefore, providing a mechanistic account of why people seek information from unreliable sources is an important future direction.

The main finding of this chapter – confirmation bias for false information - applies to Large Language Models (LLMs) like ChatGPT as well. These AI systems are trained to match user preferences using methods like Reinforcement Learning from Human Feedback (RLHF). Basically, the models get rewarded when they produce responses users deem helpful (Ouyang et al., 2022). While this makes the AI easier to use, it also makes it sycophantic (Rathje et al., 2025), telling people what they want to hear instead of giving accurate information (Perez et al., 2022; Santurkar et al., 2023). My findings add another layer to this. When an LLM gives someone confirmatory false information, that person is more likely to learn from that information. At a larger scale, LLMs could create customised false information for each user based on their beliefs, creating mini echo chambers detached from reality.

My results shed light on both the benefits and limitations of debunking misinformation. On the bright side, debunking could be effective: participants learned less from false vs true information. This is in line with other studies showing that debunking could be an effective intervention (M.-P. S. Chan et al., 2017; Walter et al., 2020). The only downside is that debunking is less effective when false information tells us what we want to hear. This helps the continued influence effect, where misinformation keeps affecting people's thinking even after they've been corrected and accepted the correction (Johnson & Seifert, 1994). Previous research suggests misinformation leaves a mental 'trace' that's hard to erase from memory (Walter & Tukachinsky, 2020). My findings complement this by showing that this mental trace is stronger for misinformation that confirms our beliefs, potentially because we learn it better in the first place, making it harder to undo later.

In light of such limitations, an alternative approach is offered called prebunking (Van Der Linden, 2024; van der Linden et al., 2017). This approach is akin to a vaccine whereby you expose people to a small, weakened version of misinformation and explain why it's wrong beforehand, building mental resistance, the effectiveness of which has been shown in several studies (van der Linden et al., 2017; Roozenbeek & van der Linden, 2019). For instance, a prebunking version of the current RL task would be to bring the information cue presentation *before* the outcomes are observed (Vidal-Perez et al., 2025). Prebunking, however, is not a perfect solution. For instance, using RL, Vidal-Perez et al. (2025) found that even when people were warned in advance that a source was unreliable, they still learned from it - especially when they were under higher cognitive or working memory load. They also found that people showed stronger positivity bias when learning from sources they knew had low credibility. This raises important questions for future research: Is prebunking better than debunking at stopping people from learning false information? Does the confirmation bias I found with debunking also happens with prebunking?

Another factor that may contribute to learning from false information is cognitive or working memory (WM) load. Efficient learning in the current RL task requires tracking the values of four option pairs while simultaneously monitoring their veracity - filtering out irrelevant false information and integrating true information - which taxes WM. WM is indispensable to RL, even in simple instrumental tasks like the two-arm bandit (Collins, 2018; Collins et al., 2017; Yoo & Collins, 2022). In my RL task, WM could help filter out false information, but under higher load, this filtering capacity may fail, leading to greater integration of false information. Evidence from Vidal Perez et al. (2025) supports this hypothesis. They tested two versions of a prebunking paradigm that differed in cognitive load: their "Discovery Study" required participants to learn one option pair per block (lower load), while their "Main Study" required learning three pairs simultaneously (higher load). Under lower load, participants showed no significant learning from an unreliable agent - basically, the False cue in my task. However, under higher load, significant learning from the unreliable agent emerged. The positivity bias persisted across both studies when encountering the unreliable agent. These findings suggest that filtering out false information becomes more difficult as load increases, which might apply to my debunking paradigm despite the obvious methodological differences. Given that processing information on environments replete with misinformation such as social media is also cognitively demanding (Pittman & Haley, 2023), where people should ignore the abundant bots and shoddy accounts (Chuai et al., 2023b), it is plausible that one reason why people are vulnerable to false information in such environments is failure to filter out misleading content, especially when they confirm one's beliefs. Future studies could shed light on this by directly manipulating WM load (e.g., by varying the number of stimuli to learn - set size) and testing whether false information integration changes as a function of it.

While the RL framework employed here provides a useful method for quantifying confirmation bias, it is important to acknowledge its limitations. One limitation is the restriction of learning rates (α) to values between 0 and 1, meaning that a negative learning rate is impossible. In a standard Rescorla-Wagner model, a positive outcome (e.g., +10 points) generates a positive prediction error, which mathematically forces an increase in the expected value of the associated option. However, when the same outcome turns out false it could create a computational paradox. Consider this example: a participant chooses Option A, sees a +10 win, but is then shown a Cross (False). If the participant interprets this as a negative scenario, they might lower their estimate of Option A. Yet, to reduce the value of Option A the model would mathematically require a negative learning rate (multiplying the positive error by a negative number). Since this parameter is conceptually invalid in standard RL, the model cannot capture this potential scenario. Instead, the model is forced to either increase the value towards the displayed reward or to suppress the learning rate toward zero, meaning that it cannot capture a case where the participant lowers the value of the option. Further, standard RL algorithms are associative, meaning that they update values incrementally based on the reward history. However, processing false information in this task involves retrospective judgement. For example, a participant understands that a "Cross" changes the meaning of the +10 they just saw. This limitation suggests that while RL can model how values change, it does not capture the reasoning participants use to determine what those values actually are when information is unreliable. To address these limitations, future studies could employ Bayesian learning models (Behrens et al., 2007; Diaconescu et al., 2014a) or Latent Cause Models (Gershman & Niv, 2010). Bayesian models track the reliability of the information source. Unlike RL, which assumes inputs are always "true" rewards, a Bayesian agent estimates the probability that a signal is correct. If the agent infers that the False cue indicates a reliability of 0%, the model mathematically inverts the prediction error in the update step. Alternatively, Latent Cause models shift the mechanism from value updating to structural inference. Rather than assuming all outcomes belong to a single state, the agent infers distinct "hidden states" that generate observations. In this framework, the agent can infer that the False cue signals a specific "Noise State," allowing the model to ignore the false feedback and protect the true value estimate.

These findings paint a bleaker picture than earlier studies, demonstrating that misinformation can exert a noticeable influence on our learning even when tested with basic, abstract stimuli in controlled settings. The confirmation bias for false information, in particular, is a finding that showcases how vulnerable we are to false claims that are in line with our existing viewpoints. Given that the algorithms are tailored to our beliefs (Glickman & Sharot, 2024; Vellani et al., 2024), these kinds of confirmatory information are bound to be common, a sizeable chunk of which will be misleading or false. The computational modelling approach I employed here was useful in understanding both the potential effectiveness of debunking interventions and the reason behind their limitations. Combating misinformation therefore requires a better understanding of the reward-based cognitive biases that shape how we learn from false information.

Chapter 4: Optimistic update bias in response to false information

4.1 Introduction

Encountering false information is a phenomenon animals and humans have long had to grapple with. From monkeys misleading their peers where food sources have been hidden (Mitchell, 1986), human scientists falsifying and publishing false data (Gopalakrishna et al., 2022) to governments spreading propaganda in the pursuit of furthering their own political agendas (Waight et al., 2025), encountering misleading information has been a long-standing problem. But whilst not a new phenomenon, exposure to false information has – amongst humans at least – become a more urgent and pressing concern in recent years due to the increased volume, velocity, and variety I now encounter on a regular basis (Ceylan et al., 2023; Pennycook & Rand, 2021; Van Der Linden, 2024). This has been driven by technological advances that have lowered the cost of producing realistic false information (open-source AI tools can be used to generate highly realistic deepfakes, for instance), decreased cost attached to disseminating false information (automated 'bots' can be used, for instance) (Ceylan et al., 2023; Pennycook & Rand, 2021), and low levels of regulatory oversight that exist to constrain content posted and shared via social media. Alongside this, recommendation systems technology platforms use look to prioritise engagement (rather than prioritising accuracy), which has been argued to further facilitate the spread of misinformation (Pennycook & Rand, 2021; Vosoughi et al., 2018) which can often be more novel, surprising and likely to garner attention and engagement.

This recent surge of exposure to false information has been blamed for a range of negative outcomes, including: failures to respond to climate change (Brulle & Roberts, 2017; van der Linden et al., 2017), political unrest (Ruohonen, 2024) and poor medical decision making, such as vaccine hesitancy (Loomba et al., 2021; Pierri et al., 2022; Roozenbeek et al., 2020; Zimmerman et al., 2023). Given the threat it poses, there has been considerable debate surrounding how best to counteract it. One proposal (Christner et al., 2024; Van Der Linden, 2024) is for content to be moderated by attaching labels to information, which inform cases where information is suspect and questionable. This has been contentious of late as prominent media platforms Meta and X (formerly Twitter) have abolished attempts to do this, citing ineffectiveness, proneness to bias and conflicts with free speech (Isaac & Schleifer, 2025; J. Taylor, 2023).

Here I investigate the degree to which such labels could enable humans to moderate the degree to which they learn from new information. On the one hand, a considerable body of evidence from psychology and neuroscience suggest that the extent to which individuals learn is not fixed. Rather, the degree to which information is integrated and used to update beliefs varies with a host of factors, including arousal (Filipowicz et al., 2020; Nassar et al., 2012), how volatile an environment I are in (Behrens et al., 2007; Pearce & Hall, 1980; Pulcu & Browning, 2019), arousal (Browning et al., 2015; Garrett et al., 2018), working memory (Z. Cheng et al., 2024), and mood (Kao et al., 2023). But on the other hand, there are many examples where it seems learning cannot be completely curtailed when people get information that they explicitly told is irrelevant or false (Ecker et al., 2010; Ross et al., 1975; Tversky & Kahneman, 1974). A well-known case is the *continued influence of misinformation* (Ecker et al., 2010) whereby individuals continue to recall narratives they encountered that were subsequently retracted or discredited.

One factor that could influence the degree to which learning can be adjusted in the face of false information – which I explore here – is that of valence; whether the information is better or worse than expected to begin with. A range of evidence suggests that individuals tend to integrate better than expected information over worse than expected information (Garrett et al., 2014, 2014; Garrett & Sharot, 2017a; Kuzmanovic et al., 2019c; Kuzmanovic & Rigoux, 2017; Sharot et al., 2011), a pattern which has been argued to help generate and sustain optimism over time. If learning can be “undone” in cases where information is revealed to be false but was better than expected in the first instance, potentially this act of “mentally undoing” the initial learning requires a greater degree of effort, compared to instances in which the initial information was worse than expected (and therefore didn’t actually change beliefs by much in the first place). If, in addition, individuals engage in motivational reasoning (Kunda, 1990) as part of this revisionist process, they are less likely to possess the motivation to undo learning that restores them back towards a worse belief (which will be the case when better than expected information is revised) compared to cases where such a revision restores them back towards a better belief (which will be the case when worse than expected information is revised).

To test this idea, I adapted a classic belief updating paradigm (Sharot et al., 2011) in which participants were presented with information about the likelihood of experiencing different adverse life events in the future. This information was then explicitly labelled as either true or false. This manipulation allowed us to answer 2 key questions: (1) whether belief updating was attenuated when information was shown to be false; (2) whether the valence bias shown in the past to exist in response to true information (Garrett et al., 2014; Garrett & Sharot, 2017a; Kuzmanovic et al., 2019c; Kuzmanovic & Rigoux, 2017; Sharot et al., 2011; Sharot & Garrett, 2016) persists when information is shown to be false.

4.2 Methods

4.2.1 Participants

127 students were recruited from the University of East Anglia SONA subject pool to participate in the study. 19 of the 127 were excluded from all analysis (1 participant did not pass attention checks during the task, 3 did not complete the task, 13 failed catch questions in the questionnaires, and 3 did not have trials I could assign to each of our four experimental conditions). The final sample size was 108 (mean (SD) age: 19.48 (1.51)). Power analysis based on effect sizes from a previous study (d=0.40) using this paradigm (Garrett & Sharot, 2017a) indicated that a sample size of 80 was sufficient to detect an effect size of with 95% power and an alpha level of 0.05 (paired sample t-test comparing good versus bad news). I increased the sample size further to enable us to use Bayes Factors where a sample size of at least 100 participants is recommended (Fu et al., 2021, 2022) to infer evidence for/against the null. The study was approved by the University of East Anglia's ethics committee. Participants received study credits as compensation for participating in the study.

4.2.2 Behavioural Task and Stimuli

The task was adapted from a belief updating task previously used to investigate optimistic updating biases (Khalid et al., 2024; Kuzmanovic et al., 2019c; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Sharot et al., 2011; Sharot & Garrett, 2016) and followed best practices for using the task (Sharot & Garrett, 2022).

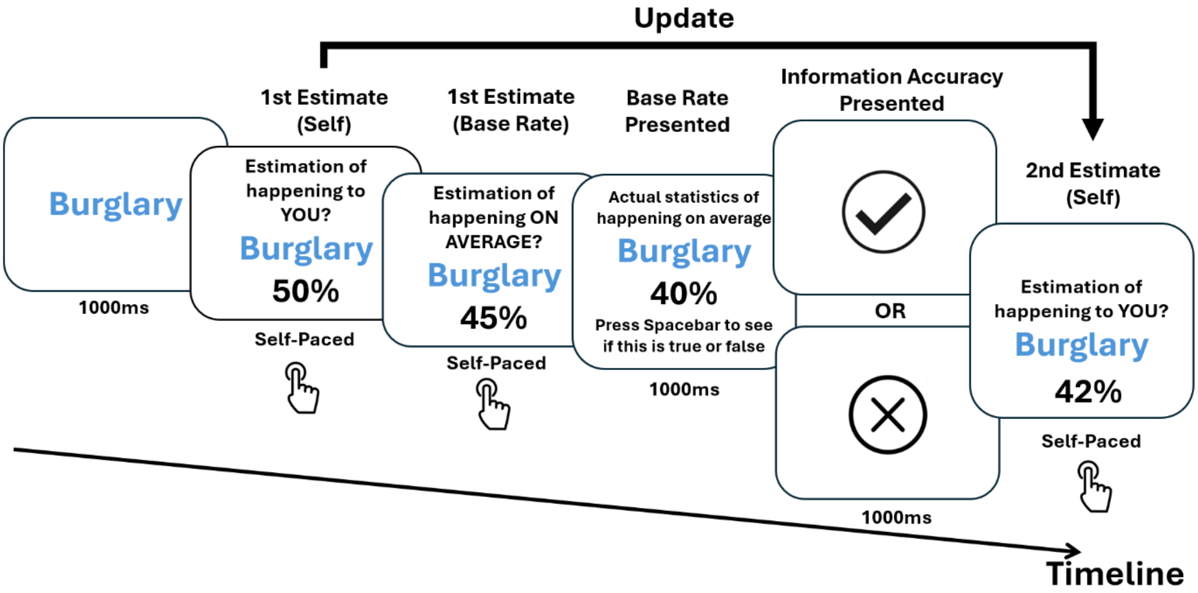

Update
1st Estimate (Self)
1st Estimate (Base Rate)
Base Rate Presented
Information Accuracy Presented
2nd Estimate (Self)
Burglary
1000ms
Estimation of happening to YOU?
Burglary
50%
Self-Paced
Estimation of happening ON AVERAGE?
Burglary
45%
Self-Paced
Actual statistics of happening on average
Burglary
40%
Press Spacebar to see if this is true or false
1000ms
OR
1000ms
Estimation of happening to YOU?
Burglary
42%
Self-Paced
Timeline

Figure 4.1: Behavioural task. On each trial, participants were presented with a short description of an adverse event and asked to imagine the event happening to them in the future. They were then asked to estimate how likely this event was to occur to them in the future and then to estimate how likely the event was to happen to them on average (the order of these two estimates was randomised). They were then presented with the probability of that event occurring on average (in a demographically similar population) and then with a cue indicating whether this statistic had been true (tick) or false (cross). Finally, participants are asked to provide their estimate how likely this event was to occur to them in the future. Shown here is an example of Good News (as the Average Presented to Participants is lower than their initial self-estimate). In other cases (where the Average Presented to Participants is higher than their initial self-estimate) participants received Bad News. See Supplementary Materials for Examples of the 4 different trial types (Good News True, Bad News True, Good News False and Bad News False). Update is quantified as the change in 1st and 2nd Self Estimates. Estimates of the Base Rate are used in the computational modelling to infer relative personal knowledge participants might privately hold about each event.

Stimuli consisted of two lists of 25 different negative life events (e.g., domestic burglary - see **Appendix 4.6** for events used). Events and their statistics were obtained from stimuli used in the original study (Sharot et al., 2011), which compiled the statistics of each event occurring at least once to someone from the United Kingdom from reputable online resources (including the Office for National Statistics and PubMed). Very rare or very common events were not included; all event probabilities lay between 10 and 70% and were normally distributed around the midpoint of the range (Sharot & Garrett, 2022), see **Appendix 4.2.** To ensure that the range of possible overestimation was equal to the range of possible underestimation, participants were told that the range of probabilities lay between 3 and 77%, and they were only permitted to enter estimates within this range. Each participant was randomly assigned to one of the two lists for true trials and the other list for false trials. When a list was designated as the false list, the event statistics were randomly shuffled so that the statistics didn't match the event descriptions (e.g., the likelihood for domestic burglary was shown for bicycle theft), but the statistical properties of the base rates (the median, range, distribution, etc.) remained unchanged. The two lists were then merged to create a final list of 50 events with 25 accurate statistics and 25 false statistics.

4.2.3 Behavioural Task

On each trial, one of the 50 events was shown. Participants were then prompted to estimate how likely the event was to happen to them personally in the future (E_1) and how likely it was to occur on average in the population (eBR; their estimate of the base rate). In half of the trials, the order of the two estimations (E_1 and eBR) was reversed (i.e. E_1 followed by eBR). After this, the base rate of the event happening to someone in the same socioeconomic environment as the participant (BR; the base rate) was provided. Participants were then prompted to press the spacebar to see the accuracy of the statistic. They then saw a stimulus indicating whether the statistic they had been shown was accurate (tick) or false (cross). Finally, they were told to estimate how likely the event was to happen to them again (E_2). There were no time constraints for submitting responses (E_1, eBR, E_2). The task was created using Qualtrics.

4.2.4 Questionnaires

At the end of the experiment, participants were asked to complete three psychiatric questionnaires: Beck Depression Inventory (Beck et al., 1961), Obsessive Compulsive Inventory (Foa et al., 2002), and the Schizotypy short scale (Mason et al., 2005). In each of the questionnaires, a catch check question was included, which was used to exclude inattentive participants (Zorowitz et al., 2023). Participants (N=13) were excluded if they got 1 or more of the catch questions wrong.

4.2.5 Controls

At the end of the experiment, participants rated stimuli on six-point scales for Negativity ("How negative you found this event?" From 1 = Not at all to 6 = Very), Prior Experience ("Has this event happened to you before?" From 1 = never to 6 = very often), Vividness ("How vividly could you imagine this event?" From 1 = not vivid to 6 = very vivid), Familiarity ("Regardless of if this event has happened to you before, how familiar do you feel it is to you from TV, friends, movies and so on?" From 1 = not at all familiar to 6 very familiar); and Arousal ("When you imagine this event happening to you how emotionally arousing is the image in your mind?" From 1 = not arousing at all to 6 = very arousing). These Ratings were included as covariates in control analysis.

4.2.6 Behavioural Analysis

Trials were categorised into 4 types (Good News True, Bad News True, Good News False and Bad News False) in a 2*2 within-subject repeated measures design with accuracy and valence as factors. Accuracy (True/False) was determined according to whether the trial had presented an accurate or false cue. Valence (Good/Bad) was determined based on whether participants underestimated or overestimated the likelihood of an event happening to them personally (i.e. E_1) relative to the provided base rate (BR). Trials where the initial estimate was equal to the base rate (E_1 = BR), ~ 2% of total trials, were excluded as they could not be classified. Update was calculated for each trial such that positive updates indicate a change toward the probability presented [update (good news) = first estimate - second estimate] and negative updates indicate a change away from the probability presented [update (bad news) = second estimate - first estimate].

Update scores were entered into a 2*2 repeated measures ANOVA with valence (good news or bad news) and accuracy (true or false) as factors. To determine whether there was evidence of biased updating for true trials, false trials or both, I followed up with planned paired sample t-tests on good news vs bad news separately for true (Good True vs Bad True) and false (Good False vs Bad False) trials. The ANOVA was run using the ezANOVA package (Lawrence, 2016). I then repeated this ANOVA analysis, this time applying a stricter classification of good and bad news (Garrett & Sharot, 2014). Specifically, I excluded any trials where BR was higher than E_1 but lower than eBR (as this could be perceived as bad news if comparing BR with E_1 but good news if comparing BR with eBR) and trials where BR was lower than E_1 but higher than eBR (as this could be perceived as good news if comparing BR with E_1 but bad news if comparing BR with eBR).

Next, I complemented this analysis using linear mixed-effects models (LMM). Trial-by-trial updates were entered as the dependent variable, with valence (coded as 1 = good news, -1 = bad news) and accuracy (1 = true, -1 = false) as predictors along with their interaction. Predictors were included as random effects and allowed to vary by participant. The model was implemented in the syntax of R as follows

LMM1: Update ~ Valence*Accuracy + (1 + Valence*Accuracy | Participant)

Next, I ran a second LMM and now included potential confounds (Sharot & Garrett, 2022).

For each trial, an estimation error (EE) term was calculated as the absolute difference between the presented probability (BR; the new information) and the participant's initial estimate (E_1) for that trial:

EE = $|E_1 - BR|$

EE's on each trial and the 5 subjective ratings for each event shown on each trial (Vividness, Past Experience, Arousal, Familiarity and Negativity) were entered as additional predictors in the model. Once again, all predictors were included as random effects. The model was implemented in the syntax of R as follows:

LMM2: Update ~ Valence*Accuracy + EE + Vividness + Arousal + Familiarity + Negativity + PastExperience + (1+ Valence*Accuracy + EstErr + Vividness + Arousal + Familiarity + Negativity + PastExperience | Participant)

The two models were fitted in R using the lmer package (Bates et al., 2015) with significance tests implemented by lmerTEST (Kuznetsova et al., 2017).

To test whether null effects for the Valence*Accuracy interaction in each model (LMM1 and LMM2) provided evidence in favour of the null (i.e. that the strength of the valence effect was the same for true and false trials), I used the brms package (Bürkner, 2017) with weakly informative priors, 16000 iterations (3000 for warmup), and four chains to generate posterior distributions for all model parameters. These were then used to generate Bayes Factors (BFs) in bayestestR (Makowski et al., 2019). I used Jeffreys' scale to interpret BFs (Jeffreys, 1961; Wagenmakers et al., 2011) according to which: BFs < 1/30 are interpreted as extreme evidence supporting the null hypothesis; BFs 1/10-1/30 reflect strong evidence in favour of the null; BFs: 1/3-1/10 reflect moderate evidence in favour of the null; BFs 1/3-1/1 reflect anecdotal evidence in favour of the null.

4.2.7 Computational Modelling

I adapted the computational model used by Kuzmanovic & Rigoux (2017, see also Garrett & Sharot, 2022) to test if the relationship between trial-by-trial errors and subsequent updates was modulated by accuracy and valence. I note that the models I test here do not attempt to make claims as to what time points in the task belief change occurs. I apply these models primarily as an analytical tool to understand the degree to which error signals in the task correspond to subsequent belief change and test whether this relationship is modulated by accuracy, valence and their interaction.

In each model, update on each trial (t), i.e. the change between participants' first and second self-estimates [calculated as: update (good news) = first estimate - second estimate; update (bad news) = second estimate - first estimate] was predicted via the following equation:

$$\text{Update}_t = \alpha * EE_t * (1 - rP_t * w)$$

rP (relative personal knowledge) was calculated by comparing the estimated base rate (eBR) and the first estimate (E_1) as:

$rP_t = (eBR_t – E_{1\,t}) / (eBR_t – 1)$ *if* $E_1 < eBR$

$rP_t = (E_{1\,t} – eBR_t) / (77– eBR_t)$ *if* $E_1 > eBR$

$rP_t = 0$ *if* $E_1 = eBR$

rP ranges from 0 to 1 where rP = 0 indicates the person does not see themselves as different from the population, rP = 1 indicates the greatest disparity possible between the individual's perceived likelihood of the event occurring to themselves and the population's likelihood (which might occur if say the individual has a range of reasons – such as specific lifestyle choices, family history, etc.) - why they believe a population statistic is not relevant for them.

w is a free parameter which indexes the degree to which rP impacts belief updating. When W is 0, rP has no influence on belief updating, while when W is 1, rP has maximum influence on belief updating.

α is a learning rate, which governs the degree to which participants updated their beliefs in response to the size of the estimation errors (Sharot et al., 2011; Garrett et al., 2014; Garrett et al., 2018). I tested 4 different models with different numbers of α (α = 2, 3 or 4). By varying the number of learning rates and how these selectively parsed information according to whether information presented was true/false, good/bad, I were able to test if estimation errors were integrated to differing degrees for these two factors – valence and accuracy. The 4 models were formulated as follows:

Model 1 (M1)

$Update_t = \alpha_{true} * EE_t * (1 – rP_t * w)$ if information accuracy cue = True

$Update_t = \alpha_{false} * EE_t * (1 – rP_t * w)$ if information accuracy cue = False

Free parameters (n=3): $\alpha_{true,}$ $\alpha_{false,}$ w

Model 2 (M2)

$Update_t = \alpha_{true} * EE_t * (1 – rP_t * w)$ if accuracy = True

$Update_t = \alpha_{false,\ goodnews} * EE_t * (1 – rP_t * w)$ if accuracy = False and valence = Good

$Update_t = \alpha_{false,\ badnews} * EE_t * (1 – rP_t * w)$ if accuracy = False and valence = Bad

Free parameters (n=4): α_{true}, $\alpha_{false,\ goodnews}$, $\alpha_{false,\ badnews}$, w

Model 3 (M3)

$Update_t = \alpha_{false} * EE_t * (1 – rP_t * w)$ if accuracy = False

$\text{Update}_t = \alpha_{\text{true, goodnews}}* \text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = True and valence = Good

$\text{Update}_t = \alpha_{\text{true, badnews}}* \text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = True and valence = Bad

Free parameters (n=4): α_{false}, $\alpha_{\text{true, goodnews}}$, $\alpha_{\text{true, badnews}}$, w

Model 4 (M4)

$\text{Update}_t = \alpha_{\text{true, goodnews}}*\text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = True and valence = Good

$\text{Update}_t = \alpha_{\text{true, badnews}}*\text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = True and valence = Bad

$\text{Update}_t = \alpha_{\text{false, goodnews}}*\text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = False and valence = Good

$\text{Update}_t = \alpha_{\text{false, badnews}}*\text{EE}_t*(1 - \text{rP}_t*\text{w})$ if accuracy = False and valence = Bad

Free parameters (n=5): $\alpha_{\text{false, goodnews}}$, $\alpha_{\text{false, badnews}}$ $\alpha_{\text{true, goodnews}}$, $\alpha_{\text{true, badnews}}$, w

In each model, I converted the predicted Update on each trial into a predicted 2nd Estimate $(\widehat{E_2})$ on each trial by adding or subtracting the predicted Update from participants' 1st Self Estimate (E_1), depending whether the base rate (BR) presented on that trial was lower than E_1 (in which case beliefs shift down) or above E_1 (in which case beliefs shift up as the base rate presented was above E_1):

$\widehat{E_2}_t = E_{1t} + \text{Update}_t$ if $\text{BR}_t > E_{1t}$

$\widehat{E_2}_t = E_{1t} - \text{Update}_t$ if $\text{BR}_t < E_{1t}$

2nd estimates predicted on each trial $(\widehat{E_2}(t))$ were compared to participants' actual 2nd estimates $(E_2(t))$ to find the best fitting parameters in the model fitting process.

4.2.7.1 Model Fitting Procedure

I fitted the models hierarchically using Expected Maximization (EM) algorithm (Huys et al., 2011) in the Julia language (Bezanson et al., 2012) version 1.9.4. Hierarchical parameter estimation has been shown to provide superior cross-validation performance on unobserved data (Scheibehenne & Pachur, 2015). See **Chapter 2** for the full details.

To find the best fitting set of parameters for each model (given the data I observed) using the EM algorithm, I used a log-likelihood function. For each observation, I calculated the probability density of the observed second estimate (E_2) given the model's predicted second estimate and other parameters. Specifically, I modelled E_2 as following a normal distribution (similar to (Nassar et al., 2021)) but truncated the distribution such that it was bounded between 3 and 77 (the range of possible values in our task):

$$\log L = \sum\nolimits_{t=1}^{N} \quad \log\left[\frac{\phi\left(\frac{E_2(t) - \widehat{E_2}(t)}{\sigma}\right)}{\sigma\left[\Phi\left(\frac{77 - \widehat{E_2}(t)}{\sigma}\right) - \Phi\left(\frac{3 - \widehat{E_2}(t)}{\sigma}\right)\right]}\right]$$

Where:

- $E_2(t)$ is the observed second estimate for trial t
- $\widehat{E_2}(t)$ is the model's predicted estimate for trial t
- σ is the standard deviation as a free parameter
- ϕ is the standard normal cumulative distribution function (CDF)
- Φ is the standard normal probability density function (PDF)

4.2.7.2 Model Comparison

I then compared the fit of the four models by calculating unbiased subject-level leave-one-out cross-validation (LOOcv) scores for each participant for each model. The LOOcv scores were fed into the mbb-vb-toolbox in MATLAB (Daunizeau et al., 2014). See **Chapter 2** for the full details.

4.2.7.3 Statistical Tests on the Learning Rates

Once the winning model was identified, I reparametrized it such that one parameter indexed the optimistic update bias for true information, one for false information, and a third captured their interaction. I then estimated these parameters again using hierarchical EM. See the full details in **Chapter 2.**

4.2.7.4 Model Recovery

I ran model recovery to validate the degree I could robustly identify each of our 4 models (M1-M4). The basic logic of this approach is that if the underlying data has verifiably been generated by one of the four specific models, this model should outperform the other 3 in model comparison. See **Chapter 2** for the full details.

4.2.7.5 Parameter Recovery

To evaluate parameter identifiability, I conducted a parameter recovery analysis on the winning model. I simulated behaviour for 200 synthetic participants, each completing 200 trials, using parameter values randomly drawn from uniform distributions spanning the empirically observed ranges. The same model-fitting procedure used for real participants was then applied to the simulated data. Recovery success was assessed by comparing the true and recovered parameter values using Pearson correlations, and values higher than 0.80 were deemed high enough for a successful recovery. See **Chapter 2** for the full details.

4.2.7.6 Simulations

To qualitatively examine each model's capacity to reproduce the behavioural patterns I observed, I simulated data for each of the four models (M1–M4). For each model, I generated data for 500 synthetic participants, each completing 50 trials. Trial characteristics (base rates, number of true/false trials, range of estimations) matched the structure and range of the actual task. Parameter values used for each simulation were set to the average value of the parameters observed in the real data.

4.3 Results

Biased updating in response to true and false information. Update scores from each participant were entered into a 2*2 repeated measures ANOVA with Accuracy (True/False) and Valence (Good/Bad) as within-subject factors. This showed a main effect of valence (F(107) = 297.27, p < 0.001) with greater updating for good news compared to bad news, a main effect of accuracy (F(107) = 43.47, p < 0.001) with greater updating for true compared to false trials, and no interaction between valence and accuracy (F(107) = 2.05, p = 0.15). Replicating findings from previous studies (Kuzmanovic et al., 2019a; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Sharot et al., 2011; Sharot & Garrett, 2016), I found a significant difference in belief updating between good and bad news for true (t(107) = 5.74, p < 0.001; paired t-test, 75.9% of participants updated more in response to True Good compared to True Bad). But I also found a similar difference in updating between good and bad for the false trials (t(107) = 5.09, p < 0.001, 75.9% of participants updated more in response to False Good compared to False Bad).

These results are not explained by differences in the distribution of statistics presented to participants (Garrett & Sharot, 2017a, 2023; Sharot & Garrett, 2022) – these were deliberately engineered (by the experimental design, see **Methods**) such that statistics labelled as true and statistics labelled as false during the experiment each assumed a normal distribution centred around the midpoint of the scale (see **Appendix 4.2** for histogram plots of the base rates presented). The results are also not the result of potential misclassification of trials into good or bad (Garrett & Sharot, 2014). To check this, I reran this analysis, this time applying a more stringent method of classifying trials into good and bad news (See **Methods** and **Appendix 4.7** for full details). This again revealed a main effect of valence (F(1,107) = 318.76, p < 0.001), a main effect of accuracy (F(1,107) = 37.88, p < 0.001) and no significant interaction (F(1,107) = 1.38, p = 0.24). Again, paired t-tests showed greater updating for good compared to bad news for true (t(107) = 5.26, p < 0.001) and for false trials (t(107) = 4.86, p < 0.001).

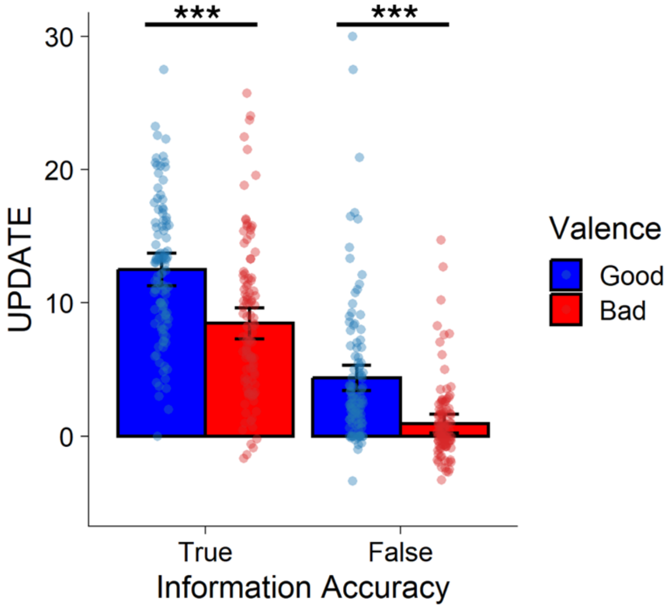


Figure 4.2: Biased updating in response to true and false information. Participants reduced the degree to which they used the information presented and updated their beliefs following the receipt of false compared to true information (Main Effect of Accuracy: F(107) = 43.47, p < 0.001). Participants also updated their beliefs more when information was good news (presented an opportunity to adjust beliefs in a positive direction) than after receiving bad news (that called for adjustments in a negative direction, Main Effect of Valence: F(107) = 297.27, p < 0.001). Planned paired sample comparisons showed that the valence effect was present both for true (t(107) =

5.74, p < 0.001, replicating previous results) and for false trials (t(107) = 5.09, p < 0.001). There was no interaction between information accuracy and valence (F(107) = 2.05, p = 0.15). Error bars represent SEM. ***p < 0.001, two-tailed paired sample t-test.

Together, these results suggest that whilst participants can modulate the degree to which they update beliefs in response to new information (according to how accurate the information is), biased updating exists both in response to true and false information. Next, I complemented this analysis using linear mixed-effects models (LMM). The motivation for this was that it allowed us to conduct a Bayes Factor analysis to interrogate whether the lack of an interaction I observed above provided evidence in favour of the null (i.e. evidence in favour that the valence effect was similarly strong for true and for false trials). Trial-by-trial updates were entered as the dependent variable, with valence, accuracy and their interaction as predictors. This again revealed a main effect of valence (t(101.32) = 7.47, p < 0.001), a main effect of accuracy t(107.84) = 18.18, p < 0.001) and no accuracy by valence interaction (t(128.63) = 1.13, p = 0.25) (see **Table 4.1** for full statistics). I then ran a Bayesian Factor analysis on the interaction (see **Methods**). This revealed strong evidence (Jeffreys, 1961; Wagenmakers et al., 2011) in favour of the null (BF_{10} = 0.030, Bayes factors < 1 indicate support for the null over the alternative hypothesis), suggesting that the valence effect for false trials was indeed of a similar magnitude to the valence effect for true trials.

Predictor	Estimate	std. Error	df	CI	Statistic	p
(Intercept)	0.06	0.03	105.89	0.00 – 0.11	2.03	**0.043**
Valence	0.17	0.02	101.32	0.12 – 0.21	7.47	**<0.001**
Accuracy	0.35	0.02	107.84	0.31 – 0.39	18.18	**<0.001**
Valence × Accuracy	0.02	0.02	128.63	- 0.01 – 0.05	1.13	0.258
N $_{Participant}$	108					
Observations	5242					

Table 4.1: Linear mixed effects model. Fixed Effect Estimates and accompanying statistics from a linear mixed-effects model predicting updates on each trial from Valence, Accuracy, and their interaction (Valence × Accuracy).

Optimistic update bias survives after controlling for confounds. Next, I ran a second LMM, this time controlling for potential confounds (Sharot & Garrett, 2022):

Estimation Errors (the absolute difference between participants' initial estimations and the information provided) and all subjective ratings (see **Methods** and **Table 2.2**). This again revealed a main effect of valence (t(98.86) = 10.17, p < 0.001), a main effect of accuracy (t(107.15) = 19.18, p < 0.001) and no interaction (t(105.85) = 1.26, p = 0.20). There were also significant effects for Estimation Error (t(105.46) = 14.23, p < 0.001) and Past Experience (t(79.42) = -3.44, p = 0.001). The remaining subjective rating scores were not significant (Vividness: t(179.54) = -0.48, p = 0.62; Familiarity: t(207.52) = -0.93, p = 0.35; Arousal: t(216.79) = -0.26, p = 0.79; Negativity (t(100.59) = -1.29, p = 0.19) (see **Appendix 4.5** for full statistics of this model). The Bayes Factor analysis on the interaction again found evidence in favour of the null (BF_{10} = 0.038).

Together, these results suggest two key findings: (1) Belief updating can be modulated according to how true or false a piece of information is when this is made explicit; (2) Optimistic update bias (i.e. the valence effect) exists both in response to true and false information and is of a similar strength in each instance. Together, these findings result in false information having a larger impact on changing beliefs when the information is better (versus worse) than expected.

	Trial Type (Valence, Accuracy)			
	Good, True	**Bad, True**	**Good, False**	**Bad, False**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Update $^{V, A}$	12.50 (12.65)	8.46 (12.01)	4.37 (9.74)	0.94 (7.37)
Estimation Errors	17.94 (13.78)	24.29 (15.03)	21.34 (16.06)	24.87 (15.42)
N trials	8.39 (3.19)	16.1 (3.14)	8.32 (3.47)	15.7 (3.19)
Subjective Ratings (All scales 1 = low to 6 = high)				
Familiarity $^{V, A}$	3.42 (1.40)	2.96 (1.42)	3.51 (1.38)	2.93 (1.43)
Past Experience $^{V, A}$	1.64 (1.09)	1.24 (0.63)	1.81 (1.18)	1.20 (0.55)
Vividness $^{V, A}$	3.12 (1.43)	2.72 (1.36)	3.38 (1.40)	2.68 (1.36)
Emotional Arousal $^{V, A}$	3.03 (1.45)	2.89 (1.37)	2.93 (1.40)	2.89 (1.43)
Negativity $^{V, A}$	4.51 (1.34)	4.43 (1.26)	4.35 (1.33)	4.49 (1.31)

^{V} Main effect of valence $p < 0.05$
^{A} Main effect of valence $p < 0.05$

^{V*A} Interaction effect (Valence * Accuracy) p < 0.05

Table 4.2 Participants' Updates, Estimation Errors, Number of trials and subjective ratings of familiarity with stimuli, past experience, vividness, arousal, negativity.

Formal models suggest that estimates are updated by a prediction error signal (that quantifies the difference between prior expectations and outcomes) and a learning rate which governs the rate at which prediction errors drive belief change (Sutton & Barto, 2018). Next, I turned to computational modelling (Kuzmanovic & Rigoux, 2017) to examine the relationship between updates to beliefs and learning rates inferred during learning using an error term analogous to the prediction error in our task, the estimation error (Sharot et al., 2011), which quantifies the difference between prior beliefs (E_1) and the information provided (BR).

I adapted and extended the computational modelling approach used by Kuzmanovic & Rigoux (2017) to test how participants beliefs changed on each trial as a function of the size of the estimation error (i.e. how much the difference in prior beliefs and the information provided motivated subsequent belief change) and the degree to which this process was modulated by information accuracy and valence.

I did this by testing 4 different models, which were identical except for the number of learning rates and how these learning rates parsed out different types of information. Briefly (see **Methods** for full details), Model 1 (M1) had 2 learning rates: one for true and one for false information; Model 2 (M2) had 3 learning rates: one for true information and two for false information – one for false good news and another for false bad news; Model 3 (M3) also had 3 learning rates: one for false information and two for true information – one for true good news and another for true bad news. Model 4 (M4) had 4 learning rates (one for true good news, true bad news, false good news and false bad news). Each model also had 2 additional parameters: w and σ. w enabled the model to dampen the effect that estimation errors had in changing beliefs according to participants' personal knowledge about each event (e.g., personal risk factors for developing an illness, such as family history). σ was used to estimate the Gaussian that the estimates were drawn from in the model fitting process (see **Methods**). Model recovery tests I conducted, where I fit data simulated under one of the four models to all 4 models and examined whether the winning model is reliably identified as the model used to generate the data, validated that data generated by each model would be identifiable as the winning model (see **Methods** and **Appendix 4.3**).

Four learning rate model the best fit to the data. Leave-one-out cross-validation (LOOcv) scores from the 4 models were compared using a Variational Bayesian Approach (see **Methods**). This inferred that out of the 4 models, M4 provided a superior fit to the data (see **Figure 4.3(b)**). M4 had the highest model frequency (~75%), well above the chance level (25%), and an exceedance probability of 1. This indicates strong evidence that M4 - featuring learning rates that facilitate optimistic update bias for both true and false information - captured participants' behaviour best compared to the simpler models (M1–M3). As an additional check, I also compared M4 to the other 3 models using paired sample t-tests (FDR corrected for multiple comparisons) which also suggested M4 was a superior fit to participants estimations (M4 vs M3: t(107) = -2.11, p_adj = 0.03; M2 vs M4: t(107) = -5.17, p_adj < 0.001; M1 vs M4: t(107) = -5.45, p_adj < 0.001). 32.4% of participants had the lowest LOOcv score for M4 (18.5% for M1, 19.4% for M2 and 29.6% M3). Together, this model comparison analysis suggests that participants used estimation errors differentially to update beliefs depending on *both* whether the information that generated these errors was true or false and whether the information was better or worse than expected.

Higher learning rate for good news vs bad news for both true and false information. Next, I examined the pattern of learning rates from the winning model (M4) and tested for differences between them. This showed an optimistic update bias present both for true and false information (see **Figure 4.3(b)**). Specifically, participants learn more from good than bad news for true (t(107) = 8.56, p < 0.001, hierarchical t-test comparing $\alpha_{true\_goodnews}$ with $\alpha_{true\_badnews}$) and false (t(107) = 6.74, p < 0.001, hierarchical t-test comparing $\alpha_{false\_goodnews}$ with $\alpha_{false\_badnews}$) information. There was no interaction in the magnitude of the learning rates between valence (good versus bad) and information accuracy (true vs false): t(107) = 0.59, p = 0.55.

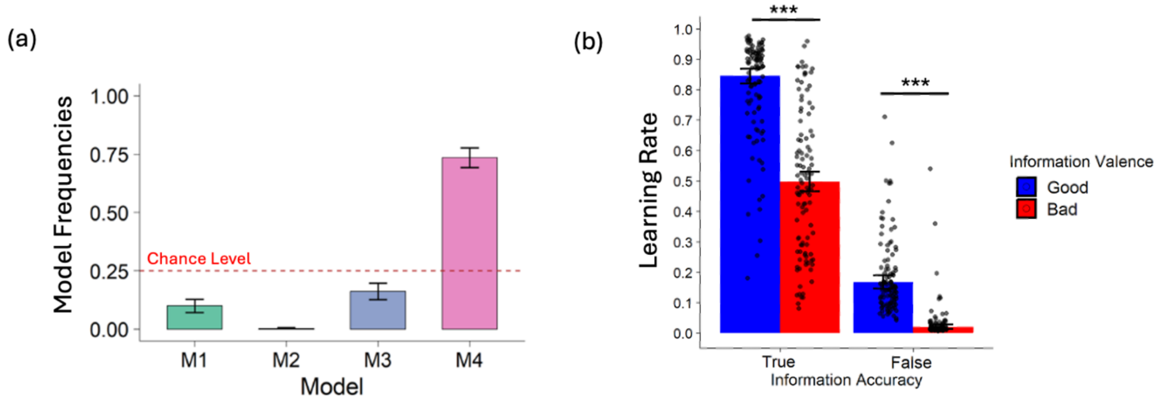


Figure 4.3: Model Frequencies and Estimates. (a) Estimated model frequencies from the VBA model comparison. Model 4 (M4) had the highest frequency, selected for approximately 74% of participants, with an exceedance probability (XP) of 1. This frequency is substantially higher than the chance level of 25% (indicated by the red dashed line), which represents the expected frequency if model selection were random across the four models (M1–M4). The model frequency reflects the proportion of the population best accounted for by each model (see Supplementary Material Figure 3 for additional model diagnostics). **(b)** Four learning rates from the winning model (M4) showed an optimistic update bias present both for true (t(107) = 8.56, p < 0.001, hierarchical t-test comparing $\alpha_{true\_goodnews}$ with $\alpha_{true\_badnews}$) and false (t(107) = 6.74, p < 0.001, hierarchical t-test comparing $\alpha_{false\_goodnews}$ with $\alpha_{false\_badnews}$) information. ***p < 0.001, hierarchical t-test.

Finally, to check how updating patterns varied under each model (given the parameters fit to the data), I simulated updating under each of the four models using the mean parameters from the model fitting process (see **Table 4.3)**. This showed a clear pattern whereby M1 enabled estimation errors to produce updating that was greater for true than false, but there was no valence bias (i.e. updating from good being greater than for bad) for either. M2 enabled estimation errors to produce updating that was greater for true than false and a valence bias for false only. M3 enabled estimation errors to produce updating that was greater for true than false and a valence bias for true only. By having 4 learning rates that allowed learning to vary for each combination of true, false, good, and bad, M4, the winning model enabled estimation errors to produce updating that was greater for true then false and a valence bias for both types of information in a pattern qualitatively similar to the actual updating I observed in the data (**Figure 4.4**).

Model	α_{true}	α_{false}	$\alpha_{true\_goodnews}$	$\alpha_{true\_badnews}$	$\alpha_{false\_goodnews}$	$\alpha_{false\_badnews}$	w	σ	LOOcv
M1	0.63 [0.58-0.68]	0.042 [0.02 – 0.06]	-	-	-	-	0.67 [0.57 – 0.76]	8.02 [7.43-8.67]	166 [162 - 169
M2	0.62 [0.57 – 0.68]	-	-	-	0.16 [0.12 – 0.21]	0.02 [0.009 – 0.04]	0.80 [0.72 – 0.87]	7.65 [7.14 – 8.20]	165 [162-169]
M3	-	0.051 [0.03-0.072]	0.84 [0.79-0.88]	0.50 [0.43 – 0.56]	-	-	0.72 [0.63-0.80]	7.24 [6.73-7.79]	164 [160-167]
M4	-	-	0.84 [0.79 – 0.88]	0.49 [0.43-0.56]	0.16 [0.12 – 0.21]	0.02 [0.01-0.039]	0.70 [0.61-0.79]	7.13 [6.63 – 7.65]	163 [160-167]

Table 4.3. Mean parameter estimates and Leave-One-Out cross-validation (LOOcv) scores from each of the 4 models. 95% confidence intervals are shown in square brackets. The mean parameter estimates were used for simulating data under each model shown in **Figure 4.4**.

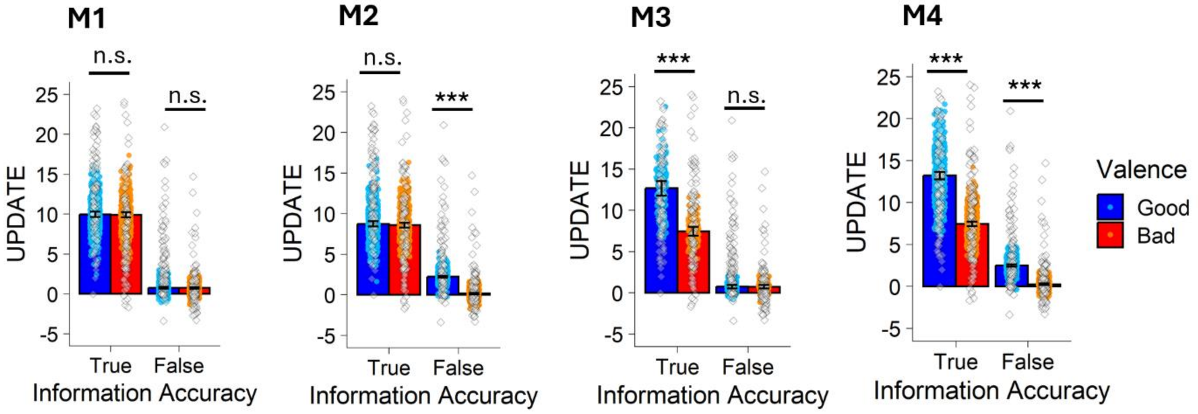


Figure 4.4: Simulations from each of my four models. M1 (which has 2 learning rates: one for true and one for false information) recovers the main effect of accuracy but cannot generate an effect of valence. M2 (which has 3 learning rates: a single learning rate for true information and two for false information – one for false good news and another for false bad news) recovers a main effect of accuracy but a valence effect for false information only. M3 (which also has 3 learning rates: a single learning rate for false information and two for true information – one for true good news and another for true bad news) recovers a main effect of accuracy and a valence effect for true information only. M4 (the winning model when fitted to participants responses) which has 4 learning rates (one for true good news, true bad news, false good news and false bad news) is able to capture the updating pattern I observe in the real data – a main effect of information accuracy and a valence effect both in response to true information and false information. Grey diamonds plot the real data from participants to be able to compare with the simulated data (plotted as coloured dots).

4.4 Discussion

Understanding the circumstances under which individuals are prone to integrating false information is important for understanding how erroneous beliefs such as conspiracy theories can persist in the face of flawed evidence (Douglas et al., 2017) and developing effective strategies to counter the prominent rise of misinformation (Hanley & Durumeric, 2024; Lewandowsky et al., 2012; Nyhan & Reifler, 2010) . Here, by adapting a classic belief updating task which presents individuals with information that can vary on two dimensions – valence (whether the information is better or worse than expected) and accuracy (whether the information is true or false) – I show that both of these factors exert important roles in governing the degree to which beliefs are influenced by the information and change as a consequence. First, I show that information is used to update beliefs to a greater degree when that information is revealed to be true compared to false. This suggests that warnings about the reliability of a piece of information may contribute towards determining whether that information serves to have an impact on altering beliefs or not over the long term. Second, I show that information is integrated to a greater degree when it presents a shift in beliefs towards a "good" (i.e. better than first thought) compared to a "bad" (i.e. worse than first thought) direction. This is consistent with past findings (Garrett et al., 2014, 2014; Garrett & Daw, 2020; Kappes et al., 2018; Korn et al., 2012b, 2016; Kube & Rozenkrantz, 2021; Kuzmanovic et al., 2018, 2019c; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Oganian et al., 2019b; Sharot et al., 2011; Sharot & Garrett, 2016), suggesting that learning is often biased in a positive direction. What is new in the findings I present here is that this bias is equally prevalent in response to false information (in fact, I find evidence against any modulation of the bias by accuracy). This suggests that whilst debunking information after it has been encountered can act to mitigate its influence, such measures are likely to be less successful in cases where false information presents news that is better than expected.

Using computational modelling of participants' responses suggested that both of these patterns of belief updating arose out of the differential use of error signals, generated when new information is encountered and deviates from prior expectations. In general (Garrett & Daw, 2020; Sharot et al., 2011; Sutton & Barto, 2018), the more surprising a piece of information is, the more beliefs can be expected to shift up or down.

Here I show that this relationship between degree of surprise (i.e. the size of the error signal) and subsequent change in beliefs (parametrised in computational models as learning rates) is stronger when information is revealed to be true compared to when it is revealed to be false and when information is better compared to worse than expected. I show this by comparing models which parse information according to its accuracy only, according to valence only and according to *both* accuracy and valence with this latter model providing the best fit to the data both quantitively - using Bayesian model comparison - and qualitatively - by comparing how well belief updating generated by each model (using simulations) was able to capture the updating pattern I observe in the human data (Palminteri, Wyart, et al., 2017).

In the set of computational models I tested, updates were implemented as a single step, which (effectively) collapses over two distinct stages of a trial in our task. Specifically: (1) when information is first received and an estimation error generated; (2) when cues about the reliability of the information are received. One possibility is that the process of belief change occurs in a single step in this way. Under this scenario, information presented (at (1)) would need to be maintained in working memory before the true/false cue is shown (at (2)), which could act much akin to a go/no-go signal (Logan et al., 1984) in determining whether to then integrate the information (i.e. implement an update), possibly via recruiting the brain circuits previously suggested to be involved in belief updating in this task to differing degrees dependent both on whether the information was true or false and whether the information was better or worse than expected. These brain regions include the left inferior frontal gyrus and medial frontal cortex (Garrett et al., 2014; Sharot et al., 2011) for good news and the right inferior frontal gyrus (Sharot et al., 2011) and right inferior parietal lobule (Garrett et al., 2014) for bad news. But an alternative possibility is that separate updates to beliefs occur at each of these two stages; an initial update to beliefs at (1) before the accuracy of the information is known. And then a revision to this updated belief at (2), depending on the identity of the true/false cue. At this second stage, learning from the first stage could either be undone (in the presence of a false cue) and/or further boosted (in the presence of a true cue). Building a complete temporal picture of the updating process as it unfolds and understanding which stages of this process could cause biases to emerge remains to be tested both by incorporating other neuroscience methods such as functional Magnetic

Resonance Imaging (Glover, 2011) alongside experimental paradigms which would enable these different potential underlying processes to be dissociated from one another.

I adapted a widely used belief updating task (Garrett et al., 2014, 2014; Garrett & Daw, 2020; Garrett & Sharot, 2023; Kappes et al., 2018; Korn et al., 2012b, 2016; Kube & Rozenkrantz, 2021; Kuzmanovic et al., 2018, 2019c; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Oganian et al., 2019b; Sharot et al., 2011; Sharot & Garrett, 2016) shown to test for the presence of biases in belief updating by adding explicit cues on each trial which signalled to participants whether the information they had just observed had been true or false. This allowed me to test whether the bias exists and is similar in strength when true and false information is received. A natural question that arises from these findings is how might biases in belief updating manifest in the absence of any such cues or in the presence of a third "unknown accuracy" cue? The former is how original versions of the belief updating task have been run in the past. Interestingly, the optimistic pattern of belief updating observed when using this original version (without any information accuracy cues) qualitatively (Sharot et al., 2011) resembles the pattern of belief updating I observe in the true condition here. This might lead one to hypothesise that participants might treat information explicitly signalled to be true similarly to cases where information is presented without any cues provided about accuracy. However, caution is warranted. Even though cues about information accuracy were not provided, the default position of participants undertaking original instantiations of the belief updating task may be to assume that the statistics presented were true (rather than unknown). Indeed, the statistics presented in the original design were factually correct and task instructions primed participants to believe they were accurate. Hence, it remains to be tested how belief updating from true and false information each differs from cases where information accuracy is withheld.

Nonetheless, the results here show that sensitivity to new information is able to adapt in response to signals about how accurate the information is or is not. This extends past findings showing that sensitivity to new information is often not fixed but can flexibly adjust to a range of factors including: volatility of the environment (Behrens et al., 2007), surprise (Pearce & Hall, 1980), age (Moutsiana et al., 2013), psychiatric symptoms (Garrett et al., 2014; Ossola et al., 2020), reward uncertainty (Chen et al., 2022; Nassar et al., 2012), working memory (Collins & Frank, 2012), social context (Diaconescu et al., 2014b), arousal levels (Eldar et al., 2016; Garrett et al., 2018) and confidence (Desender et al., 2019; Meyniel et al., 2015; Rollwage et al., 2020; Yeung & Summerfield, 2012). Potentially some or all of these factors may also exert roles in governing the degree to which information known to be false can impact beliefs making certain groups of individuals more susceptible than others to the effects of malicious misinformation attempts which in turn might warrant greater measures being put in place to protect those likely to be more vulnerable.

My findings suggest that warnings and labels that call out cases of potential false information could be a means to help prevent false beliefs being generated and sustained in the face of accurate information to the contrary. However, explicit labels and the like are unlikely to be a panacea. Not least because - as I show here - individuals are likely be more suspectable to false information (even when this is made explicit) in cases where it provides a better-than-expected view of the future. Indeed, recent complementary findings (Vidal-Perez et al., 2025) suggest that a related ('positivity') bias also exists whereby individuals learn to a greater degree from false information that confirms past choices and decisions. An important concern is that alongside these biases in how individuals learn when they receive false information, individuals also have increasing agency over where they choose to source their information from in the first place, particularly in digital environments. This can result in selective exposure to certain types of information and skewed informational environments (Flaxman et al., 2016). This skew can be further compounded if individuals are also selective about what information they choose to share with peers in their network, choosing to predominantly share false information that perpetuates a specific view of the world that one finds desirable (Pennycook & Rand, 2022). There is then a potential dual challenge to be met to counter disinformation – how biased the information is that I receive and the role that biases play in how I then choose to learn from that information. Other promising avenues exist to meet this challenge such as using Large Language Models to deploy reasoning to move individuals away from deep seated false beliefs (Costello et al., 2024), providing rewards (such as 'likes') in return for sharing accurate information (Vellani et al., 2024) and 'prebunking' (Van Der Linden, 2024). Awareness of the powerful role that biases can play and how these can arise from core learning principles are important factors to consider in evaluating the effectiveness of these and others as the scientific community looks to develop and test different ways to counteract inaccurate information from successfully perpetuating and sustaining false beliefs.

Chapter 5: When we value false information: the interaction between information accuracy and confirmation in the ventromedial prefrontal cortex

5.1 Introduction

Imagine reading online that your favoured political party was likely to win the next election, only to later find out that this information was in fact false. How would finding this out make you feel? Grateful for the chance to be able to apply a retrospective correction to the fake news? Or disappointed that this information (which was in line with what you had hoped) was no longer valid? If the latter, would this be accompanied by a lingering temptation to ignore or downweight the evidence that the information you had received was suspect?

Whilst it might seem like the most rational thing to place a high value on the opportunity to correct false beliefs, previous findings (Garrett & Daw, 2020; Garrett & Sharot, 2017b; Korn et al., 2012b; Kuzmanovic & Rigoux, 2017; Lefebvre et al., 2017; Palminteri, 2023, 2025b; Palminteri, Lefebvre, et al., 2017; Sharot et al., 2011) suggest that even when individuals get information from legitimate bona fide sources, they attend to this selectively and in such a way that enables them to maintain beliefs that are biased in a desirable direction given their idiosyncratic motivations, goals, desires and past decisions (Hart et al., 2009; Kunda, 1990; Tappin et al., 2017). And when it comes to false information, behavioural evidence alongside computational modelling (Chapters 3 and 4) has revealed that when individuals find out that information they encountered was false, they do not correct their beliefs in a symmetric manner. Rather, they continue to learn from and use false information in cases where that false information had been confirmatory – confirming that their past choices and decisions were correct. But at the same time, they are adept at ignoring false information in cases where the information had been disconfirmatory and called into question past decisions.

A potential account of these findings is that people are motivated, value, and pay attention to cues warning them that the information they've encountered is suspicious, faulty and should not be heeded. But only in certain cases. These cases are when the information to begin with is disconfirmatory; hence, fake news warnings provide a welcome opportunity to ignore the information that had served to challenge past decisions. But in cases where the information encountered has been confirmatory, revelations that this information was inaccurate act as aversive, which in turn prevents any necessary motivation to try and correct them. Adapting the degree to which we value and pay attention to signals about whether information can or cannot be trusted, potentially provides a mechanism by which information can be selectively used to help beliefs align with our desires (i.e. what we want to be the case) rather than the reality.

To investigate this, I combined brain imaging with a reinforcement learning paradigm (Chapter 3) in which individuals made repeated choices, received feedback which could either confirm or disconfirm whether their choices were correct and were then told whether this feedback was true or false. My computational modelling findings previously showed that individuals integrated information revealed to be true and false to a greater degree when it was confirmatory compared to disconfirmatory. If this relates to how agents value finding out the veracity of the information, this makes an interesting and neural prediction for populations of neurons that encode subjective value. Specifically, finding out that information is true ought to act as rewarding if the information is confirmatory (as this serves to validate both the information and the decision) relative to when the information is disconfirmatory (as this validates the information but challenges the decision in the process). Conversely, finding out the information is false ought to act as rewarding if the information is disconfirmatory (as this serves to challenge the information calling the decision into disrepute) relative to when the information is confirmatory (as this challenges the information and suggests the decision may not have been correct after all).

To see whether this revelation of whether the feedback was true or false selectively activated voxels associated with subjective value, I focused my analysis on the ventromedial prefrontal cortex (vmPFC), a region known to correlate with value across a range of domains including primary rewards like food, secondary rewards like money, abstract rewards like social approval (Bartra et al., 2013; Clithero & Rangel, 2014;

Levy & Glimcher, 2012), aesthetic judgments of art (Kawabata & Zeki, 2004), and viewing attractive faces (O’Doherty et al., 2003).

5.2 Methods

5.2.1 Participants

Forty participants (mainly students at the University of East Anglia) took part in the study for credits and up to a £5 bonus based on performance. Data from eight participants were excluded from the analysis: three for failing to pass behavioural attention checks (less than 50% accuracy) and five due to excessive head motion (defined as mean framewise displacement higher than 0.3 mm or absolute mean displacement higher than 2mm). The final sample for analysis consisted of 32 participants ([25 female, 7 male]; mean age = 20.5 years). All participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological or psychiatric conditions. The study was approved by the School of Psychology Ethics Committee at the University of East Anglia. All participants provided written informed consent prior to participation.

5.2.2 The Task

Participants performed the probabilistic instrumental learning task (Chapter 3) with several changes (**Figure 5.1**). First, the duration of the first and second fixation crosses was randomized to last between one to three seconds and four to five seconds, respectively. Second, choices were no longer self-paced but restricted to a 3-second time limit, with a warning message appearing if participants responded too slowly. This helped to ensure all sessions had the exact same length. Third, the trial structure was set so that each of the four blocks had 12 trials for each of the four unique pairs of stimuli, bringing the total to 196 trials for the entire task. Finally, one attention check trial was added to each block. The task was programmed and presented using PsychoPy 2.2 (Peirce et al., 2019).

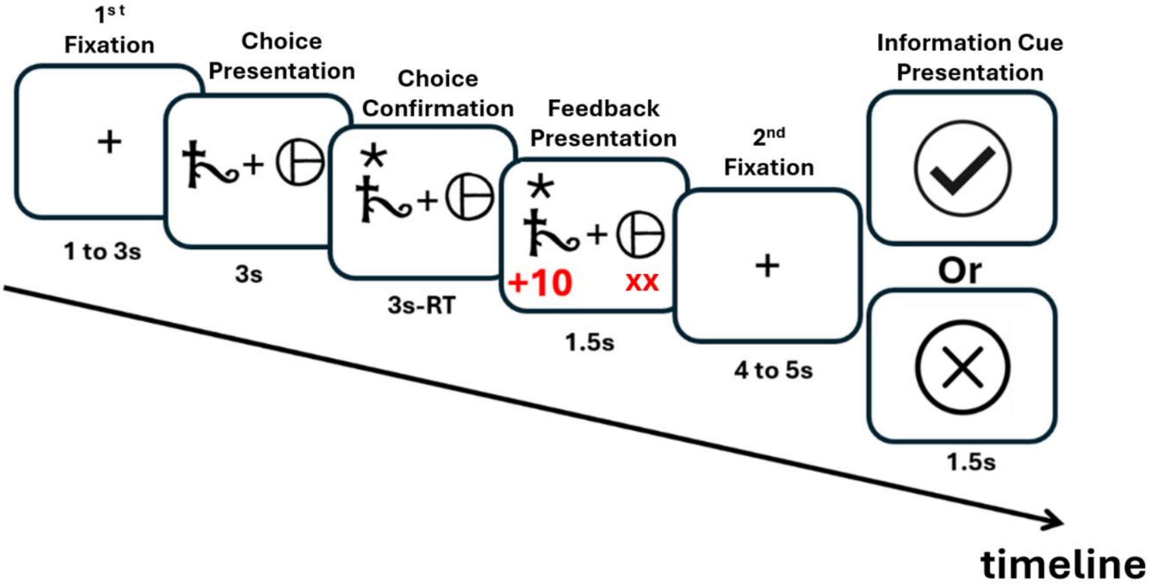


Figure 5.1: Timeline of the task. Participants are initially shown two options and select one within a 3-second limit. Upon selection, a star appears above the chosen option as confirmation. Following this, the outcome of one of the two options (either chosen or unchosen) is displayed, with "xx" marking the outcome not shown. Whether the outcome was true (tick) or false (cross) is then indicated.

Before the main experiment, participants completed a practice session (one outside the scanner on a computer and one inside the scanner). This involved 16 trials with two unique stimulus pairs outside the scanner, followed by 8 trials with two different unique pairs inside the scanner.

5.2.3 Behavioural Analysis

To investigate how participants used both feedback and accuracy of the feedback to guide subsequent decisions, I analysed choice repetition on a trial-by-trial basis. The dependent variable was choice repetition, coded as a binary outcome (1 = the participant repeated the choice from the previous trial, -1 = they switched their choice). I fitted a linear mixed-effects model using the glmer function from the lme4 package in R (Bates et al., 2015). The model predicted the likelihood of choice repetition (t) from two fixed effects: the previous_feedback (t-1) (confirmatory coded as +1 vs disconfirmatory coded as -1) and the previous _accuracy (t-1) (true coded as +1 and false coded as -1), including the interaction between them. To account for variability across participants, the model included a random intercept for each participant as well as random slopes for both main effects and their interaction.

5.2.4 Computational Model

I tested the same four models described in Chapter 3, fitted and validated using approaches outlined in Chapter 2. For brevity, below I will only describe the winning model M4 (see Appendix 5.4 for the rest of the models).

For each subject and each block, option values $Q_i(t)$ were stored separately for each pair of options, with $i \in \{1,2\}$ denoting the option index and t the trial number. At the start of the experiment, all Q-values were initialised to zero. On trial t, the participant chose one option (i_{chosen}), rendering the other option unchosen ($i_{unchosen}$). Depending on whether the feedback was presented for the chosen or unchosen option, the corresponding Q-value was updated.

The prediction error (δ) was defined as:

$$\delta_i(t) = R_i(t) – Q_i(t)$$

where $R_i(t)$ is the observed outcome.

The update was given as:

$$Q_i(t+1) = Q_i(t) + \alpha * \delta_i(t)$$

with the learning rate α determined by both outcome accuracy (true vs false) and feedback type (confirmatory vs disconfirmatory). This yielded four distinct learning rates:

$$Q_i(t+1) = Q_i(t) + \alpha_{Conf, false} * \delta_i(t) \quad \text{if accuracy = False and feedback = Conf}$$

$$Q_i(t+1) = Q_i(t) + \alpha_{Disconf, false} * \delta_i(t) \quad \text{if accuracy = False and feedback = Disconf}$$

$$Q_i(t+1) = Q_i(t) + \alpha_{Conf, true} * \delta_i(t) \quad \text{if accuracy = True and feedback = Conf}$$

$$Q_i(t+1) = Q_i(t) + \alpha_{Disconf, true} * \delta_i(t) \quad \text{if accuracy = True and feedback = Disconf}$$

“Confirmatory” feedback was defined as an outcome that confirmed the participant made the right choice ($\delta > 0$ for the chosen option or $\delta < 0$ PE for the unchosen option), while “disconfirmatory” feedback was defined as an outcome that disconfirmed the participants decision ($\delta < 0$ PE for the chosen option or $\delta > 0$ for the unchosen option).

Choices were modelled using a SoftMax decision rule applied to the Q-values of the chosen and unchosen options on the current trial:

$$P(\text{choose } i) = \frac{\exp\big(\beta Q_i(t)\big)}{\exp\big(\beta Q_1(t)\big) + \exp\big(\beta Q_2(t)\big)}$$

where P(choose i) is the probability of selecting option i at trial t, and β is the inverse temperature parameter that controls the degree of stochasticity in choice behaviour. Larger values of β yield more deterministic choices, while smaller values reflect more exploratory behaviour.

The contribution of each trial to the likelihood was given by the log probability of the observed choice. For a choice between the chosen (i_c) and unchosen (i_u) options:

$$\mathrm{l}_t = \log\left(\frac{1}{1 + \exp\big(-\beta\left[Q_{i_c}(t) - Q_{i_u}(t)\right]\big)}\right)$$

The log-likelihood for a subject was then:

$$\mathcal{L} = \sum\nolimits_t \mathrm{l}_t$$

and the model minimised the negative log-likelihood, –L, during estimation.

In summary, M4 is a five-parameter ($\alpha_{\text{Conf, true}}$, $\alpha_{\text{Disconf, true,}}$ $\alpha_{\text{Conf, false}}$, $\alpha_{\text{Disconf, false}}$, β) model that distinguishes learning rates for confirmatory vs disconfirmatory feedback and true vs false information.

5.2.5 fMRI Image Acquisition

Scanning was performed at the University of East Anglia scanning centre UWWBIC, using a 3T Siemens MAGNETOM Prisma MRI scanner equipped with a Siemens head coil. The imaging session began with a high-resolution T1-weighted structural scan with MPRAGE sequence. This was followed by four functional runs, each lasting 12 minutes, amounting to 374 scans, the first 6 of which were discarded. Functional images were acquired using a T2-weighted echo-planar imaging (EPI) sequence with multi-band acceleration. The following parameters were used: Repetition Time (TR) = 2000 ms; Echo Time (TE) = 30 ms; Flip Angle = 78°; 50 interleaved axial slices; Voxel size = 3.0 × 3.0 × 3.0 mm; Slice thickness = 3.0 mm (no gap); Field of View (FOV) = 192 × 192 mm; Matrix size = 64 × 64; multi-band acceleration factor = 2. The session concluded with a 1-minute gradient echo (GRE) field map with the same resolution and slice locations as the functional images to correct for geometric distortions caused by magnetic field inhomogeneities.

5.2.6 fMRI Data Preprocessing

Statistical Parametric Mapping (SPM12, Wellcome Trust Centre for Neuroimaging) was used for image processing and analysis. Raw DICOM images were first converted to NIfTI format. After discarding the first 6 dummy volumes, images were realigned to the 7th volume. Movement plots were studied to ensure scan-to-scan translations greater than one-half of a voxel (1.5 mm) or rotations greater than 1° did not cause artifacts in the corresponding scans. Structural images were reregistered to mean EPI images and segmented into grey and white matter. These segmentation parameters were then used to normalise and bias-correct the functional images to a standard EPI template based on the Montreal Neurological Institute (MNI) reference brain using a nonlinear (7th-degree B-spline) interpolation. Normalised images were spatially smoothed with an 8 mm Full-Width at Half-Maximum (FWHM) Gaussian kernel. A high-pass filter of 1/128-Hz was applied to the time-series data to remove low-frequency artifacts.

5.2.7 fMRI General Linear Models

GLM1 (main analysis). For each participant, a design matrix was created with event onsets time-locked to the temporal positions of Choice Presentation, Feedback Presentation and Information Cue Presentation. Events were modelled as delta functions and convolved with a canonical hemodynamic response function to create regressors of interest. The onset regressor for Information Cue Presentation was subdivided into 4 conditions: Confirmatory True (where confirmatory feedback turned out true), Disconfirmatory True (where disconfirmatory feedback turned out true), Confirmatory False (where confirmatory feedback turned out false), and Disconfirmatory False (where disconfirmatory feedback turned out false). This resulted in six regressors for each session. Six motion correction regressors estimated from the realignment procedure were entered as covariates of no interest.

To identify regions that tracked the subjective value of the two options participants chose between on each trial, I entered the absolute difference in the learned Q-values

($|\Delta Q|$) of the two options presented (with trial by trial Q values of each option extracted from the winning computational model for each participant) as parametric regressors, modulating the events in which choice pairs were presented (Choice Presentation).

To identify correlates of prediction errors at the time that information accuracy was revealed, unsigned prediction errors ($|\delta|$) at the time the information accuracy cue was presented were entered as parametric modulators at the Information Cue Presentation timepoint, parsing out separately for Confirmatory True, Disconfirmatory True, Confirmatory False, and Disconfirmatory False. Note that I used *unsigned* PE here as a parametric modulator, but confirmatory and disconfirmatory PEs both have a mixture of positive and negative PEs. For instance, confirmatory feedback is generated both when a positive PE occurs for factual feedback (i.e. feedback for the option chosen) and when a negative PE occurs for counterfactual feedback (i.e. feedback for the unchosen option), and vice versa for disconfirmatory. By using unsigned PEs, I can identify brain regions in which BOLD responses scale with the extent to which the error term suggests participants choose correctly (for confirmatory trials) or incorrectly (for disconfirmatory trials), independently of whether the PE is positive or negative.

GLM2. A second GLM was used to be able to separately extract BOLD response at the time of the information cue, by separating this into eight bins according to whether feedback was for the option chosen or unchosen, whether the PE had been positive or negative and whether the information cue revealed the feedback to have been true or false. This was to be able to explore and plot out effects present in GLM1 at the time of information cue presentation at a more granular level and see whether the direction of the BOLD response changed according to all 3 factors.

For each participant, a design matrix was created with event onsets time-locked to the temporal positions of Choice Presentation, Feedback Presentation and Information Cue Presentation. As for GLM1, events were modelled as delta functions and convolved with a canonical hemodynamic response function to create regressors of interest. The onset regressor for Information Cue Presentation was subdivided into eight conditions:

(1) Positive PE, Chosen, True (trials where the prediction error was positive, feedback was provided for the option chosen and the information cue revealed the feedback to have been true)

(2) Negative PE, Chosen, True (trials where the prediction error was negative, feedback was provided for the option chosen, and the information cue revealed the feedback to have been true)

(3) Positive PE, Chosen, False (trials where the prediction error was positive, feedback was provided for the option chosen, and the information cue revealed the feedback to have been false)

(4) Negative PE, Chosen, False (trials where the prediction error was negative, feedback was provided for the option chosen, and the information cue revealed the feedback to have been false)

(5) Positive PE, Unchosen, True (trials where the prediction error was positive, feedback was provided for the unchosen option, and the information cue revealed the feedback to have been true)

(6) Negative PE, Unchosen, True (trials where the prediction error was negative, feedback was provided for the unchosen option, and the information cue revealed the feedback to have been true)

(7) Positive PE, Unchosen, False (trials where the prediction error was positive, feedback was provided for the unchosen option, and the information cue revealed the feedback to have been false)

(8) Negative PE, Unchosen, False (trials where the prediction error was negative, feedback was provided for the unchosen, and the information cue revealed the feedback to have been false)

Note that (1), (3), (6) and (8) represent cases of confirmatory feedback which transpire to be true in the cases of (1) and (6) but false for (3) and (8). Whereas (2), (4), (5) and (7) represent cases of disconfirmatory feedback which transpire to be true in the cases of (2) and (5) but false in the cases of (4) and (7).

This resulted in 10 regressors for each session. Six motion correction regressors estimated from the realignment procedure were entered as covariates of no interest. Also, one subject had to be excluded from this analysis for insufficient data in one of the conditions.

5.2.8 Region of Interest (ROI) Definition

The use of a Region of Interest (ROI) approach in neuroimaging can be summed up as a trade-off between statistical power and spatial exploration. By restricting the analysis to a priori hypothesized regions, this approach significantly reduces the severity of the multiple comparisons problem in whole-brain analyses, increasing the sensitivity to detect subtle effects (Poldrack, 2007). However, this sensitivity relies on the validity of the ROI selection strategy. ROIs can be defined in several ways. Anatomical definition uses standardized atlases (e.g., Automated Anatomical Labelling) and assumes that functional boundaries map perfectly onto structural ones. Functional definition involves using an independent localizer task to identify the region within each specific subject (Saxe et al., 2006). While this accounts for individual variability in functional topography, it increases scanning time and relies on the assumption that the localizer task taps into the exact same neural computation as the main experiment.

In the absence of a functional localizer, ROIs can be defined using independent coordinates from the literature. This means a choice between deriving peaks from a meta-analysis or a single representative or relevant study. Meta-analytic coordinates offer robustness by aggregating across widespread data, smoothing out study-specific noise (Yarkoni et al., 2011). However, the resulting consensus regions can be spatially broad or general. On the other hand, coordinates from a single, high-quality study allow for greater specificity to the exact psychological process being investigated, which in my case is subjective value. Therefore, I prioritized specificity and created an independent ROI mask for an a priori brain region known to signal subjective value, the vmPFC (Bartra et al., 2013; Chib et al., 2009; Kable & Glimcher, 2007; Lebreton et al., 2009; Lefebvre et al., 2017; Levy & Glimcher, 2012; Padoa-Schioppa & Assad, 2006; Rangel et al., 2008). This ROI was created using the MarsBaR software (Brett et al., 2002) by defining a 6-mm radius sphere centred on peak coordinates (Montreal Neurological Institute (MNI) space coordinates (x, y, z) = (12, 56, 4)) from a previous independent study showing robust vmPFC activation in response to subjective value (De Martino et al., 2013). I extracted the betas using the same software. Given that the coordinates are derived from an independent dataset, this approach avoids the circularity or double-dipping error (Kriegeskorte et al., 2009). This occurs when the same dataset is used to both select the region of interest (e.g., finding the peak voxel of activation in a specific contrast) and to test the hypothesis within that region (e.g., extracting Beta values to determine if the effect is significant). Because fMRI data contains noise, selecting the "best" voxels inherently selects those with noise that aligns with the hypothesis. Therefore, running statistical tests on these pre-selected voxels inflates effect sizes and significance levels, rendering the statistics invalid for inference. The use of a fixed sphere (6mm) is to capture the core of the functional region while accommodating minor inter-subject variability in functional anatomy. However, a limitation of this approach is that it assumes spatial consistency across populations. If the functional region in the current sample drifts slightly from the published coordinates, the fixed sphere may capture less signal than a subject-specific functional localizer.

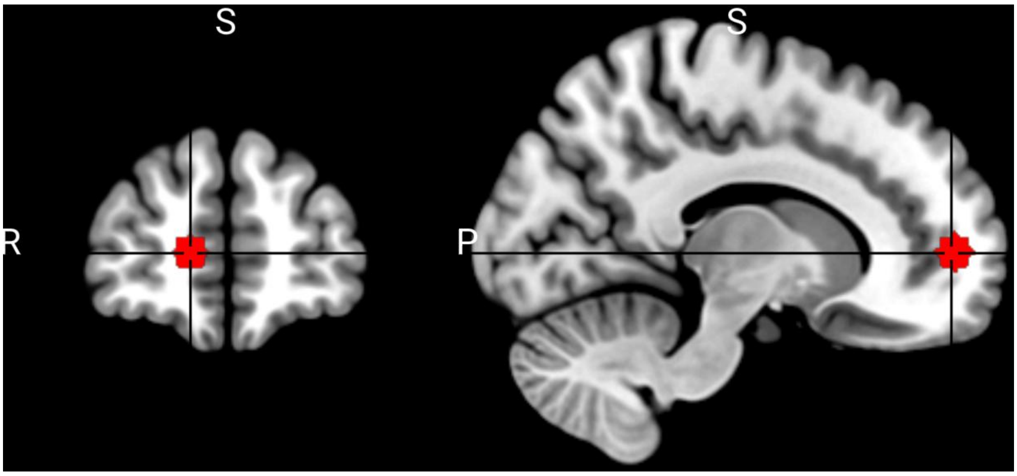


Figure 5.2: The vmPFC ROI mask. This mask was created by a 6-mm radius sphere around peak coordinates of an independent study (Montreal Neurological Institute (MNI) space coordinates (x, y, z) = (12, 56, 4); highlighted in red).

5.2.9 Main Analysis

Choice Presentation. At the time of choice presentation, I investigated if previous findings of vmPFC tracking the difference in value when choosing between 2 options (Boorman et al., 2009; Gläscher et al., 2009; Hare et al., 2008; Hunt et al., 2012) replicated in this study. To do this, I extracted (from GLM1) the individual participant beta coefficients for the parametric modulator (modulating the choice time point by the trial-by-trial absolute difference in Q-values between the two options under consideration) from the vmPFC ROI and tested for significance at the group level using a one-sample t-test (vs 0).

Information Cue Presentation. At the time of information presentation (when participants learned whether the feedback they had just seen was true or false), I investigated if BOLD response in the vmPFC varied according to whether both the information was true or false and whether the feedback just seen was confirmatory or disconfirmatory. I did this in three ways. First, I extracted (from GLM1) the beta coefficients capturing the average (i.e. unmodulated) BOLD response (vs baseline) for the four conditions (true confirmatory, true disconfirmatory, false confirmatory and false disconfirmatory) from the vmPFC ROI and entered these into a 2*2 repeated measures ANOVA to test for main effects and interaction. To better understand the pattern of BOLD response in vmPFC in this analysis, I then used GLM2 to split the pattern out further (separating for chosen/unchosen, positive/negative PE and true/false information cue). Finally, to see if vmPFC activity scales proportionally with prediction error magnitude, such that larger PEs are associated with stronger BOLD responses and smaller PEs with weaker responses, I extracted the beta coefficients (from GLM1) capturing the parametrically modulated (by trial-by-trial unsigned PEs) activity from the vmPFC ROI. These were entered into a 2*2 repeated measures ANOVA.

5.2.10 Whole-Brain Analyses

I also conducted a whole-brain exploratory analysis at choice and information cue presentations. Significance was determined using a cluster-level correction (voxel-wise threshold $p < 0.001$ uncorrected, family-wise error (FWE) $P < 0.05$, cluster size (K) > 5). I used the JuBrain Anatomy Toolbox (a.k.a. SPM Anatomy Toolbox) (Eickhoff et al., 2005) to label the regions corresponding to the coordinates.

5.3 Results

5.3.1 Behavioural and Computational Modelling Results

Participants modulate their learning based on feedback and accuracy. The linear mixed effects model (see **Methods**) revealed an interaction between previous feedback (confirmatory or disconfirmatory) and previous accuracy (true or false) in predicting choice repetition (β=0.107, p=0.002), as detailed in **Table 5.1**. Further, while receiving confirmatory feedback increased the likelihood of repeating a choice (β=0.14, p<0.001), this effect was significantly amplified when the feedback turned out to be true. This suggests that participants valued both the confirmatory feedback on its own but modulated their learning depending on its veracity. Therefore, on a behavioural level, this confirms that participants integrate both the type of the feedback (confirmatory vs. disconfirmatory) and its accuracy (true vs. false) to guide their subsequent decisions. Next, I assessed whether there was any bias in integrating information using computational models.

Predictor	**Estimate**	**std. Error**	**Statistic**	**p**
(Intercept)	0.64	0.09	7.14	**<0.001**
Previous_Feedback	0.14	0.03	3.71	**<0.001**
Previous_Accuracy	-0.06	0.04	-1.42	0.15
Previous_Feedback × Previous_Accuracy	0.107	0.03	3	**0.002**
N $_{\textit{Participant}}$	32			
Observations	5593			

Table 5.1 The behavioural model. Fixed Effect Estimates and accompanying statistics from a linear mixed-effects model predicting choice repetition on each trial from previous feedback, previous accuracy, and their interaction.

Four learning rate model the best fit to the data. Using leave-one-out cross-validation (LOOcv) scores and a Variational Bayesian Approach, I found that Model 4 provided the best fit as it did in the previous chapter. This model achieved the highest frequency, at approximately 97% **(Figure 5.3(a)),** with an exceedance probability of 1.0. This frequency is significantly above the chance level of 25%, indicating that participants learn differently across the feedback and accuracy domains, echoing the behavioural results. See **Appendix 5.5** for additional model comparison results.

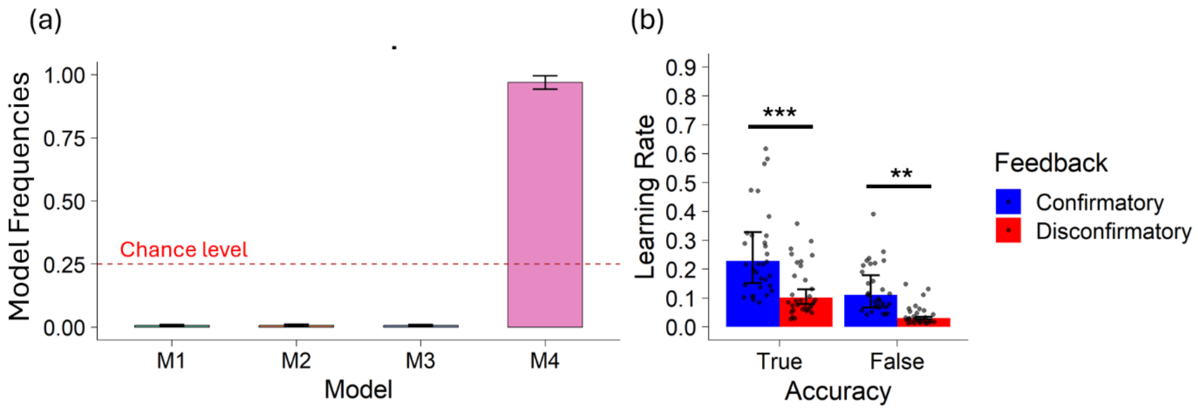


Figure 5.3: Modelling Results. (a) Results from the VBA model comparison showed that Model 4 (M4) had the highest frequency, being chosen for about 97% of participants, with an exceedance probability (XP) of 1. **(b)** Estimates from M4 revealed that learning rates were higher for confirmatory than disconfirmatory feedback for both true information (t(31) = 4.30, p < 0.001; hierarchical t-test comparing α_{true_conf} and $\alpha_{true\_disconf}$) and false information (t(46) = 3.17, p < 0.01; hierarchical t-test comparing α_{false_conf} and $\alpha_{false\_disconf}$), consistent with confirmation bias. ***p < 0.001, *p < 0.01, hierarchical t-test.

Higher learning rate for confirmatory vs disconfirmatory feedback for both true and false information. Using the winning model (M4), I examined the pattern of learning rates **(Figure 5.3(b)).** Replicating the findings from the previous chapter, participants showed confirmation bias for both false (t(31) = 4.30, p < 0.001) and true information (t(31) = 3.17, p < 0.01), learning more from confirmatory than from disconfirmatory feedback. Moreover, the magnitude of this bias did not significantly differ between true and false information (t(31) = 0.80, p = 0.42).

5.3.2 fMRI Results

Subjective value at the time of Choice. I first looked to validate the decision to use activity in the vmPFC as an indirect measure of subjective value at the time participants were presented with the information cue. I did this by first examining activity at the time of choice and looking to see if previous findings (Bartra et al., 2013; Boorman et al., 2009; Hare et al., 2008; Padoa-Schioppa & Assad, 2006) showing that vmPFC activity correlated with the difference in subjective value between two options being chosen between replicated here. Indeed, BOLD signal correlated positively with the unsigned Q-value difference in the vmPFC (peak MNI [x,y,z] = 4,62,10; Z = 3.71, cluster-level pFWE = 0.03, **Figure 5.4**) along with the Cingulate Gyrus (peak MNI [x,y,z] x=8,32,-4; Z = 5.09, cluster-level pFWE = 0.000), Cerebellum (peak MNI [x,y,z] = -36,-80,-40; Z = 4.62, cluster-level pFWE = 0.01), and Lateral Occipital Cortex (peak MNI [x,y,z] = -54,-66,34; Z = 4.54, cluster-level pFWE = 0.02) (see **Appendix 5.1** for complete statistics from the whole-brain analysis). This effect was also significant in the vmPFC ROI mask (see **Methods**) constructed using reported voxels from an independent study (t(31) = 2.32, p = 0.02, one-sample t-test against 0).

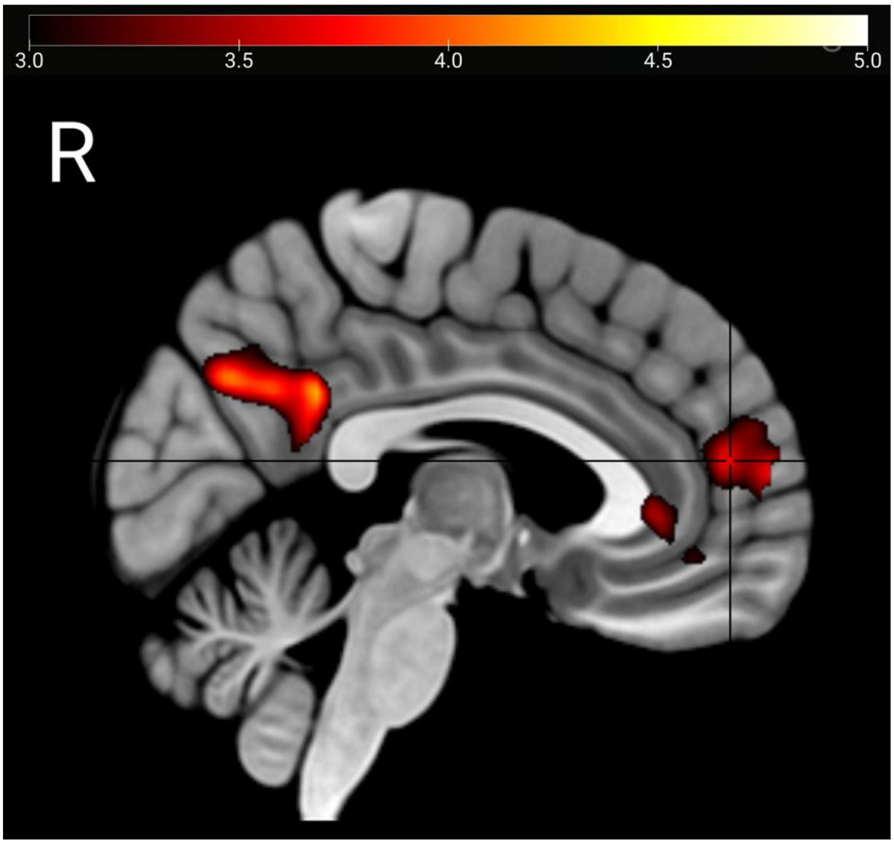


Figure 5.4: The vmPFC Activity at the Time of Choice. The vmPFC activity (peak MNI [x,y,z]: 4, 62, 10) tracks the subjective value difference between options (P < 0.05 FWE corrected at the cluster level) at the time of choice. The statistical map is displayed at a threshold of p < 0.001 uncorrected, overlaid on a standard MNI template. The colour bar indicates the Z-statistic.

Subjective value at the time of Information Cue. Next, I turned to the main question - investigating vmPFC activity at the time participants received the information cue, revealing to them whether the feedback they had just seen had been true or false. The hypothesis was that when this cue reveals information to have been true, finding this out would have higher value (indexed here as higher vmPFC BOLD response) when the feedback had been confirmatory compared to when it had been disconformity. But when the cue reveals the information to have been false, the opposite should transpire whereby this has lower value when the feedback had been confirmatory compared to when it had been disconformity.

Examining changes to the average BOLD response in vmPFC (i.e. the unmodulated effect) by extracting betas from the vmPFC ROI and entering them into a 2*2, repeated measures ANOVA with Information Accuracy (True/False) and Feedback (Confirmatory/Disonfirmatory) revealed a significant interaction between accuracy and feedback (F(1,31) = 7.47, p = 0.01, **Figure 5.5**). There was no effect of accuracy (F(1,31) = 1.99, p = 0.16) or feedback (F(1,31) = 0.0003, p = 0.98). Post hoc tests revealed that the Accuracy*Feedback interaction was the result of greater vmPFC activity when feedback had been confirmatory (versus disconfirmatory) when the information cue revealed was true (t(31) = 2.00, p = 0.05; two-tailed paired t-test between Confirmatory-True and Disconfirmatory-True). But when the information cue was revealed to be false, vmPFC activity was in the opposite direction; lower when the feedback had been confirmatory compared to when it had been disconfirmatory (t(31) = -2.28, p = 0.03; two-tailed paired t-test between Confirmatory-False and Disconfirmatory-False). The interaction was further characterised by greater vmPFC activity in response to confirmatory feedback when the information cue was true compared to when it was false (t(31) = 2.55, p = 0.01; two-tailed paired t-test between Confirmatory-True and Confirmatory-False) with no significant difference between the cues in response to disconfirmatory feedback (t(31) = -0.65, p = 0.51; two-tailed paired t-test between Disconfirmatory-True and Disconfirmatory-False). The results suggest that finding out information is true is rewarding when it validates existing beliefs relative to when it calls them into question. Conversely, finding out information is false is valuable when this negates information that had challenged prior beliefs relative to cases where this negates information that had validated them.

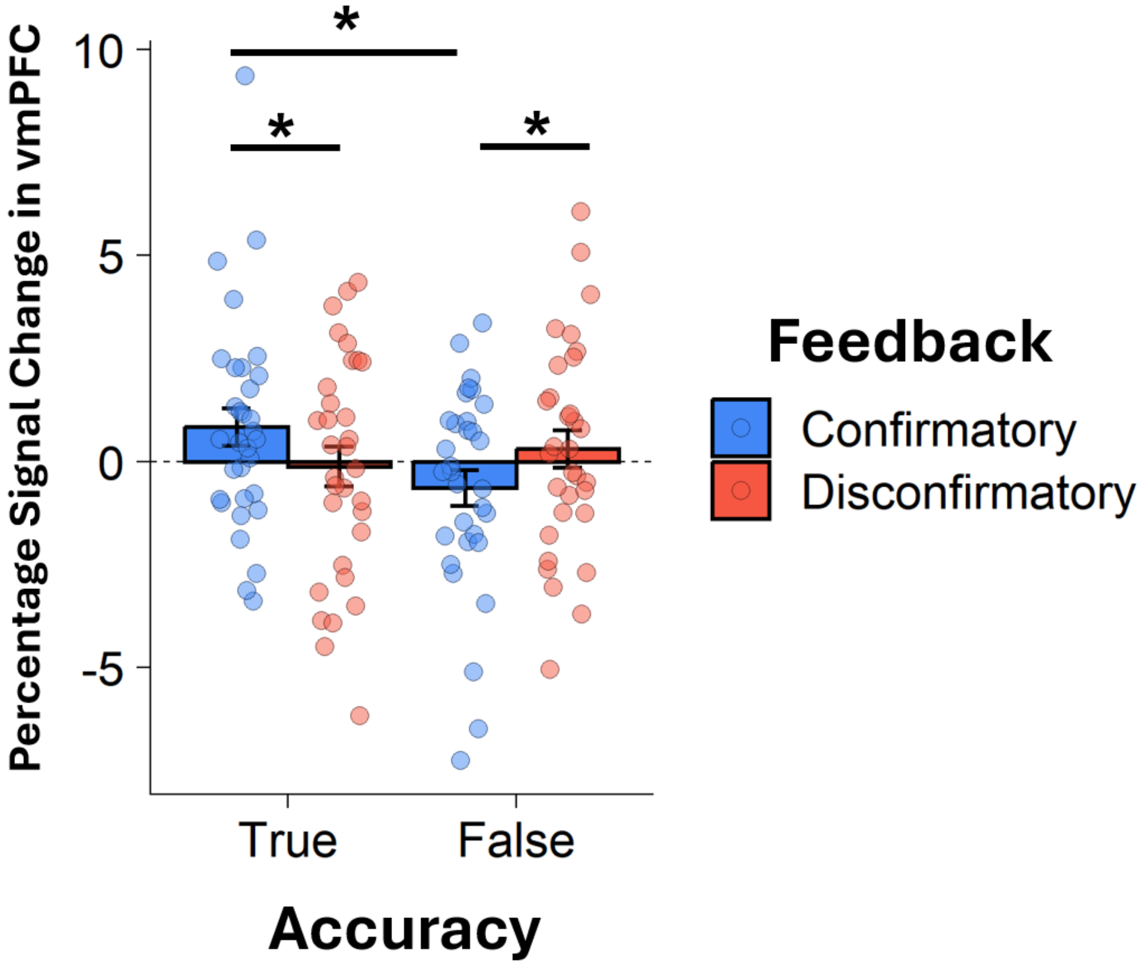


Figure 5.5: The Accuracy * Feedback Interaction Effect. Interaction between feedback and accuracy in the vmPFC ROI (F(1,31) = 7.47, p = 0.01). When the information cue is revealed to be true, vmPFC activity is higher if the feedback had been confirmatory versus disconfirmatory (t(31) = 2.00, p = 0.05; two-tailed paired t-test between Confirmatory-True and Disconfirmatory-True). But when the information cue is revealed to be false, vmPFC activity is in the opposite direction, being lower when the feedback had been confirmatory compared to when it had been disconfirmatory (t(31) = -2.28, p = 0.03; two-tailed paired t-test between Confirmatory-False and Disconfirmatory-False). *p ≤ 0.05.

An exploratory whole-brain analysis also revealed a significant interaction between Accuracy and Feedback in a cluster in the left Medial Temporal Lobe (MTL) that included the Left Hippocampus (peak MNI [x,y,z] = -36,-38,-2; Z = 4.51, cluster-level pFWE = 0.04, k = 361), the Dorsolateral Prefrontal Cortex (peak MNI [x,y,z]= 32,30,44; Z = 4.35, cluster-level pFWE = 0.01) and the Left Cerebellum (peak MNI [x,y,z] =-32,-50,-44; Z = 3.95, cluster-level pFWE = 0.04). See Appendix 5.1 for whole brain statistics including main effects of Accuracy and Feedback.

Given that confirmatory information can occur under different scenarios (positive prediction error for factual outcome, negative prediction error for counterfactual outcome) as can disconfirmatory (negative prediction error for factual outcome, positive for counterfactual outcome) the results we observed in the vmPFC (Figure 5.5) predicts a specific pattern of BOLD response depending on both on the sign of the prediction error, whether outcome was given to the option chosen or unchosen; and, crucially, whether accuracy was revealed to be true or false. In the case of true information, we would expect this to act as rewarding in cases where the prediction error has been positive for the option chosen or negative for the option unchosen (relative to the opposite cases – negative prediction error for the option chosen and positive prediction error for the option not chosen). But this pattern should be inverted for false information, whereby this acts as rewarding in cases where the prediction error has been negative for the option chosen or positive for the option unchosen (relative to the opposite cases – positive prediction error for the option chosen and negative prediction error for the option not chosen).

To unpack this, I used a second fMRI model (GLM2, see **Methods**) which separated trials into eight conditions, depending on 3 factors: the sign of the PE (positive/negative), the outcome (chosen/unchosen) and information cue (true/false) and extracted the BOLD response in my a priori vmPFC ROI at the time the information cue was presented.

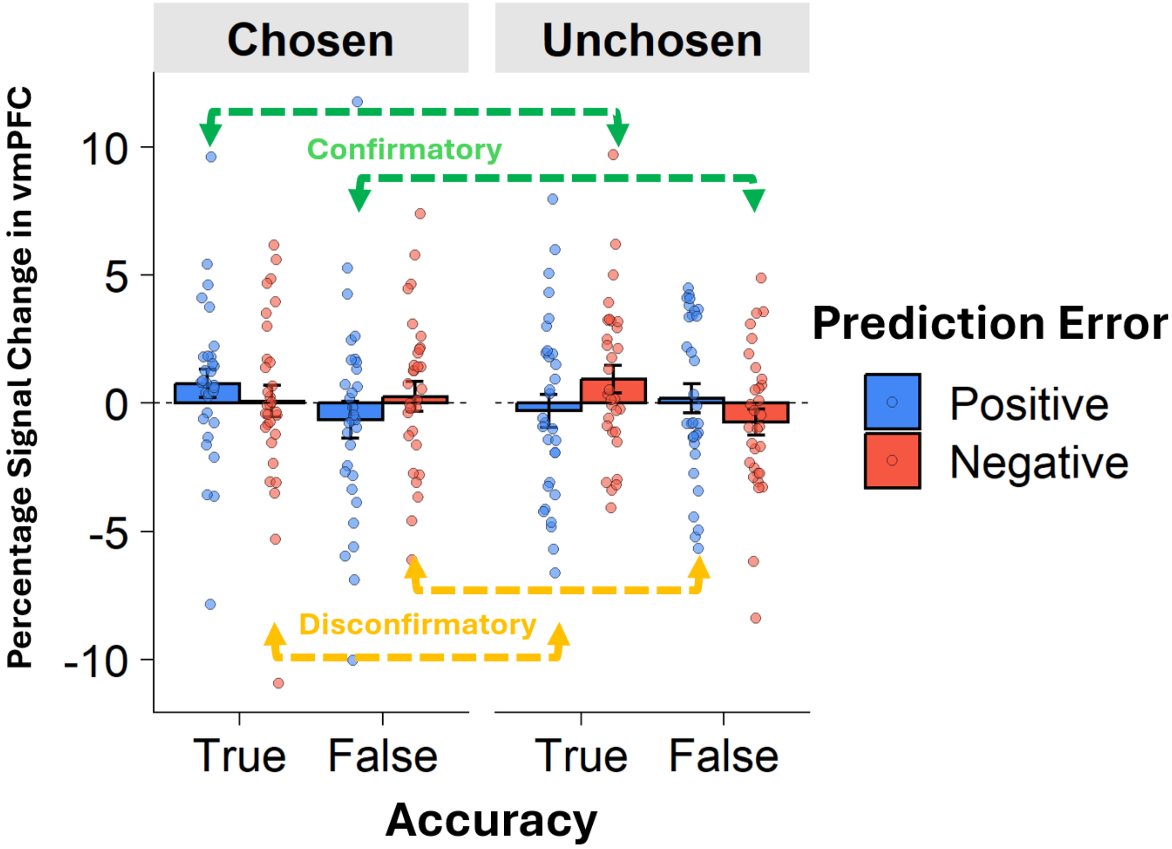


Figure 5.6: The Three-way Interaction Effect. A three-way interaction (F(1,30) = 7.12, p = 0.01) between the sign of prediction error (positive vs negative), outcome (chosen vs unchosen), and accuracy (true vs false). The green, dashed lines show confirmatory feedback (positive – chosen and negative-unchosen), the direction of which is positive for true information but negative for false information. The orange, dashed lines indicate disconfirmatory information (negative-chosen and positive-unchosen).

A 2×2*2 repeated-measures ANOVA on the vmPFC beta estimates revealed a significant three-way interaction (F(1,30) = 7.12 , p = 0.01, **Figure 5.6**). Post hoc tests showed that this was the result of a 2-way interaction between Accuracy and PE_Sign following counterfactual outcomes (i.e. outcome for the unchosen option) (F(1,30) = 5.44, p = 0.02) that was not significant (but showed a qualitative pattern of activation in the opposite direction, **Figure 5.6**) following factual outcomes (F(1,30) = 2, p = 0.16). There was no significant main effect of sign of PE (F(1,30) = 0.26, p = 0.61), Outcome (F(1,30) = 0.05, p = 0.81) or Accuracy (F(1,30) = 2.16, p = 0.15), full stats are reported in **Table 5.2**.

Effect	df	F	p
Intercept	1, 30	0.03	0.861
PE_Sign	1, 30	0.27	0.61
Accuracy	1, 30	2.16	0.152
Outcome	1, 30	0.06	0.811
PE_Sign × Accuracy	1, 30	0.13	0.718
PE_Sign × Outcome	1, 30	0.01	0.929
Accuracy × Outcome	1, 30	0.01	0.957
PE_Sign × Accuracy × Outcome	1, 30	7.12	.012*

Table 5.2 The Three-Way ANOVA. A 2×2*2 repeated-measures ANOVA for the sign of PE (positive or negative), Accuracy (true or false), and Outcome (shown for the chosen or unchosen option) on the beta estimates of the vmPFC ROI. * p < 0.05.

Finally, I examined whether there were parametric effects of unsigned PEs at the time of information cue presentation, extracting the betas from the vmPFC ROI (GLM1) and entering them into a 2*2, repeated measures ANOVA with Information Accuracy (True/False) and Feedback (Confirmatory/Disonfirmatory) as factors. There was no significant effect of accuracy (F(1,31) = 0.03, p = 0.82), Feedback (F(1,31) = 1.62, p = 0.21) or Accuracy*Feedback interaction (F(1,31) = 0.02, p = 0.88). See **Appendix 5.1** for full whole-brain statistics, including main effects of Accuracy and Feedback.

5.4 Discussion

The main analysis revealed a significant interaction between feedback (where a piece of information confirmed or disconfirmed one's choice) and information accuracy (whether the feedback turned out true or false) in the vmPFC – a key region for assigning subjective value (Bartra et al., 2013; Chib et al., 2009; le & Glimcher, 2007; Lebreton et al., 2009; Lefebvre et al., 2017; Levy & Glimcher, 2012; Padoa-Schioppa & Assad, 2006; Rangel et al., 2008).vmPFC activity was significantly higher if confirmatory vs disconfirmatory feedback turned out true. Conversely, when the feedback turned out false, the vmPFC activity showed the opposite pattern; being significantly higher if the feedback had been disconfirmatory compared to confirmatory. Together, this suggests that participants assigned greater value to finding out information was false when it invalidated earlier disconfirmatory evidence compared to when it invalidated earlier confirmatory evidence, but assigned greater value to finding out information was true when it validated earlier confirmatory compared to disconfirmatory evidence.

The neural findings have implications for advancing the computational models of the thesis. Although there is a significant Feedback*Accuracy interaction effect using the average BOLD response, there is no significant parametric effect of prediction errors on this interaction at the time of information cue. These prediction errors, however, were generated at the *feedback timepoint*, which raises the possibility that separate learning signals are generated at the *information cue* timepoint. This points toward an alternative framework that my current models cannot test: a two-step learning process. In such a framework, a first update would occur when the feedback is received. This would be followed by a second update upon the reveal of the accuracy, which could be modelled in several ways. For information revealed to be false, the initial learning could be revised, and I could quantify this by a specific 'undo' parameter (ω). Another model could have two distinct 'undo' parameters: one for confirmatory feedback and another for disconfirmatory feedback. Conversely, for information revealed to be true, the initial update is confirmed. A different model could explore whether this confirmation is passive (i.e., the absence of undoing) or an active 'boosting' of the initial learning, governed by its own 'boost' parameter (γ). Therefore, the neural results of this chapter could provide insight into the modelling of the behavioural data.

My exploratory, whole-brain analysis culminated in two significant regions for the Feedback*Accuracy interaction - right dorsolateral prefrontal cortex (DLPFC) and a cluster that included the left medial temporal lobe (MTL) – and a main effect of accuracy in the right DLPFC. I propose two explanations, albeit speculative, for the involvement of these regions. First, regarding the main effect of accuracy, is the DLPFC's role in belief-updating under uncertainty (Hofmans & van den Bos, 2025; Moutsiana et al., 2015; Schulreich & Schwabe, 2021). In one transcranial direct current stimulation (tDCS) study, Schulreich and Schwabe (2021) enhanced right DLPFC activity in participants performing a continuous belief-updating task. The results showed increased value updating when it was normatively expected from a Bayesian perspective, meaning that they got better at changing their beliefs when the evidence dictated they should. The current finding of a main effect of accuracy in the DLPFC, with higher activity for true versus false information, aligns with this work as it is more rational: participants needed to update their beliefs more when faced with true information. The MTL has also been implicated in learning and belief updating (Moutsiana et al., 2015) as part of a frontal-subcortical circuit in the original belief update task. Specifically, structural connectivity between the left hippocampus (an MTL subregion) and left inferior frontal gyrus (IFG) was associated with a greater tendency to update beliefs in response to good news vs bad news. Individuals with stronger physical connectivity between these regions showed higher updating in response to good vs bad news. In the current task, MTL activity was higher for confirmatory vs disconfirmatory true information, with the opposite pattern for false information. This suggests the MTL may use desirability as a distinguishing factor in belief updating. My second speculation is the involvement of both regions in retrospective confidence judgements when information cue is presented, at which point the brain asks, "Given the feedback and its accuracy, did I make the right decision?”. In other words, it is evaluating the correctness of the initial beliefs formed at the feedback timepoint. A sub-region of MTL, the Left parahippocampal gyrus, has been implicated in retrospective confidence judgement (Martín-Luengo et al., 2021), with higher activity correlating with higher confidence. The MTL in the current task has higher activity when the initial confirmatory feedback turns out true (confirmatory - true) or the contradictory information could be ignored (disconfirmatory - false), both of which could signal one has made the right choice (as opposed to disconfirmatory true and confirmatory false), increasing retrospective confidence. The DLPFC has also been implicated in retrospective confidence judgment (Fleming, 2024; Fleming et al., 2018; Fleming & Dolan, 2012; Martín-Luengo et al., 2021), where in one meta-analysis it was shown to correlate with one’s confidence in the prior decision (Martín-Luengo et al., 2021). Further, in a causal study using transcranial magnetic stimulation (Shekhar & Rahnev, 2018), disrupting the DLPFC caused participants to report *lower confidence* in their decisions. The researchers proposed a model where the DLPFC's job is to "read out" the strength of the sensory evidence. When the DLPFC is disrupted, this readout becomes noisy, signalling to other areas that the evidence was ambiguous, which in turn leads to a lower feeling of confidence. In the current results, the DLPFC has the highest activity for confirmatory true and disconfirmatory false information, both of which convey a feeling of confidence. However, unlike in the MTL, confirmatory true is not significantly higher than disconfirmatory true, casting doubt on this speculation.

In the present study, I inferred the subjective value of information from vmPFC BOLD responses. A promising future direction would be to create a more direct link between brain activity and subjective feeling by collecting explicit behavioural ratings. For instance, participants could be asked to rate "How positive or negative did you feel finding out this information was true/false?" on a continuous scale after the information cue is revealed. The hypothesis would be that these ratings would mirror the neural data I have collected. Another approach would be using a willingness-to-pay (WTP) task, which has been used to probe the neural representations of value (Plassmann et al., 2007), and see how feedback and accuracy influence willingness to pay and how this is represented in the brain.

Together, the results here are the first to show how the brain processes false information depending on whether it had confirmed or disconfirmed one’s beliefs. I focused on a brain region known for valuation, the vmPFC, to assess its BOLD response to finding out that earlier feedback is false. One might expect the brain would dislike learning it had been misled or deceived, but the results revealed that participants do assign a higher value to disconfirmatory false information compared to when confirmatory information proves false. These findings suggest that people might be differentially motivated to scrutinise the accuracy of information depending on whether it aligns with their existing beliefs. When information appears to confirm one's beliefs (e.g., experimental results that support a scientist's predictions), one may be less inclined to verify its validity. Conversely, when information contradicts expectations (e.g., experimental results going against a scientist’s predictions), one may be more motivated to question its veracity.

Chapter 6: General Discussion

6.1 Summary and Limitations

6.1.1 Chapter 3: Confirmation Bias in Response to False Information

6.1.1.1 Summary

In this study, I investigated whether confirmation bias, the tendency to overweight evidence that supports one's existing beliefs and underweight that which disconfirms it, persists when individuals process information they know to be false. I combined behavioural testing with computational modelling across two studies, using a novel learning task where participants made choices and received feedback that was explicitly cued as either true or false. The behavioural results revealed that participants learned even from information explicitly cued as false. When guided by this false feedback, their selection of the misleading option was significantly above the 50% chance level.. However, this learning was modulated, as performance was significantly better when feedback was true. This demonstrates that while participants paid attention to the accuracy cues, they were unable to completely ignore the influence of false information on their decisions.

The computational modelling results suggested that the learning from false information is biased. I tested a suite of models and found that a model with four distinct learning rates - for confirmatory true, disconfirmatory true, confirmatory false, and disconfirmatory false feedback - provided the best fit to the data in both studies. The parameters from this winning model revealed a robust confirmation bias for both true and false information, where the learning rate for belief-confirming feedback was significantly higher than for belief-disconfirming feedback. The bias for false information could not be explained away by gradual perseveration, was robust across Gain and Loss contexts, and the patterns of factual and counterfactual learning rates confirmed it was separate from positivity bias. Further, the strength of the confirmation bias for false information was the same as the bias for true information as demonstrated by the lack of interaction between the learning rates across feedback (confirmatory vs disconfirmatory) and accuracy (true vs false). These findings identify a mechanism that supports the persistence of biased beliefs, even in the face of information that is explicitly discredited.

6.1.1.2 Limitations

A limitation of this study is the lack of a neutral or unknown cue with which I could compare the current results. Although the confirmation bias in the face of false information is robust, it would be bolstered by having this third Unknown condition as in the real world seeing a piece of information whose veracity is unknown is common. Similarly, the True and False labelling of the information is authoritative and in the real world we don't get such clear-cut accuracy statements of information.

6.1.2 Chapter 4: Optimistic Update Bias in Response to False Information

6.1.2.1 Summary

In this study, I investigated the degree to which explicit labels marking information as false – debunking - enable humans to reduce belief updating in response to false information. I adapted a classic belief-updating paradigm, the update bias task (UBT), in which participants were presented with information about the likelihood of experiencing adverse life events, which was then explicitly labelled as either true or false. This setup allowed me to answer two main questions: first, do people reduce their belief updating when information is labelled as false? And second, does the well-documented optimistic update bias - the tendency to learn more from good news than bad news - persist even for information known to be false? The behavioural results revealed a main effect of accuracy, with greater updating for true compared to false information, and a main effect of valence, with greater updating for good news compared to bad news, and no interaction between valence and accuracy. Further, separate paired t-tests between good news and bad news for true and false information showed that optimistic update bias exists for false information as well. As a result, participants were less successful in discounting false information when it was better than expected compared to worse than expected. Finally, the optimistic update bias survived after controlling for potential confounds like subjective ratings and how much one deems themselves different from others, with the latter measured through computational modelling.

Computational models explained belief updating through estimation errors - the discrepancy between people's initial estimates and the information presented. These models propose that the belief updating process is governed by learning rates, which determine how strongly estimation errors influence belief changes. Different models incorporate varying numbers of learning rates, and the winning model had four learning rates for True Good News, True Bad News, False Good News, and False Bad News. The results from this model revealed that learning rates were significantly higher for good news than for bad news for both true and false information.

These findings identify a potential mechanism - optimistic update bias - that explains why false good news has a greater impact on belief change than false bad news, which could be important for understanding vulnerability to misinformation.

6.1.2.2 Limitations

Similar to chapter 3, the main limitation of the study is that the "true" and "false" labels provided to participants were definitive and authoritative. In real-world environments, the veracity of information is often ambiguous, and individuals must infer credibility from uncertain cues. Therefore, the ability to modulate learning observed here might not generalise to more ecologically valid contexts where the truthfulness of information is not explicitly stated.

6.1.3 Chapter 5: When we value false information: the interaction between information accuracy and confirmation in the ventromedial prefrontal cortex

6.1.3.1 Summary

In this study, I used Functional Magnetic Resonance Imaging (fMRI) to investigate how the ventromedial prefrontal cortex (vmPFC) values a piece of confirmatory vs disconfirmatory information that turns out true vs false. The behavioural and computational results replicated my previous findings; participants' choices were influenced by an interaction between feedback and its accuracy, and their learning was best captured by a model with four distinct learning rates. This model confirmed the existence of a confirmation bias- a higher learning rate for confirmatory versus disconfirmatory feedback - for both true and false information.

The neuroimaging results showed a significant interaction in the vmPFC between the accuracy of the information (true/false) and whether the initial feedback confirmed or disconfirmed the person's beliefs. Essentially, the value the brain assigned to information accuracy depended on its relationship with the earlier feedback. When participants learned that the feedback they received was true, their vmPFC activity was significantly higher if that feedback had originally confirmed vs disconfirmed their belief. In contrast, when they learned the feedback was false, vmPFC activity was higher if the feedback had originally disconfirmed vs confirmed their belief. This suggests it's rewarding to learn that confirming vs disconfirming evidence was correct, and it's also rewarding to learn that the evidence that challenged vs confirmed you was wrong. I then unpacked the feedback*accuracy interaction by separating the confirmatory vs disconfirmatory feedback into its subcomponents: Outcome (shown for the Chosen vs Unchosen options) and the sign of PE (Positive vs negative) such that I had eight onset regressors across PE_Sign (Positive vs Negative), Accuracy (True vs False), and Outcome (Chosen vs Unchosen). The results showed an interaction across these three dimensions, indicating that vmPFC is encoding information value based on confirmation rather than the sign of PE.

6.1.3.2 Limitations

The main limitation is the same as the other two chapters: lack of a no-cue condition. Currently, we do not know how the brain responds to receiving no information on the accuracy of confirmatory vs disconfirmatory information. For instance, would we see the same pattern of vmPFC activity as false cue?

6.2 Synthesis

Across several computational studies in two different learning environments – RL and one-shot learning - I showed that not only do people learn from false information but also are biased in this learning whereby they learn more from *desirable* false information than *undesirable* false information. This learning asymmetry, driven by differential prediction error encoding in the RL task and differential estimation error encoding in the UBT, gave rise to confirmation bias in the former – where desirable information confirmed your choices - and optimistic update bias in the latter– where desirable information gave you an optimistic outlook about the future. Across both paradigms, a model featuring four distinct learning rates - separating desirability (Confirmatory Feedback and Good News vs Disconfirmatory Feedback and Bad News) and accuracy (True vs False) of information - provided the best explanation for participants' behaviour. Estimating the winning model revealed that in both paradigms, desirable information was integrated to a higher degree than an undesirable information no matter the veracity. This biased information integration has been shown in a plethora of belief-updating (for a review see Sharot & Garrett, 2016) and reinforcement learning (for a review see Palminteri & Lebreton, 2022) studies for true information, and now I have shown it also exists in the face of false information.

Across all studies, the strength of both biases was not significantly different between true and false information. This means that in two totally different learning environments, participants exhibited similar levels of biased information integration regardless of veracity. One potential explanation for this is that the brain has one control for desirability that is asymmetric, and a separate 'accuracy' control that acts like a global volume knob, reducing the impact of *all* false information without altering the underlying bias. Therefore, at the information presentation timepoint, desirability is integrated in a biased manner – learning rate for the desirable (confirmatory or good news) information higher than learning rate for the undesirable (disconfirmatory or bad news) information – and then, when it turns out false, the initial learning is turned down but the bias persists, hence the lack of interaction for the strength of the bias for true vs false information.

To probe the neural processing across desirability and accuracy dimensions, I conducted an fMRI study using the RL task, the results of which implicated the vmPFC in biased information valuation. Specifically, upon accuracy reveal, the vmPFC valued true and false information differently depending on whether initially, at the feedback timepoint, the information confirmed or disconfirmed participants' choice. if initially desirable, true information was deemed more valuable than false information; if initially undesirable, however, vmPFC shows no difference between true and false information. Similarly, it assigned greater value to undesirable information that turned out false compared to the desirable one that proved false. This could suggest that, for false information, participants are motivated to better ignore the undesirable information that turns out false relative to the desirable one.

Although the biased false information integration and valuation is concerning, participants did learn *less* from false information than true information – main effect of accuracy – across all studies in the thesis. This means that debunking (warning about the veracity of information) could be somewhat effective. This is in line with the consensus in the misinformation literature that debunking is an effective, albeit imperfect, intervention (M.-P. S. Chan et al., 2017; Van Der Linden, 2024). I contribute to this line of research by providing a computational account of debunking, showing under what circumstances it could be less effective, and offering a neural account of how the brain processes false information the moment it is debunked. In what follows I will describe a few open questions that could deepen our understanding of how misinformation is learned and how it could be curbed.

6.3 Key Questions for Future Research

6.3.1 Is pre-bunking more effective than debunking in reducing learning from false information?

Chapters 3 and 4 showed that debunking, where information *turns out* false, is effective in reducing the degree to which people learn from false information no matter if the information is abstract (i.e., two-arm bandit symbols) or realistic like future negative life events. The misinformation literature has offered another approach that warns people *ahead of* a potential falsehood, inoculating them like how a vaccine inoculates one from infections. This is known as prebunking (Christner et al., 2024; Ecker et al., 2022; Van Der Linden, 2024). The prebunking version of my tasks would be if the information cue is presented first - telling them if the upcoming information (e.g., the outcome of the option) is true or false - and then participants are asked to make a choice and observe the feedback.

A possible, within-subject design would be intermixing "pre-bunking" and "debunking" trials to compare the effectiveness of both strategies within the same individual. In the debunking trials, where information turns out false, a retrospective correction might be involved, a cognitively demanding process where an already encoded, affectively charged learning signal should be suppressed or undone. As my current results demonstrate, this correction is often incomplete, leading to residual learning from false information. In contrast, on pre-bunking trials, I would first warn participants that the upcoming feedback is false. This "inoculation" allows for the engagement of proactive filtering. Instead of correcting an error, the brain can prepare to treat the subsequent information as irrelevant. I hypothesize that this proactive stance would alter feedback processing in two ways. First, it could effectively gate the learning signal, preventing the prediction error generated by the outcome from being used to update value representations in the vmPFC. Second, by framing the outcome as meaningless from the start, it could dampen the initial affective response to a win or a loss, neutralizing the very biased valuation shown in Chapter 5 that potentially contributes to this form of irrational learning. Therefore, learning from false information would be significantly lower on pre-bunking trials compared to debunking trials. Specifically, the learned preference for the misleadingly "good" option in the unsolvable condition should be drastically reduced, if not eliminated, on pre-bunking trials. It should be noted that Vidal-Perez et al., (2025) have already shown that learning from unreliable information is still present in a prebunking paradigm.

This behavioural difference should be captured by the parameters of the four-learning-rate model. I could also extend the model to estimate separate learning rates for pre-bunking and debunking trials. My prediction is that the learning rates for false information (α_conf_false and α_disconf_false) would be significantly smaller for pre-bunking trials, ideally approaching zero, providing formal evidence that the learning signal was successfully gated. Demonstrating these effects would provide mechanistic evidence for why inoculation-based interventions are effective. Further, the findings would reveal that preventing a biased belief from taking root is cognitively more efficient than attempting to revise it after the fact.

Beyond simply reducing overall learning from false information, another key question is whether pre-bunking can specifically neutralize the optimistic and confirmation biases that Chapters 3 and 4 show are so robust. The persistence of these biases for false information in my studies suggests they are driven by the initial, stronger response to desirable or choice-confirming information. A debunking cue arrives too late, after this biased emotional signal has already been processed. Pre-bunking, however, is perfectly positioned to intervene in this process. By warning that upcoming information is false, it can dampen the affective response at its source. A piece of good news or a confirmatory outcome that is known in advance to be fake should lose its rewarding quality, thereby neutralizing the very signal that drives the bias. Further, I predict that pre-bunking will specifically attenuate the confirmation bias for false information. Computationally, this means the difference between the learning rates (α_conf_false - α_disconf_false) should be significantly smaller on pre-bunking trials compared to debunking trials. This same logic applies to a pre-bunking version of my belief-updating task. I hypothesize that the optimistic update bias - the greater belief updating for "False Good News" versus "False Bad News" - would also be significantly reduced or eliminated on pre-bunking trials.

6.3.2 Does withholding information cues affect learning from false information?

A limitation of my studies is that information was always explicitly labelled as either true or false. In the real world, however, we frequently encounter information with unknown veracity. The next step would be to introduce a third, no-cue or Unknown case to investigate how people learn from false information when it is perceived relative to the more uncertain information (see Figure 6.1 for the proposed design). In essence, the probabilities behind the False and Unknown cases are the same (Figure 6.1 (b) and (c)) but participants are not privy to these. This design creates a test of several competing hypotheses about how we treat ambiguous information that is common in the real world. Do we operate under a "truth default," treating unknown information as if it's true? Or do we assume it is true if it confirms our beliefs and false if it disconfirms them? A third possibility is that we engage in graded learning, adjusting our learning based on the reliability signal.

Using a 3 (Accuracy: True, False, Unknown) x 2 (Feedback: confirmatory, disconfirmatory) within-subject design, I hypothesize that participants will still learn from false information, but the learning rates will be graded by certainty such that learning from unknown information will be higher than false information, falling between the other two, resulting in an ordered pattern: ɑ_true > ɑ_unknown > ɑ_false (Figure 6.1 (d)). This would suggest that people modulate their learning and belief updating in proportion to the stated certainty of the information, but still, would learn from false information, especially if it confirms their beliefs. Second, I hypothesize that the biases I identified across my studies – confirmation bias and optimistic update bias - would still be present for the "unknown" information. Specifically, for the reinforcement learning task (Chapter 3), I predict that learning from "Confirmatory-Unknown" feedback would be greater than from "Disconfirmatory-Unknown" feedback **(Figure 6.1 (d)).** Similarly, for the belief-updating task (Chapter 4), I predict that belief change would be greater in response to "Good News - Unknown" than to "Bad News - Unknown". Finally, I hypothesize that the strength of the optimistic and confirmation biases across True, False, and Unknown would be the same, just like how there was no interaction between True and False in the strength of the biases in the current studies. Overall, the uncertainty that comes with the Unknown cue may provide the perfect environment for the biases to flourish as there is no explicit factual ground to constrain our desire to believe good news or confirm our choices.

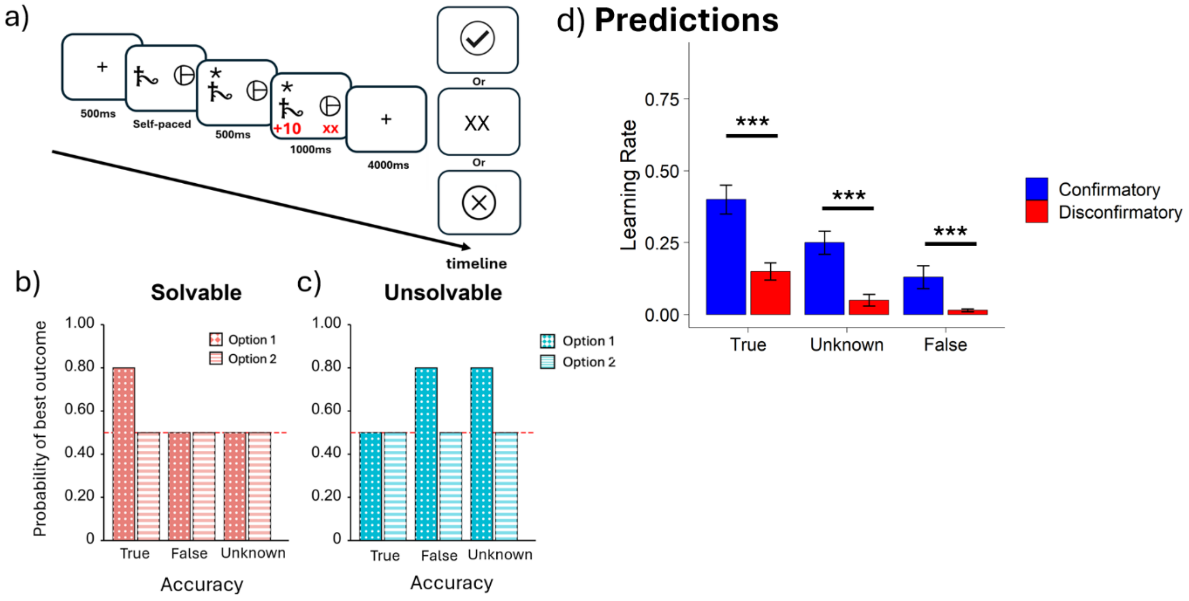


Figure 6.1: The experimental design and its predictions. (a) shows the same reinforcement learning (RL) task as in Chapter 3, but now with three cases for information accuracy: True, False, and Unknown. A key aspect of the design, shown in panels (b) and (c), is that the underlying reward probabilities for the False and Unknown conditions are identical, but participants are not privy to these probabilities. The predictions, visualized in panel (d), are twofold. First, learning is expected to be graded by certainty, with learning rates following the pattern of true > unknown > false. Second, confirmation bias will exist no matter the accuracy, whereby learning from confirmatory feedback will be significantly greater than learning from disconfirmatory feedback.

I could also try different computational models with this design. While estimating a single learning rate (α_unknown) is an obvious first step, more advanced models could reveal the cognitive strategies underlying this parameter. The "Unknown" condition forces a choice: how should this ambiguity be resolved? An alternative model could formalize this as a dynamic process. For instance, α_unknown might not be a fixed value but rather a mixture of α_true and α_false, weighted by a personal "trust" parameter that ranges from 0 to 1. If a person's w is close to 1, it means they are resolving the ambiguity by treating the "Unknown" information as if it were "True." If their w is close to 0, they are treating it as if it were "False." This approach also allows me to see how this tendency is shaped by their confirmation or optimistic update biases. Another idea would be to have a model that proposes we resolve ambiguity in a self-serving manner based on the nature of the feedback itself. When 'Unknown' feedback is confirmatory the model suggests we treat the information as if it were true, applying the confirmatory true learning rate. Conversely, when the feedback is disconfirmatory we dismiss the ambiguous information as unreliable, applying the disconfirmatory false learning rate. This approach formalizes the idea that we interpret ambiguous evidence through the lens of our existing beliefs, readily accepting it as true when it supports our choices and rejecting it as false when it doesn't.

6.3.3 From Valuation to Regulation: The Network Dynamics of Misinformation Valuation

Chapter 5 identified the ventromedial prefrontal cortex (vmPFC) as a hub for the biased computation of value, showing that the vmPFC treats false information that disconfirmed one's choice as more valuable than that which confirmed it. Therefore, it seems that the vmPFC dynamically recalculates the value of information when its veracity is revealed, generating different neural signals when a confirmatory vs disconfirmatory feedback turns out true vs false. These findings establish the vmPFC's central role in the *valuation* aspect of biased learning from true and false information. The vmPFC, however, does not operate in isolation. The interaction I observed at the information cue stage suggests a potential process of retrospective re-evaluation, where the initial value representation at the feedback stage is modulated. The next logical step; therefore, is to investigate the network-level dynamics that support this process. Specifically, how does the brain's metacognitive network interact with the vmPFC to suppress or update value signals in light of new information about their reliability?

The process of evaluating the "True/False" cue in my task could be conceived as an act of metacognitive judgment. As outlined in a recent review by Fleming (2024), such judgments can be deconstructed into several distinct computational components, which are likely supported by different neural systems. These components include: (i) the initial representation of uncertainty about the world, (ii) the transformation of this uncertainty into a propositional confidence judgment about one's own performance, (iii) the global broadcast of this confidence signal to other brain systems for control and communication, and (iv) the influence of a self-model that provides top-down beliefs about one's own abilities. My study is suited to investigate the interplay between these components. The initial feedback (confirmatory/disconfirmatory) generates a first-order choice - actual performance - which my results link to the vmPFC. The subsequent accuracy cue then prompts a second-order metacognitive evaluation, forcing the system to reflect on and potentially revise its initial state.

The literature points to a network of prefrontal regions responsible for this regulation. This network can be subdivided based on the components of a metacognitive judgment. One component is about representing propositional confidence - the brain's estimate of the probability that a specific choice or belief is correct, which is distinct from lower-level sensory uncertainty (Fleming, 2024). My finding of a biased valuation signal in the ventromedial prefrontal cortex (vmPFC) aligns with work showing this region encodes signatures of propositional confidence (Bang & Fleming, 2018). Next, for a confidence estimate to be useful, it must be subject to global broadcast and communication with other brain systems to guide subsequent thought and behaviour. This broadcasting and strategic use of confidence is thought to involve the rostrolateral prefrontal cortex (RLPFC), anterior prefrontal cortex (aPFC), and frontopolar cortex (Fleming, 2024). The aPFC is specifically implicated in mediating the impact of post-decision evidence on subjective confidence, a process central to my task (Fleming et al., 2018), making it a prime candidate for receiving the initial value signal from the vmPFC and initiating a revision. Finally, performance monitoring and control are central functions of the dorsal anterior cingulate cortex (dACC) and dorsolateral prefrontal cortex (dlPFC), which monitor for conflict and apply rules to guide behaviour (Miller & Cohen, 2001; Shenhav et al., 2013). The "False" cue in my task could be thought of as an error signal that engages this network, which in turn would provide the necessary top-down signals to execute the belief update.

To formally test hypotheses about these network-level interactions, I propose to use Dynamic Causal Modelling (DCM). DCM allows for the testing hypotheses about how brain regions influence one another's activity and how these connections are modulated by experimental conditions (K. J. Friston et al., 2003). This makes it the ideal tool to move beyond asking "what areas are active?" to asking "how do these areas work together to revise beliefs?" For instance, one could ask how does the frontoparietal metacognitive network regulate the vmPFC to update beliefs when the reliability of feedback is revealed?

Using DCM, I can test a plausible model of this network. The model would include nodes for the vmPFC (representing propositional confidence) and key metacognitive regions like the RLPFC, aPFC and dlPFC/dACC. This leads to several testable hypotheses. First, I predict a top-down correction signal, where the presentation of the "False" cue will significantly modulate the effective connectivity from a cognitive control region (likely the dlPFC) to the vmPFC. This directed influence would represent a top-down "correction" or "gating" signal that implements the revision of the initial value representation. Second, concerning the role of the RLPFC and aPFC in initiating the update, I predict that these regions will play a mediating role, showing modulated connectivity from the vmPFC and to the dlPFC. This would be consistent with their proposed role in monitoring initial confidence signals and initiating a revision by recruiting the cognitive control side of the dlPFC (Fleming, 2024). Finally, I predict that the neural signature of "relief" will manifest as a network reconfiguration. I expect that the strength of this top-down modulation from the metacognitive network to the vmPFC will be significantly stronger for disconfirmatory false and confirmatory true feedback. This provides a mechanistic explanation for the "relief" signals I observed in the vmPFC's activity. Confirming these hypotheses would provide a network-level account of belief revision, dissociating the initial, biased, bottom-up valuation signal computed in the vmPFC from a subsequent, top-down regulatory process by the prefrontal cortex.

6.3.4 Is learning from false information dependant on working memory?

A large body of research has established that human reinforcement learning is not a monolithic process but relies on the interplay between multiple cognitive systems (see Yoo & Collins, 2022 for a review). Work by Anne Collins and her colleagues has demonstrated that a slow, incremental RL system operates in parallel with a fast, flexible, but capacity-limited working memory (WM) system (Collins & Frank, 2012; Collins, 2018). This line of research has shown that as the number of items to learn - the set size - increases, the WM system becomes taxed, impacting learning strategies (Collins & Frank, 2012). In my task, the challenge is twofold: participants must not only learn the values of the options but also actively suppress learning from information they know is false. This suppression is an executive function that relies on the same limited cognitive resources as WM. Therefore, I propose that by increasing the cognitive load on the WM system, I can test its role in gating irrelevant information and resisting cognitive biases. I can achieve this by varying the set size - having different groups learn a low number of pairs (e.g., 2), a medium number (the original 4), or a high number (e.g., 6).

If ignoring false feedback requires these limited cognitive resources, then increasing the load on that system should impair this ability. This leads to two specific hypotheses. First, overall learning from false information will increase with cognitive load. I predict that as the set size increases, participants will be more influenced by the misleading feedback. This would be evident by a stronger preference for the "best" option in the unsolvable condition for the group with a set size of 6 compared to the group with a set size of 2. Second, the confirmation bias for false information will be magnified under high load. I hypothesize that the bias itself will become stronger as cognitive resources are depleted. When the working memory system is taxed, the brain may rely more on default, heuristic-based learning. Computationally, this would manifest in the parameters of my model: the difference between the learning rates for false confirmatory and false disconfirmatory feedback (α_false_confirmatory - α_false_disconfirmatory) should be significantly larger in the high-load (set size 6) condition. For true information, however, the bias might not vary between the set sizes as ignoring true information is not a goal.

Beyond predicting how existing learning rates will change, a more powerful approach would be to develop models that explicitly represent working memory like how previous works have done (Collins & Frank, 2012; Collins, 2018). One approach would be to model the suppression of false information as an active "gating" process that is dependent on cognitive resources. This could be formalized by introducing a new "cognitive control" parameter, κ (kappa), which represents the efficacy of the gating mechanism. The effective learning rate from false feedback would then be dynamically modulated by this parameter (e.g., effective α_false = (1 - κ) * base α_false). Under low load, κ would be high, effectively driving learning from false information towards zero. As cognitive load (set size) increases, κ would decrease, representing the depletion of resources and causing a failure in the gating mechanism. This model would allow for a direct test of whether cognitive load impairs the *control process* itself, rather than just altering the learning rates. Another approach would be to use a dual-system model that has separate RL and WM components. In this framework, the capacity-limited WM system could be responsible for both fast learning *and* for maintaining the task rule ("ignore false feedback"). As set size increases, the WM component's contribution to learning diminishes. If this component is also responsible for the suppression signal, its failure under load would predict a corresponding increase in learning from false information. This would provide a unified account of how set size affects both learning and the ability to ignore misinformation.

Confirming these hypotheses would provide evidence that the ability to resist misleading information is not fixed but depends on the availability of cognitive resources. It would suggest that individuals are likely more vulnerable to misinformation when they are distracted, tired, or otherwise cognitively taxed - a finding with real-world implications for how we consume information in our daily lives, especially on social media.

6.3.5 Confirmation Bias in Learning from False Information in a Social Context

My thesis has established a confirmation bias in how individuals learn from misinformation in a non-social context. A timely extension of this work is to investigate whether this same bias operates when information comes from a social source. In the modern world, much of the information we consume - and the misinformation we encounter - is transmitted through social networks like X (formerly Twitter) (Vosoughi et al., 2018). The architecture of these networks, which allows for rapid, widespread dissemination of user-generated content, creates a fertile ground for biases like confirmation bias to take hold (Aral & Van Alstyne, 2010). Further, their incentive structures (likes and shares) are not designed to prioritise accuracy (Globig & Sharot, 2024). Instead, they reward content that is popular, emotionally evocative, or identity-affirming. This creates a social reward landscape that can be orthogonal, or even antithetical, to truth. Algorithmic personalization and the tendency for users to self-select into ideologically aligned groups can create "filter bubbles" and "echo chambers," which reduce exposure to diverse viewpoints and amplify confirmatory content (Glickman & Sharot, 2024). Turner et al. (2025) applied a computational model of reward learning to real-world X data, showing that users' posting behaviour is modulated by the social rewards they receive. This creates an environment where a confirmation bias is particularly potent. If users are socially rewarded for sharing confirmatory content, and our brains are already wired to preferentially learn from it, the result is a feedback loop that can drive the rapid formation of polarized communities. For instance, in a large-scale analysis of Facebook users, Zollo et al. (2017) found that people form highly segregated "tribes" around specific narratives, creating echo chambers where attempts to debunk misinformation are largely ineffective because users preferentially engage with content that confirms their group's identity, leading to the wide-scale propagation of misinformation (Zollo et al., 2017).

Learning from other people is fundamentally different than learning from non-social probabilistic feedback, requiring us to build a model of the person providing information and engaging in "mentalizing" or "Theory of Mind" - the fundamental human ability to attribute unobservable mental states like beliefs, desires, and intentions to others (Frith & Frith, 2006). When learning from a social source, we process the information itself and build a predictive model of the other person's mind. We constantly, and often unconsciously, ask ourselves: Are they knowledgeable? Are they trying to be helpful, or do they hold a bias? This continuous inference about the minds of others is what makes social learning computationally complex and distinct from learning from simple environmental feedback (Saxe, 2006).

To test whether the confirmation bias for false information persists in a social context, I propose adapting the multiplayer RL paradigm from Zhang and Gläscher (2020). This task is suited for dissociating private belief from social influence. Its multi-phase design allows for the separate measurement of an initial choice, the influence of social information, and a final, updated choice and confidence level. The task begins with the participant making an initial choice in a learning problem and then placing a bet to indicate their confidence in that decision. Following this, they are shown the choices made by four other players. With this new social information, the participant is given an opportunity to either stay with their original choice or switch to a different one. They are also allowed to update their confidence bet. Finally, the correct outcome for the trial is revealed to all players, allowing them to learn from the result.

To study the impact of false information, I will manipulate the reliability of the social sources. This approach is grounded in the advice-taking literature, which shows that people are typically sensitive to an advisor's past performance and adjust how they weigh their advice accordingly (Bonaccio & Dalal, 2006). Participants will be told they are playing with a group where some "players" are unreliable bots. This creates two types of social sources: Reliable players and bots. Reliable players are programmed to perform well (e.g., choosing the correct option 80% of the time), representing a source of "true" social information. Bots are programmed to choose at random (50/50), representing a source of noisy, "false" social information, akin to the False cues in the tasks of the current thesis. Over the course of the experiment, the participant should learn not only the value of the task options but also the credibility of their co-players. This allows me to test whether participants learn to discount the choices of the unreliable bots, a process that can be formally captured by extending the computational model from the original paper to include learned reliability weights for each social partner. Layered on top of the reliability manipulation, I will program the social feedback on a subset of trials to be either unanimously confirmatory (all four co-players agree with the participant's Choice 1) or unanimously disconfirmatory (all four co-players disagree). This unanimous feedback will come from a group that the participant has either learned is reliable or learned is unreliable.

This design creates a direct conflict between a rational assessment of source credibility and the pull of confirmation. The key dependent variable is the participant's change in confidence (the difference between Bet 2 and Bet 1). My prediction is that the confirmation bias will override the rational assessment of the unreliable source. When feedback comes from a reliable group, confidence should increase after confirmation and decrease after disconfirmation. This would replicate standard findings in social influence and collective intelligence (Bahrami et al., 2010). When feedback comes from an unreliable group, a rational agent should ignore it and show no change in confidence. However, I hypothesize that participants will still show a significant boost in confidence after receiving confirming feedback from the unreliable group but will successfully ignore disconfirming feedback from that same group, mirroring the results of Chapter 3. This would be a demonstration that the confirmation bias for false information also exists in a social context and that we are willing to take a confidence boost from any source, providing a cognitive mechanism for our vulnerability to confirmatory misinformation online.

6.3.6 Do large language models exhibit confirmation bias for false information?

My research demonstrates a robust confirmation bias in humans, which persists even when they know the information confirming their choices is false. I could investigate whether this specific pattern of biased learning extends to artificial agents, particularly Large Language Models (LLMs). Recent work has begun to establish a field of "AI Psychology," showing that LLMs exhibit many human-like cognitive biases during in-context learning. First, LLMs, much like humans, exhibit asymmetric belief updating (Schubert et al., 2024). Critically, when an LLM was given full feedback on both its chosen and unchosen options, it displayed a classic confirmation bias, learning more from outcomes that confirmed its past decisions. This bias vanished when the LLM lacked a sense of agency (i.e., when observing choices made by "someone else"), highlighting the importance of the model's perceived role in the decision-making process. Second, Hayes et al. (2025) found that LLMs are susceptible to relative value encoding biases and, using computational modelling, also found evidence for a confirmation bias in how the models learned from feedback. However, these foundational studies operate on the assumption that the feedback provided to the model is true. This leaves open an important question: is the confirmation bias in LLMs so fundamental that it persists even when the model "knows" the information is false?

To test this, I would adapt the RL paradigm from human trials (detailed in Chapter 3) for an LLM environment. The experiment will be operationalized through an interactive prompt that establishes the LLM's agency and objective (e.g., "You are a participant in a decision-making experiment. Your goal is to maximize your points by learning which of two abstract symbols is 'correct'. On each trial, you will choose a symbol and receive feedback."). It would be a 2x2 experimental design that manipulates the accuracy of the feedback and its nature as either confirmatory or disconfirmatory. In "True" trials, the feedback provided to the LLM is genuine and reliable. In contrast, during "False" trials, the feedback is explicitly labelled as unreliable or a "system glitch," yet it is still presented to the model. "Confirmatory" feedback supports the model's selection (e.g., "You chose Symbol A, and the feedback is Positive"), while "Disconfirmatory" feedback contradicts it (e.g., "You chose Symbol A, and the feedback is Negative").

I predict the LLM will learn less from false information than from true information. That is, its preference for the "best" option will be stronger in the solvable condition (where true feedback is informative) than in the unsolvable condition (where false feedback is informative), replicating the current findings. Further, I predict the LLM will exhibit a confirmation bias for both true and false information. It will learn more from feedback that confirms its choices, regardless of whether that feedback is explicitly labelled as genuine or a "glitch."

Discovering such a bias would be a significant finding. It would suggest that the architectural or training principles that lead to confirmation bias in LLMs exist even in the face of explicit falsehoods, mirroring my human data. This would have implications for AI safety, revealing a potential vulnerability where models could irrationally persist in a course of action based on desirable but demonstrably false feedback. An LLM that reinforces its actions based on desirable but false feedback could be susceptible to manipulation, reward hacking, or irrationally persisting with a flawed strategy.

6.3.7 Using EEG to Dissect the Temporal Dynamics of Learning from False Information

An unresolved question from my thesis is around the timing of learning or belief updating. Does the brain immediately update value representations upon receiving feedback, requiring a subsequent correction if that feedback proves false? Or is the initial feedback held in working memory, with the update gated until after its veracity is revealed? To test these competing hypotheses, I propose a study combining Electroencephalography (EEG) with a novel experimental manipulation. By recording EEG while participants perform the reinforcement learning task from Chapter 3, I can examine the event-related potentials (ERPs) - time-locked neural responses - elicited by the two events in each trial: the feedback and the information cue presentations. Further, I will introduce a manipulation that further probes this temporal dynamic: I will vary the duration of the fixation cross between the feedback and the information cue. This interval will be short (e.g., 250ms), medium (e.g., 1s), or long (e.g., 4s). This manipulation allows me to control the time available for post-feedback processing, and thus to test whether belief updating is an immediate or a delayed process.

The first important event is when a participant receives the outcome (e.g., "+10" or "-1"). At this point, a prediction error (PE) is generated, but the participant does not yet know if the outcome is reliable. I can measure the neural signature of this initial, potentially biased, PE using the Feedback-Related Negativity (FRN) (Holroyd & Coles, 2002; Walsh & Anderson, 2012). The FRN is a well-established ERP component that peaks approximately 250-300ms after feedback onset over frontocentral scalp sites. It is thought to originate from the medial frontal cortex, including the anterior cingulate cortex, and is considered a robust neural correlate of PEs, being larger (more negative) for worse-than-expected outcomes like losses or non-rewards (Holroyd & Coles, 2002; Walsh & Anderson, 2012). I predict that the amplitude of the FRN will be modulated by confirmation. Specifically, disconfirmatory feedback will elicit a significantly larger FRN than confirmatory feedback. This would provide a neural index of the initial, biased PE *before* the feedback's veracity is known. Because the FRN is a rapid, early component reflecting a largely automatic evaluation of the outcome (Ullsperger et al., 2014), its amplitude should *not* be affected by the subsequent delay manipulation. It provides a clean measure of the brain's initial reaction, independent of the time allowed for later deliberation. The second important event is when the participant sees the "True" or "False" information cue. This information potentially forces the participant to either solidify or revise their belief. I can measure the neural signature of this potential updating process using the P300 (or P3b). The P300 is a large, positive-going ERP component peaking 300-600ms after a task-relevant stimulus, with a parietal scalp distribution. Its amplitude reflects the allocation of attentional resources to update one's mental model of the environment and is larger for more surprising or motivationally significant events (Polich, 2007).

The delay manipulation targets the processing that occurs before the information cue, allowing me to test two competing models of belief updating. First, If the brain follows an "Update and Correct" model, an immediate, biased update occurs at feedback. A longer delay (4s) allows this initial belief to consolidate, making a subsequent correction at the information cue more effortful. The P300 amplitude should be *larger* after long delays in response to the "False" cue, reflecting a more significant revision of a consolidated belief trace, consistent with its role in processing task-updating information (Donchin & Coles, 1988). Further, learning from false information should be *stronger* after longer delays, as the initial biased update has had more time to "stick" and become resistant to revision. Second, if the brain follows a "Hold and Update" model, the feedback is held in working memory until the information cue triggers the update. A longer delay (4s) would lead to the decay of this information in working memory. The P300 amplitude should be smaller after longer delays, consistent with findings that P300 amplitude is reduced when processing stimuli that are less certain or built on degraded memory traces (Kok, 2001).

This combined EEG and delay manipulation study would provide a complete temporal narrative of learning from false information. It would dissociate the initial, automatic response to feedback (indexed by the FRN) from the later, more controlled process of belief revision (indexed by the P300). Most importantly, by observing how the delay causally affects both the P300 and behaviour, I can provide strong evidence for one of two distinct computational models of how and when we update our beliefs in the face of misinformation.

6.4 Conclusion

Through a combination of behavioural experiments, computational modelling, and neuroimaging, the studies in Chapters 3, 4, and 5 converged on the finding that humans are biased in learning from false information – optimistic update bias (Chapter 4) and confirmation bias (Chapters 3 and 5) - whereby they learn more from misinformation that gives them desirable vs undesirable information. Chapter 5 complimented this finding by showing that vmPFC activity, known for valuation, was modulated by the interaction between confirmation and accuracy, placing a higher value on information that validated prior desirable vs undesirable beliefs or invalidated prior opposing vs supporting evidence, and placing a lower value on confirmatory vs disconfirmatory information that turned out false, suggesting self-serving patterns of valuation at the expense of accuracy. These tendencies for desirable information likely serve to generate and maintain positive affective states, even at the expense of information accuracy. Algorithms then exploit these biases for engagement, tailoring content to one's needs and desires and rendering interventions such as debunking less effective. Finally, a limitation across all studies is the use of authoritative accuracy cues. I will address this limitation by introducing a no-cue condition to make the experiment more ecologically valid where the veracity of information is more ambiguous.

Appendices

Appendices for Chapter 3

Appendix 3.1: Supplementary Model 1 - Disentangling Confirmation Bias from Positivity Bias

The main text presents Model 4, which operationalises confirmation bias using four learning rates that depend on information accuracy and whether feedback is confirmatory or disconfirmatory. However, the definition of "confirmatory feedback" in Model 4 combines two distinct events: a positive prediction error (PE) for a chosen option (factual outcome) and a negative PE for an unchosen option (counterfactual outcome), and the reverse for disconfirmatory feedback. This parameterisation, while elegant, confounds a true confirmation bias with simpler underlying biases, such as a positivity bias, which is a general tendency to learn more from positive PEs than negative PEs. The purpose of Supplementary Model 1 is to de-confound these potential mechanisms. It expands on Model 4 by assigning a unique learning rate to each combination of accuracy, the outcome shown for chosen or unchosen option, and PE sign, resulting in eight learning rates. This allows me to isolate the specific influence of each factor and determine if the patterns attributed to confirmation bias in Model 4 are better explained by positivity bias.

The eight learning rates are partitioned according to three trial-by-trial conditions: the Accuracy of the information (cued as True or False), the outcome shown (for the Chosen or Unchosen option), and the sign of the prediction error (Positive or Negative).

For Factual Outcomes (Chosen Option):

If Accuracy = True:

$$Q(t+1) = Q(t) + \alpha_{\text{Pos,Chosen, true}} * \delta(t) \qquad \delta(t) = \text{Positive}$$

$$Q(t+1) = Q(t) + \alpha_{\text{Neg,Chosen, true}} * \delta(t) \qquad \delta(t) = \text{Negative}$$

If Accuracy = False:

$$Q(t+1) = Q(t) + \alpha_{\text{Pos,Chosen, false}} * \delta(t) \qquad \delta(t) = \text{Positive}$$

$Q(t+1) = Q(t) + \alpha_{Neg,Chosen\ false} * \delta(t) \qquad \delta(t) = \text{Negative}$

For Counterfactual Outcomes (Chosen Option):

If Accuracy = True:

$Q(t+1) = Q(t) + \alpha_{Pos,Unchosen,\ true} * \delta(t) \qquad \delta(t) = \text{Positive}$

$Q(t+1) = Q(t) + \alpha_{Neg,Unchosen,\ true} * \delta(t) \qquad \delta(t) = \text{Negative}$

If Accuracy = False:

$Q(t+1) = Q(t) + \alpha_{Pos,Unchosen,\ false} * \delta(t) \qquad \delta(t) = \text{Positive}$

$Q(t+1) = Q(t) + \alpha_{Neg,Unchosen\ false} * \delta(t) \qquad \delta(t) = \text{Negative}$

Choice probabilities are generated using the SoftMax function with an inverse temperature parameter β. The model is fitted by minimizing the negative log-likelihood of the participant's sequence of choices, as described in the main text.

Free parameters (n=9): $\alpha_{Pos,Chosen,\ false}$, $\alpha_{Neg,Chosen\ false,}$ $\alpha_{Pos,Chosen,\ true}$, $\alpha_{Neg,Chosen,\ true}$, $\alpha_{Pos,Unchosen,\ false}$, $\alpha_{Neg,Unchosen\ false}$, $\alpha_{Pos,Unchosen,\ true}$, $\alpha_{Neg,Unchosen,\ true}$, β

Appendix 3.2: Supplementary Model 2 - Confirmation Bias Across Gain and Loss Contexts

In this model my goal was to see if confirmation bias is robust across Gain (outcomes: +10, +1) and Loss (outcomes: -10, -1) contexts, so I created separate learning rates for each. The model had eight learning rates partitioned according to three conditions: the Context (Gain or Loss), the Accuracy of the information (True or False), and the sign of the prediction error (Positive or Negative).

For the Gain Context:

If Accuracy = True:

$Q(t+1) = Q(t) + \alpha_{Pos,Gain,\ true} * \delta(t) \qquad \delta(t) = \text{Positive}$

$Q(t+1) = Q(t) + \alpha_{Neg,Gain,\ true} * \delta(t) \qquad \delta(t) = \text{Negative}$

If Accuracy = False:

$Q(t+1) = Q(t) + \alpha_{Pos,Gain,\ false} * \delta(t) \qquad \delta(t) = \text{Positive}$

$Q(t+1) = Q(t) + \alpha_{Neg,Gain,\ false} * \delta(t) \qquad \delta(t) = \text{Negative}$

For the Loss Context:

If Accuracy = True:

$$Q(t+1) = Q(t) + \alpha_{Pos,Loss,\ true} * \delta(t) \qquad \delta(t) = \text{Positive}$$

$$Q(t+1) = Q(t) + \alpha_{Neg,Loss,\ true} * \delta(t) \qquad \delta(t) = \text{Negative}$$

If Accuracy = False:

$$Q(t+1) = Q(t) + \alpha_{Pos,Loss,\ false} * \delta(t) \qquad \delta(t) = \text{Positive}$$

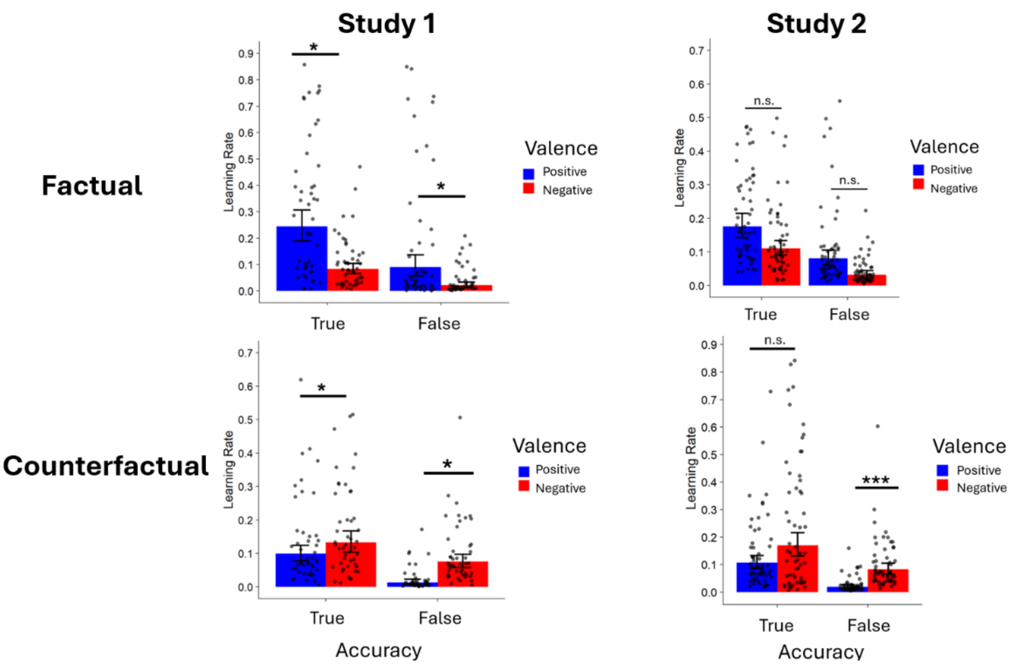
$$Q(t+1) = Q(t) + \alpha_{Neg,Loss,\ false} * \delta(t) \qquad \delta(t) = \text{Negative}$$

Choice probabilities are generated using the SoftMax function with a single inverse temperature parameter β fitted across both contexts. The model is fitted by minimizing the negative log-likelihood.

Free parameters (n=9): $\alpha_{Pos,Gain,\ false}$, $\alpha_{Neg,Gain,\ false,}$ $\alpha_{Pos,Gain,\ true}$, $\alpha_{Neg,Gain,\ true}$, $\alpha_{Pos,Loss,\ false}$, $\alpha_{Neg,Loss,\ false}$, $\alpha_{Pos,Loss,\ true}$, $\alpha_{Neg,Loss,\ true}$, β

Appendix 3.3: Factual and Counterfactual Learning Rates

To ensure my results indicated confirmation bias and not positivity bias, I created a model with eight separate learning rates. This allowed me to separately look at the factual learning rate - for the chosen option's outcome - and the counterfactual learning rate - for the unchosen option's outcome. This approach expanded on my previous model (M4), which collapsed these into general 'confirmatory' and 'disconfirmatory' rates (see Methods). This separation was crucial for testing two competing predictions. A positivity bias would mean participants learn more from positive outcomes than negative ones across all trials, regardless of what they chose. In contrast, confirmation bias would mean participants should learn more from outcomes that confirm their initial decisions. This would mean learning more from positive outcomes on factual trials (to confirm their choice was good) but learning more from negative outcomes on counterfactual trials (to confirm rejecting the other option was also good).



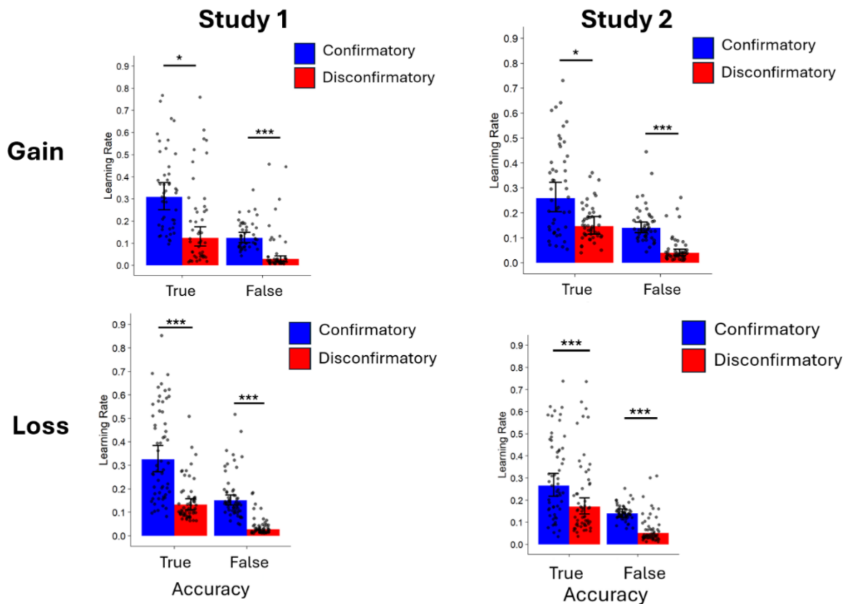
Appendix Figure 3.1: Factual and Counterfactual learning rates. The opposite pattern of learning rates for factual (above panel, though note that for Study 2 the pattern is qualitative as they are not significant) and counterfactual (below panel) trials indicating confirmation bias, *not* positivity bias, for true and false information in both studies. n.s: not significant, *p < 0.05, ***p < 0.001, hierarchical t-test.

As shown in Appendix Figure 3.1, my findings were consistent with a confirmation bias. In both studies, for factual trials, I found qualitative patterns of higher learning from positive outcomes than negative ones, which were flipped in the counterfactual trials. Specifically, in study 1, they learned more from positive than negative outcomes for factual trials (for true information: t(46) = 2.59, p = 0.01; paired t-test between positive and negative learning rates, and for false information: t(46) = 2.20, p=0.03; paired t-test between positive and negative learning rates) while this pattern flipped for counterfactual trials, learning more from *negative* that positive outcomes (for true: t(46) = −2.75, p = 0.008; paired t-test between positive and negative learning rates, and for false information: t(46) = −2.47, p = 0.01; paired t-test between positive and negative learning rates). The same patterns emerged in study 2 but some of the effects did not reach significance (for factual, true information: : t(56) = 1.88, p = 0.068; paired t-test between positive and negative learning rates, and for false information: t(56) = 1.86, p=0.067; paired t-test between positive and negative learning rates, for counterfactual true information: t(56) = −1.87, p = 0.06; paired t-test between positive and negative learning rates, and for false information: t(56) = −3.81, p < 0.010; paired t-test between positive and negative learning rates).

Appendix 3.4: Gain and Loss Learning Rates

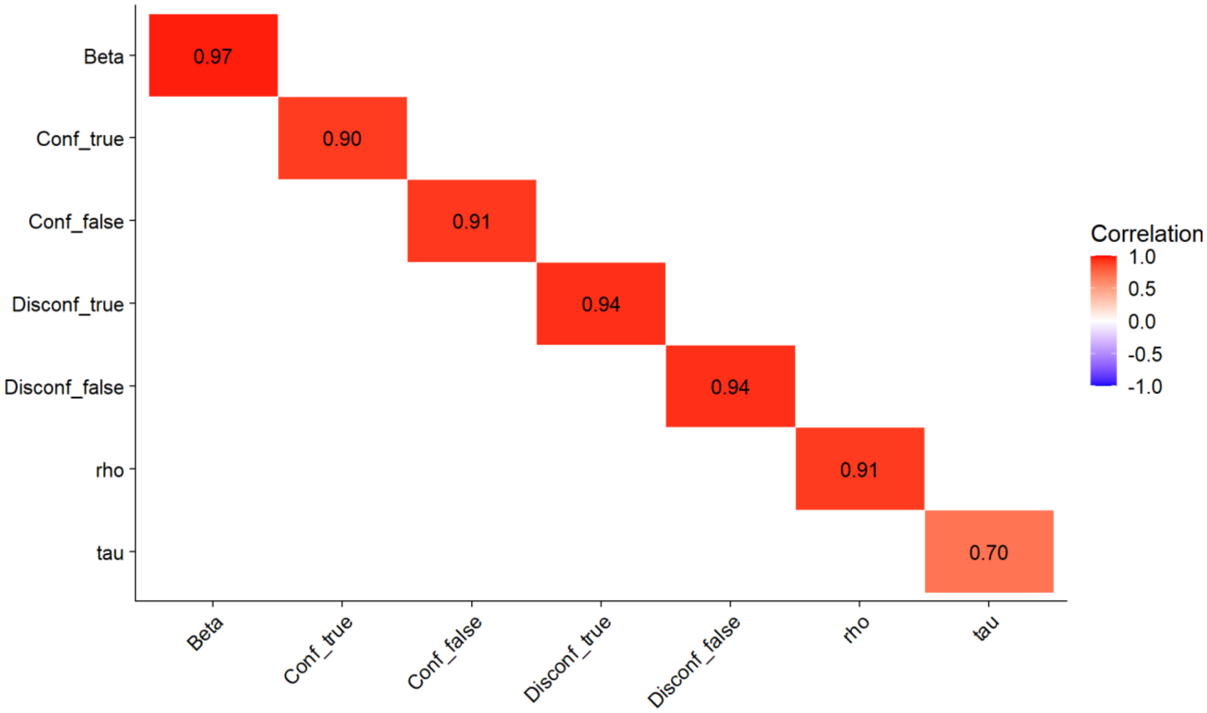
Next, I implemented a model with separate learning rates for the Gain context (where outcomes were positive: +10, +1) and the Loss context (where outcomes were negative: -10, -1). My goal was to see if the bias for true and false information would generalize across both as in one case (the gain contexts) getting a low reward (+1) needs to be framed as a loss whereas in the other (the loss context) incurring a small loss needs to be framed as a gain. Differences in approach learning between gains and losses have been reported in the past (Guitart-Masip et al., 2012) but it is unclear whether these impact confirmation bias.

As shown in Appendix Figure 3.2 confirmation bias is robust in the Gain context for both true information (t(46) = 2.51, p = 0.01; t(56) = 2.46, p = 0.01) and false information (t(46) = 3.99, p < 0.001; t(56) = 5.06, p < 0.001). Similarly, the bias exists in the Loss context for both true information (t(46) = 4.05, p < 0.001; t(56) = 3.48, p < 0.001) and false information (t(46) = 4.37, p < 0.001; t(56) = 6.25, p < 0.001), replicating previous findings on true information (Palminteri et al., 2017) and extending them to false information.



Appendix Figure 3.2: Confirmation bias for Gain and Loss contexts. The bias for true and false information is robust across Gain (above panel) and Loss (lower panel) contexts in both studies. *p < 0.05, ***p < 0.001 hierarchical t-test.

Appendix 3.5: Parameter Recovery for the Gradual Perseveration Model

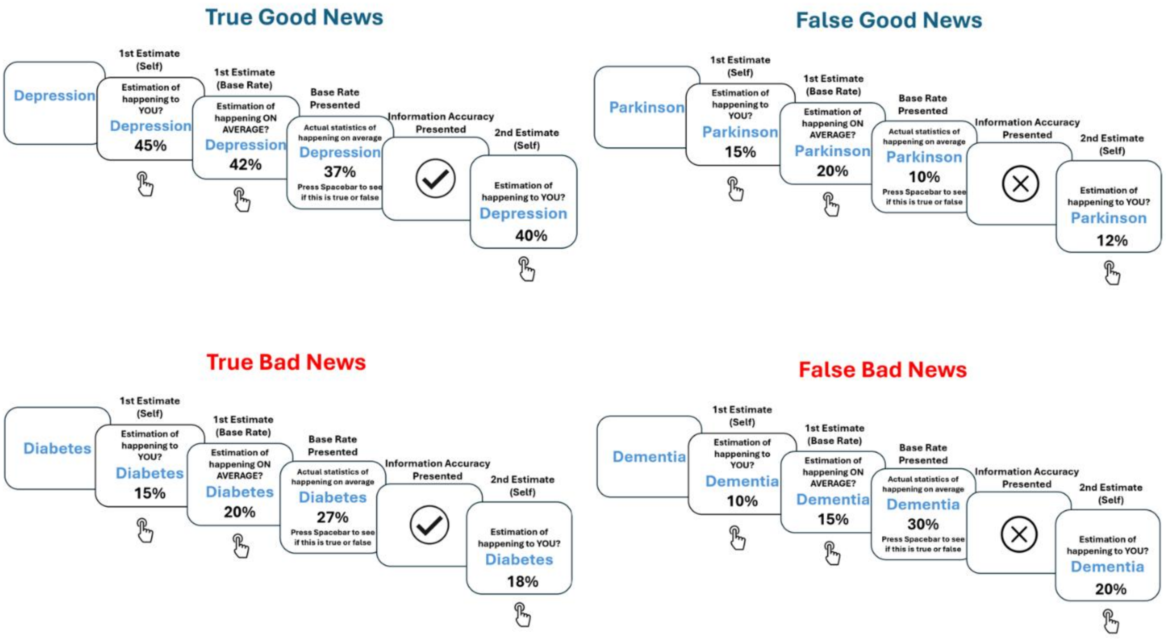


Appendix Figure 3.3: The Gradual Perseveration Model's Parameter Recovery. The choice trace parameter shows suboptimal recovery, which might make the estimates from this model unreliable

Appendices for Chapter 4

Appendix 4.1: The four possible trial types in the Update Bias Task

The following figure illustrates the four possible trial types in my experimental design:



True Good News
Depression
1st Estimate (Self)
Estimation of happening to YOU?
Depression
45%
1st Estimate (Base Rate)
Estimation of happening ON AVERAGE?
Depression
42%
Base Rate Presented
Actual statistics of happening on average
Depression
37%
Press Spacebar to see if this is true or false
Information Accuracy Presented
✓
2nd Estimate (Self)
Estimation of happening to YOU?
Depression
40%
False Good News
Parkinson
1st Estimate (Self)
Estimation of happening to YOU?
Parkinson
15%
1st Estimate (Base Rate)
Estimation of happening ON AVERAGE?
Parkinson
20%
Base Rate Presented
Actual statistics of happening on average
Parkinson
10%
Press Spacebar to see if this is true or false
Information Accuracy Presented
⊗
2nd Estimate (Self)
Estimation of happening to YOU?
Parkinson
12%
True Bad News
Diabetes
1st Estimate (Self)
Estimation of happening to YOU?
Diabetes
15%
1st Estimate (Base Rate)
Estimation of happening ON AVERAGE?
Diabetes
20%
Base Rate Presented
Actual statistics of happening on average
Diabetes
27%
Press Spacebar to see if this is true or false
Information Accuracy Presented
✓
2nd Estimate (Self)
Estimation of happening to YOU?
Diabetes
18%
False Bad News
Dementia
1st Estimate (Self)
Estimation of happening to YOU?
Dementia
10%
1st Estimate (Base Rate)
Estimation of happening ON AVERAGE?
Dementia
15%
Base Rate Presented
Actual statistics of happening on average
Dementia
30%
Press Spacebar to see if this is true or false
Information Accuracy Presented
⊗
2nd Estimate (Self)
Estimation of happening to YOU?
Dementia
20%

Appendix Figure 4.1: Trial Types in the Update Bias Task

True Good News: In this example, participants initially estimated their risk of depression at 45% (1st Estimate). When presented with the actual base rate (37%), which was lower than their initial estimate, they received "good news" that their risk was overestimated. The checkmark (✓) indicates that the provided statistic was true, and participants were asked for a 2nd estimate after learning this information.

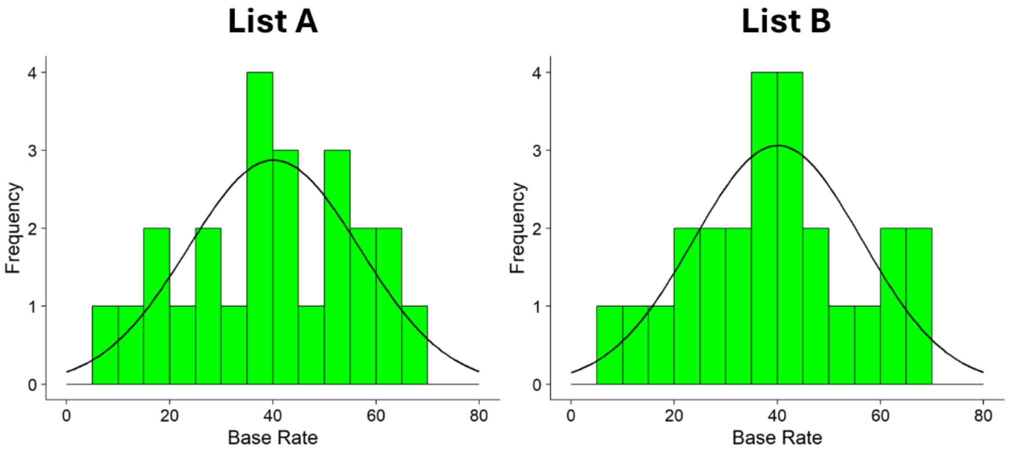
True Bad News: Here, participants initially estimated their risk of diabetes at 15% (1st Estimate). The actual base rate (27%) was higher than their initial estimate, representing "bad news" that they had underestimated their risk. The checkmark (✓) indicates that the provided statistic was true.

False Good News: In this example, participants initially estimated their risk of Parkinson's disease at 20% (1st Estimate). They were shown a purported base rate of 15%, suggesting "good news." However, the cross (✗) indicates that this statistic was false information.

False Bad News: Here, participants initially estimated their risk of dementia at 10% (1st Estimate). They were shown a purported base rate of 30%, suggesting "bad news." The cross (✗) indicates that this statistic was false information.

In all trials, participants provided an initial self-risk estimate (1st Estimate), were presented with the purported average risk (Base Rate) along with an indication of whether this information was true or false and then were asked to provide a second self-risk estimate (2nd Estimate). The relationship between the 1st Estimate and Base Rate determined whether the trial represented "good news" or "bad news" for the participant.

Appendix 4.2: The distribution of base rates in the Update Bias Task



Appendix Figure 4.2: The Distribution of Base Rates in the Update Bias Task

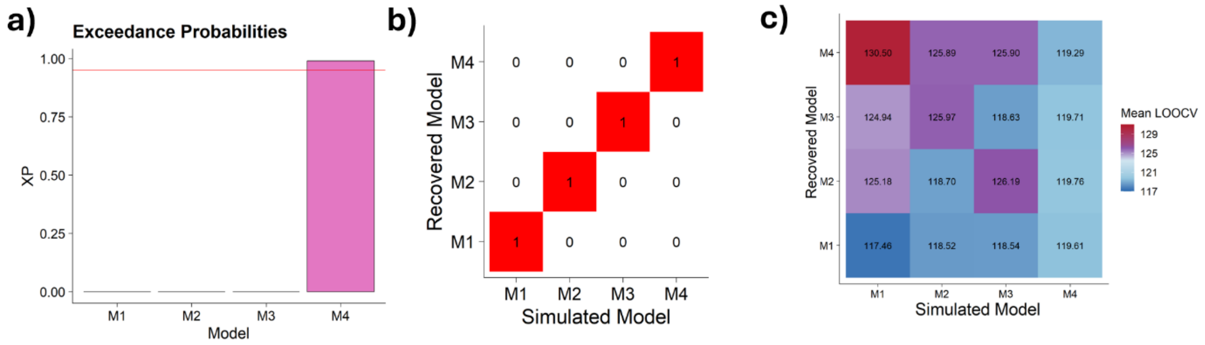
This figure illustrates the distribution of base rates (event probabilities) used in the two stimuli lists. As shown, both List A and List B contain 25 negative life events with base rates that are normally distributed, primarily ranging between 10% and 70%. The x-axis represents the base rate percentages while the y-axis shows the frequency of events at each percentage point.

Both distributions follow similar patterns, ensuring comparable statistical properties between the two lists. This balanced distribution was crucial for our experimental design, as participants were randomly assigned one list for true trials and the other for false trials.

When a list was designated for false trials, the statistics were randomly shuffled among events (e.g., pairing the statistic for domestic burglary with bicycle theft), while maintaining the overall statistical distribution shown here. This shuffling procedure ensured that the false information retained the same statistical properties (median, range, and distribution) as the true information, controlling for any potential biases that might arise from systematic differences in the probability distributions.

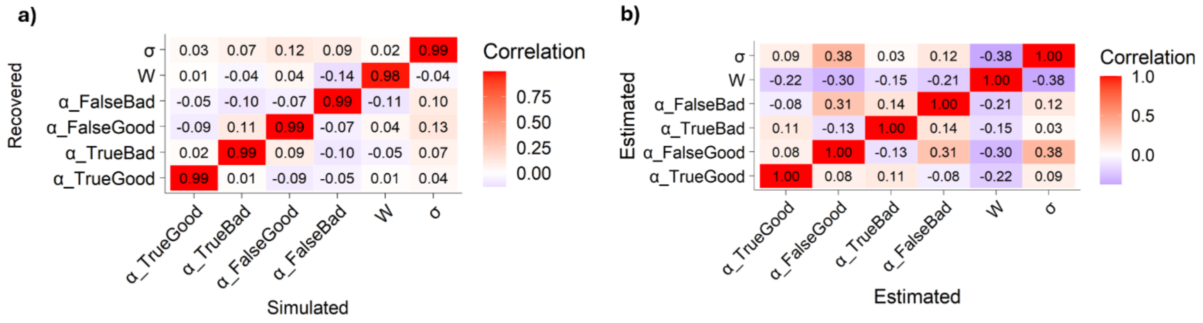
The comparable distributions between List A and List B allowed us to counterbalance the assignment of lists across participants, with each participant receiving one list with true statistics and one with shuffled (false) statistics, creating a final combined list of 50 events.

Appendix 4.3: Additional model diagnostics



Appendix Figure 4.3: Additional Model Diagnostics. (a) Shows the exceedance probabilities (XP), which quantify the confidence that each model is more likely than all other models in the set. M4 achieved an exceedance probability of nearly 1.0, indicating extremely high confidence that it outperforms the competing models. This is further supported by the frequency analysis in the main text, where M4 had an estimated frequency of approximately 0.74, substantially exceeding the chance level of 0.25. This suggests that for roughly 74% of subjects, M4 was the most likely model to have generated their data. (b) Shows the confusion matrix representing model recovery accuracy. Each simulated model (x-axis) is correctly identified by the model comparison procedure as the best-fitting model (y-axis), with all values on the diagonal equal to 1 and off-diagonal values equal to 0. This indicates perfect recoverability and discriminability between the models, confirming that the model-fitting approach can reliably distinguish among the candidate models. (c) Displays the mean Leave-One-Out Cross-Validation (LOOcv) scores for each model averaged over 100 iterations, where lower values indicate better predictive performance. The simulation results demonstrate that when data were generated from a specific model (columns), the corresponding model generally achieved the lowest LOOcv score when fitted to that data, validating our model recovery procedure. Notably, M4 showed strong recovery performance.

Appendix 4.4: Parameter Recovery



Appendix Figure 4.4: Parameter Recovery. (a) Successful parameter recovery of the winning model in the Update Bias Task with high correlations between the simulated and estimated parameters. (b) The correlation between the parameters of the winning model. The weak correlations between most parameter pairs demonstrate that parameters do not systematically trade off against each other during estimation, supporting the model's identifiability.

Appendix 4.5: Mixed-Effects model controlling for potential confounds in the Update Bias Task

The results of a mixed-effects model extending the model in the paper revealed that even after controlling for potential confounds—including estimation error (EstErr) and subjective ratings—the main effects of Valence and Accuracy remained intact with no significant interaction between the two. The subjective ratings included Negativity ("How negative you found this event?" From 1 = Not at all to 6 = Very), Prior Experience ("Has this event happened to you before?" From 1 = never to 6 = very often), Vividness ("How vividly could you imagine this event?" From 1 = not vivid to 6 = very vivid), Familiarity ("Regardless of if this event has happened to you before, how familiar do you feel it is to you from TV, friends, movies and so on?" From 1 = not at all familiar to 6 very familiar); and Arousal ("When you imagine this event happening to you how emotionally arousing is the image in your mind?" From 1 = not arousing at all to 6 = very arousing).

Predictor	*Estimate*	*std. Error*	*df*	*CI*	*Statistic*	*p*
(Intercept)	0.09	0.02	94.43	0.04 – 0.13	3.46	**0.001**
Valence	0.24	0.02	98.86	0.19 – 0.28	10.17	**<0.001**
Accuracy	0.36	0.02	107.15	0.33 – 0.40	19.18	**<0.001**
EstErr	0.32	0.02	105.46	0.28 – 0.36	14.23	**<0.001**
Vividness	-0.01	0.02	179.54	-0.04 – 0.02	-0.48	0.628
Past Experience	-0.05	0.02	79.42	-0.08 – -0.02	-3.44	**0.001**
Familiarity	-0.01	0.02	207.52	-0.04 – 0.02	-0.93	0.351
Arousal	-0.00	0.02	216.79	-0.03 – 0.03	-0.26	0.795
Negativity	-0.02	0.01	100.59	-0.05 – 0.01	-1.29	0.197
Valence × Accuracy	0.02	0.02	105.85	-0.01 – 0.05	1.26	0.207
N $_{\text{Participant}}$		108				
Observations		5242				

Appendix Table 4.1: Mixed-Effects model controlling for potential confounds in the Update Bias Task

Appendix 4.6: The stimuli used in the Update Bias Task

List of the stimuli used in the study and their respective base rates. These events were split into two lists, with one randomly assigned to true trials and the other to false trials, where the base rates for the false trials were shuffled.

Event	BaseRate
Computer crash with loss of important data	68
Hospital stay longer than three weeks	58
Bicycle theft	54
Arteries hardening (narrowing of blood vessels)	45
Miss a flight	44
Victim of violence with need to go to A&E	34
Having a stroke	23
Lose Wallet	51
Household accident	58
Insect infestation (like ants) in your home	41

Victim of bullying at work (non physical)	46
Food poisoning	40
diabetes (type 2)	27
severe insomnia	21
Sports related accident	62
artificial joint	16
Domestic burglary	39
Depression	37
Divorce	50
Being cheated by husband/wife	52
death by infection	10
ulcer	13
death before 80	41
abnormal heart rhythm	29
skin burn	56
Cancer	30
Knee osteoarthritis (causing knee pain and swelling)	54
Being fired	62
Drug abuse	17
Blood clot in vein	14
Parkinson's disease	10
Bone fracture (break)	39
Victim of violence by a stranger	37
Victim of mugging	16
Severe teeth problems when old	31
Theft from vehicle	63
Hepatitis A or B	36
Theft from person	42
Eye cataract (clouding of the lens of the eye)	61
Back pain	70
Disease of spinal cord	24
Dementia	18
Having fleas/lice	42
Sexual dysfunction	37
More than 47 thousand dollars debt	48
Witness a traumatizing accident	40
Obesity	32
Irritable bowel syndrome (disorder of the gut)	30
Hernia (rupture of internal tissue wall)	43
Fraud when buying something on the internet	70

Appendix 4.7: Excluding Potentially Misclassified Trials in the Update Bias Task

To ensure my results were not influenced by potential trial misclassification, I conducted a supplementary analysis on a dataset that excluded the following trials:

- Trials where the BR was higher than E_1 but lower than the eBR.
- Trials where the BR was lower than E_1 but higher than the eBR.

Using this dataset, I conducted the 2*2 repeated measures ANOVA with Accuracy (True/False) and Valence (Good/Bad) as within-subject factors and the paired t-tests comparing good news vs bad news for true and false trials. The pattern of results remained consistent with those reported in the main paper, such that we observed a significant main effect of valence (F(1,107) = 318.76, p < 0.001) and accuracy (F(1,107) = 37.88, p < 0.001), without a significant interaction (F(1,107) = 1.38, p = 0.24). Paired t-tests confirmed that updating was greater for good news compared to bad news for both true (t(107) = 5.26, p < 0.001) and false trials (t(107) = 4.86, p < 0.001). Therefore, the results reported in the paper cannot be explained away by potentially misclassified trials.

Appendices for Chapter 5

Appendix 5.1: Additional fMRI Results

Contrast	Timepoint	Region	Peak MNI Coordinates [x y z]	Peak Statistic (z)	Cluster level p (FWE)	Cluster Size
Positive Correlation with DV	Choice	Cingulate Gyrus	12 -52 32	5.09	0.000	1150
		Cerebellum Left, Crus II	-36 -80 -40	4.63	0.01	394
		Lateral Occipital Cortex, Superior Division	-54 -66 34	4.54	0.02	358
		Cingulate Gyrus, Anterior	8 32 -4	4.20	0.04	280
		Frontal Pole	4 62 10	3.71	0.03	305
Negative Correlation with DV	Choice	Lateral Occipital Cortex, Inferior Division	38 -84 -10	5.47	0.000	3007
		Temporal Occipital Fusiform Cortex	-36 -50 -20	5.33	0.000	2291
		Middle Frontal Gyrus	-56 12 34	4.41	0.01	366
		Paracingulate Gyrus	-6 20 42	4.38	0.004	533
		Lateral Occipital Cortex, Superior Division	24 -58 54	4.16	0.001	644
		Lateral Occipital	-22 -60 54	4.09	0.003	554
Main Effect of Accuracy (True > False)	Information Cue	Lateral Occipital Cortex, Superior Division	-34 -70 -42	4.01	0.03	425
Main Effect of Accuracy (False > True)	Information Cue	NA	NA	NA	NA	NA
Main Effect of Feedback (Conf > Disconf)	Information Cue	NA	NA	NA	NA	NA
Main Effect of Feedback (Disconf > Conf)	Information Cue	NA	NA	NA	NA	NA
Interaction (Conf_True - Conf_False) > (Disconf_True - Disconf_False))	Information Cue	Left MTL (cluster including Hippocampus)	-36 -38 -2	4.51	0.04	361
		Right DLPFC	32 30 44	4.35	0.01	501
		Cerebellum Left	-32 -50 -44	3.96	0.04	366
Interaction (Disconf_True - Disconf_False) > (Conf_True - Conf_False))	Information Cue	NA	NA	NA	NA	NA

Main Effect of PE-modulated Accuracy (True > False)	Information Cue	NA	NA	NA	NA	NA
Main Effect of PE-modulated Accuracy (False > True)	Information Cue	NA	NA	NA	NA	NA
Main Effect of PE-modulated Feedback (Conf > Disconf)	Information Cue	NA	NA	NA	NA	NA
Main Effect of PE-modulated Feedback (Disconf > Conf)	Information Cue	NA	NA	NA	NA	NA
Interaction PE-modulated (positive)	Information Cue	NA	NA	NA	NA	NA
Interaction PE-modulated (negative)	Information Cue	NA	NA	NA	NA	NA

Appendix Table 5.1 Activation Table. The coordinates [x y z] are reported in Montreal Neurological Institute (MNI) space. The statistical threshold for significance was set at a voxel-level of p<0.05, Family-Wise Error (FWE) corrected at the cluster-level for multiple comparisons across the whole brain, with a minimum cluster size of [k = 5] voxels. The "Contrast" column describes the specific statistical comparison being tested at the designated "Timepoint" of the trial. For each significant cluster, the table lists its anatomical "Region," the MNI "Coordinates" of its peak voxel, the peak statistical "z-value," the FWE-corrected "p-value," and the "Cluster Size" in voxels. Other abbreviations include: DV (Difference in Value), PE (Prediction Error), Conf (confirmatory feedback), and Disconf (disconfirmatory feedback). NA indicates nothing was significant for the contrast at the threshold.

Appendix 5.2: Additional Computational Modelling Details

This appendix provides the alternative reinforcement learning models that were tested against the winning model presented in the main text. All models are based on the standard Rescorla-Wagner learning rule and use a SoftMax choice rule to generate choice probabilities. The models differ only in the number of learning rates.

Model 1

This model tests the hypothesis that learning is driven by the accuracy of the information, regardless of whether it confirms or disconfirms prior expectations. It uses two learning rates:

$Q(t+1) = Q(t) + \alpha_{true} * \delta(t)$ if information accuracy cue = True

$Q(t+1) = Q(t) + \alpha_{false} * \delta(t)$ if information accuracy cue = False

Free parameters (n=3): $\alpha_{true,}$ $\alpha_{false,}$ β

Model 2

This model explores the possibility that the nature of feedback (confirmatory vs. disconfirmatory) is important for learning when the information turns out false but not when it is true. It uses a single learning rate for true information, but splits the learning rate for false information into two:

$Q(t+1) = Q(t) + \alpha_{true} * \delta(t)$ if accuracy = True
$Q(t+1) = Q(t) + \alpha_{Conf, false} * \delta(t)$ if accuracy = False and feedback = Conf
$Q(t+1) = Q(t) + \alpha_{Disconf, false} * \delta(t)$ if accuracy = False and feedback = Disconf

Free parameters (n=4): α_{true}, $\alpha_{Conf, false}$, $\alpha_{Disconf, false}$, β

Model 3

This model tests the alternative hypothesis that the distinction between confirmatory and disconfirmatory feedback is relevant for learning when the information turns out true:

$Q(t+1) = Q(t) + \alpha_{false} * \delta(t)$ if accuracy = False
$Q(t+1) = Q(t) + \alpha_{Conf, true} * \delta(t)$ if accuracy = True and feedback = Conf
$Q(t+1) = Q(t) + \alpha_{Disconf, true} * \delta(t)$ if accuracy = True and feedback = Disconf

Free parameters (n=4): α_{false}, $\alpha_{Conf, true}$, $\alpha_{Disconf, true}$, β

Appendix 5.3: Additional Computational Modelling Results

In addition to the model comparison results reported in the main text, I conducted paired sample t-tests with FDR correction. This analysis reinforced Model 4's superiority over all other models, showing statistically significant advantages against Model 1 (t(31) = -3.80, p_adj < 0.001), Model 2 (t(31) = -2.80, p_adj < 0.01), and Model 3 (t(31) = -2.75, p_adj < 0.01). When looking at individual participants, Model 4 also performed the best. It had the lowest LOOcv score for 31.2% of participants, a higher proportion than Model 3 (25%), Model 2 (25%), and Model 1 (18.8%).

Appendix S: Reparameterization to conduct hierarchical t-test and interaction tests

For both tasks used in Chapters 2, 3, and 4, the reparameterization of the winning model is the same. Therefore, In the following I have used the Update Bias Task's model to demonstrate the approach.

I started reparametrizing Model 4 (M4) by calculating the sum (AS_T, AS_F) and difference ($AD_{T,}AD_F$) between good and bad news learning rates, separately for true and false trials:

(1) $AS_T = \alpha_{\text{true, goodnews}} + \alpha_{\text{true, badnews}}$

(2) $AD_T = \alpha_{\text{true, goodnews}} - \alpha_{\text{true, badnews}}$

(3) $AS_F = \alpha_{\text{false, goodnews}} + \alpha_{\text{false, badnews}}$

(4) $AD_F = \alpha_{\text{false, goodnews}} - \alpha_{\text{false, badnews}}$

I then added AS_T (from (1)) and (AD_T from (2)) together to get the following equality:

$$AS_T + AD_T = \left(\alpha_{\text{true, goodnews}} + \alpha_{\text{true, badnews}}\right) + \left(\alpha_{\text{true, goodnews}} - \alpha_{\text{true, badnews}}\right)$$

$$AS_T + AD_T = 2\alpha_{\text{true, goodnews}}$$

I then divide through both sides by 2 to get the following expression for $\alpha_{\text{true, goodnews}}$:

$$\alpha_{\text{true, goodnews}} = \frac{1}{2}(AS_T + AD_T)$$

I then subtracted AD_T (from (2) AS_T (from (1)) and again divided through by 2 to get a new expression for $\alpha_{\text{true, badnews}}$:

$$AS_T - AD_T = \left(\alpha_{\text{true, goodnews}} + \alpha_{\text{true, badnews}}\right) - \left(\alpha_{\text{true, goodnews}} - \alpha_{\text{true, badnews}}\right)$$
$$= 2\alpha_{\text{true, badnews}}$$

$$\alpha_{\text{true, badnews}} = \frac{1}{2}(AS_T - AD_T)$$

Then applied the same logic to the False learning rates, resulting in the following 4 expressions for the 4 learning rates:

$$\alpha_{\text{true, goodnews}} = \frac{1}{2}(AS_T + AD_T)$$

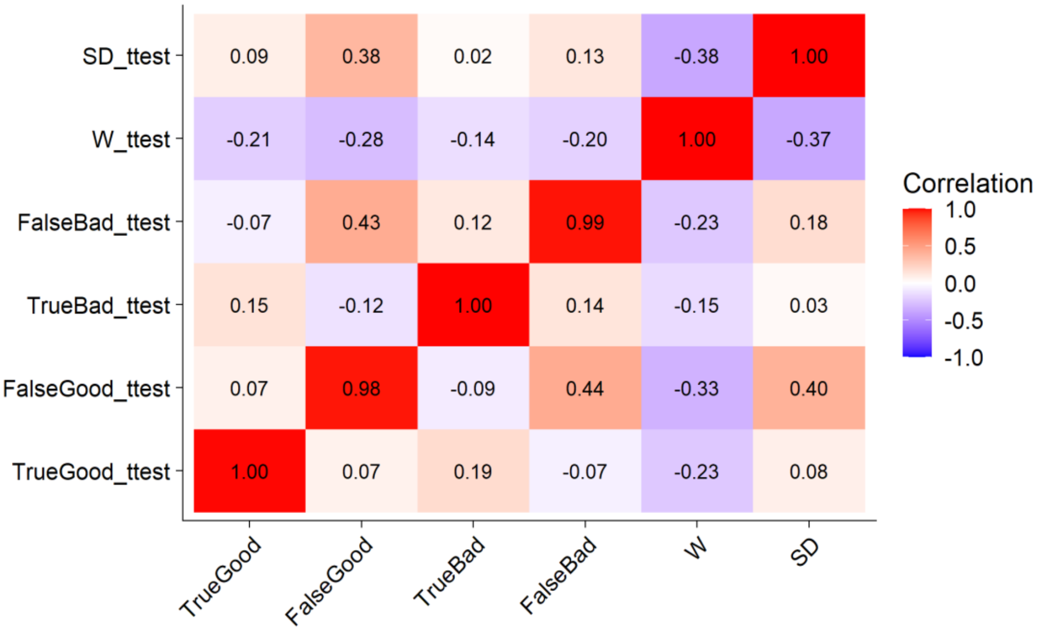
$$\alpha_{\text{true, badnews}} = \frac{1}{2}(AS_T - AD_T)$$

$$\alpha_{\text{false, goodnews}} = \frac{1}{2}(AS_F + AD_F)$$

$$\alpha_{\text{false, badnews}} = \frac{1}{2}(AS_F - AD_F)$$

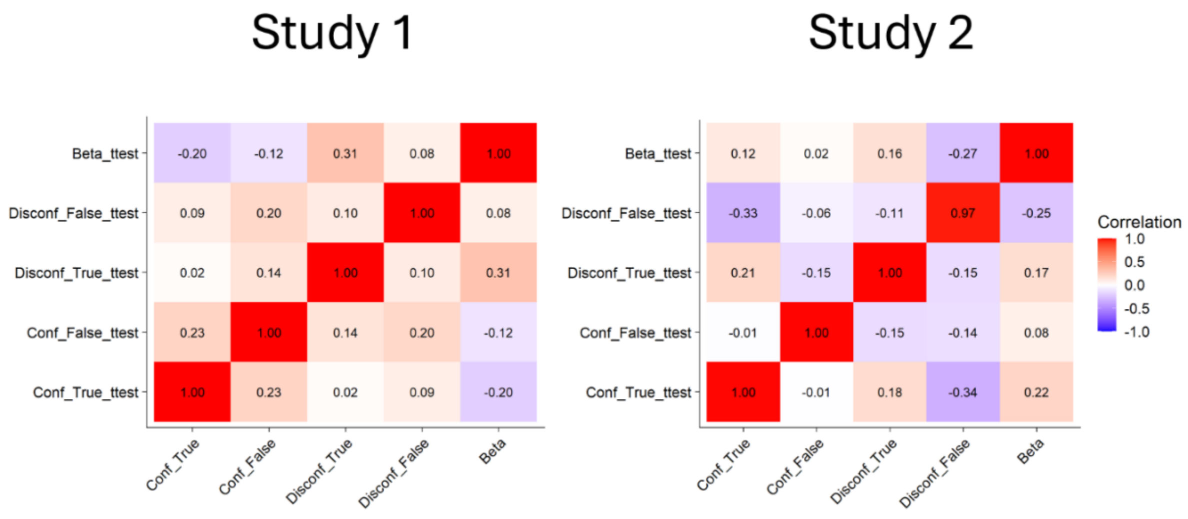
I use these 4 learning rates inside the model, but the free parameters are now: $AS_T, AD_T, AS_F, and\ AD_F$. Having the model configured in this way then allows me to conduct a hierarchical t-test (against 0) for AD_T which tests the difference between $\alpha_{\text{true, goodnews}}$ and $\alpha_{\text{true, badnews}}$ and a hierarchical t-test against 0 for AD_F which tests the difference between $\alpha_{\text{false, goodnews}}$ and $\alpha_{\text{false, badnews}}$.

I assessed the reliability of this approach by correlating the estimates from the original model with the estimates from the reparametrized model, which showed a very high correlation for all parameters:



Appendix Figure S.1: Correlation between the estimates from the original model with the estimates from the reparametrized model in the UBT.

The correlation between the original model and the reparametrized one in the reinforcement learning task used in Chapters 3 and 5 was also very high:



Study 1

Study 2

Appendix Figure S.2: Correlation between the estimates from the original model with the estimates from the reparametrized model in the RL task.

Next, I used the same approach and created a reparametrized model for the interaction test. To test whether there is an interaction in learning rates between Valence (Good News vs Bad News) and Accuracy (True vs. False), we expressed the interaction as a free parameter, defined as:

$$\text{Interaction} = \left(\alpha_{\text{false, goodnews}} - \alpha_{\text{false, badnews}}\right) - \left(\alpha_{\text{true, goodnews}} - \alpha_{\text{true, badnews}}\right)$$

Substitute the expressions:

$$\text{Interaction} = \left[\frac{1}{2}(AS_F + AD_F) - \frac{1}{2}(AS_F - AD_F)\right] - \left[\frac{1}{2}(AS_T + AD_T) - \frac{1}{2}(AS_T - AD_T)\right]$$

Simplify each term:

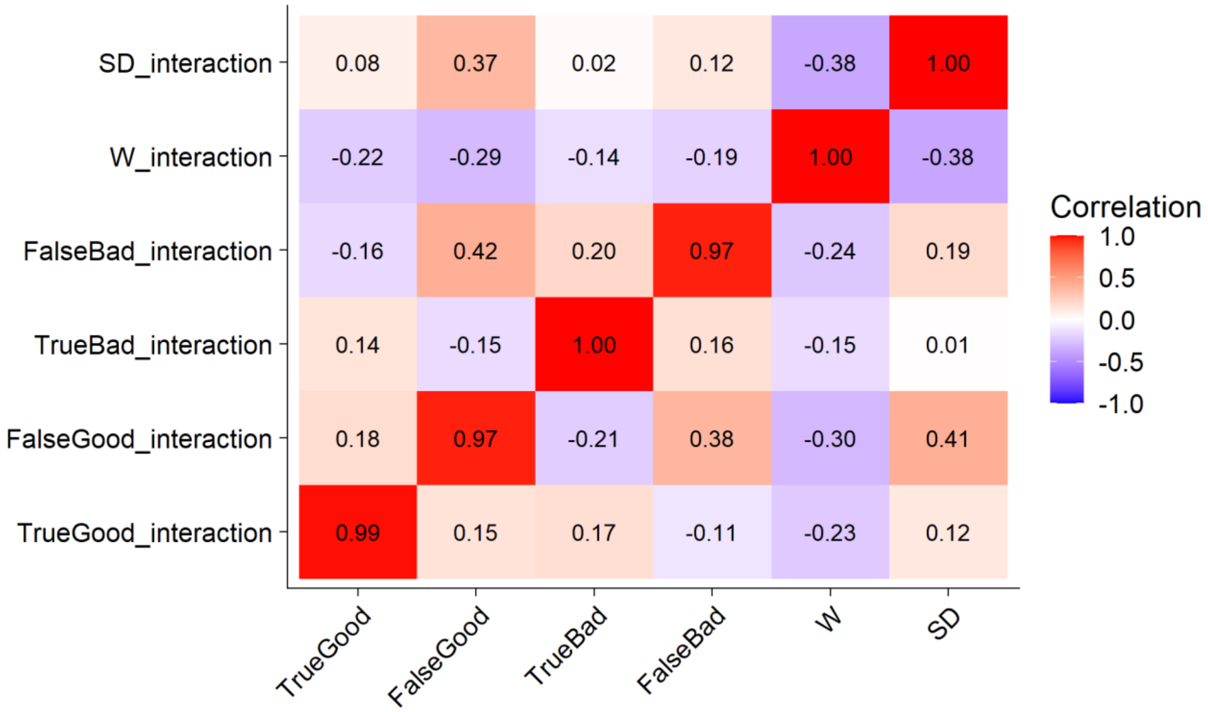
$$\text{Interaction} = \left[\frac{1}{2}AS_F + \frac{1}{2}AD_F - \frac{1}{2}AS_F + \frac{1}{2}AD_F\right] - \left[\frac{1}{2}AS_T + \frac{1}{2}AD_T - \frac{1}{2}AS_T + \frac{1}{2}AD_T\right]$$

Thus, the interaction simplifies to:

$$\text{Interaction} = AD_F - AD_T$$

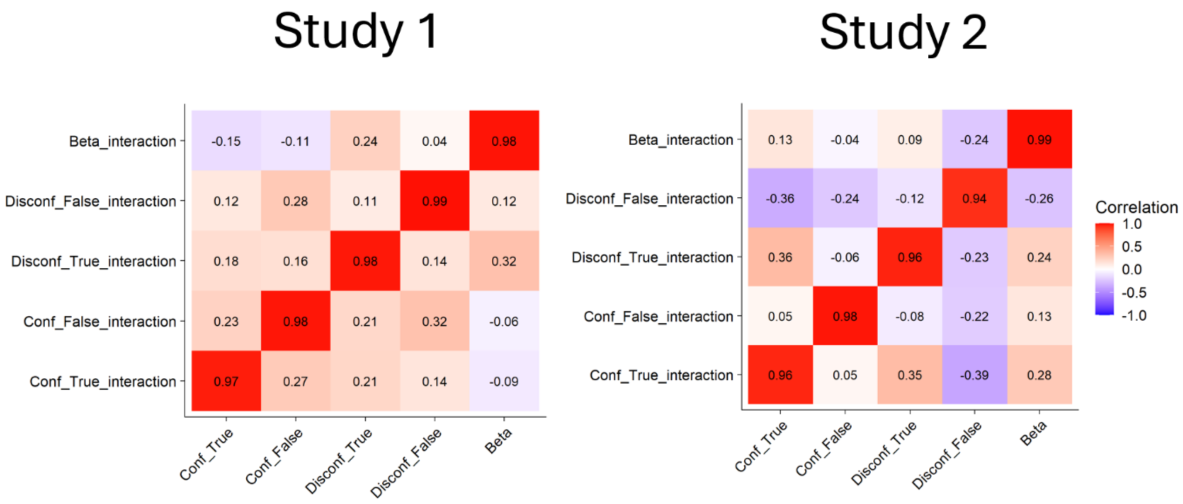
A hierarchical t-test against 0 for Interaction tells us whether there is an interaction between valence and accuracy or not.

The very high correlation between the reparametrized model and the original model confirmed the reliability of this approach:



Appendix Figure S.3: Correlation between the estimates from the original model with the estimates from the reparametrized model in the UBT.

And similarly in the RL task:



Appendix Figure S.4: Correlation between the estimates from the original model with the estimates from the reparametrized model in the RL task.

Glossary

Abbreviation	Details
AI	Artificial Intelligence
ANOVA	Analysis of Variance
BDI	Beck Depression Inventory
BOLD	Blood-Oxygen-Level Dependent
DLPFC	Dorsolateral Prefrontal Cortex
DCM	Dynamic Causal Modelling
EEG	Electroencephalography
fMRI	Functional Magnetic Resonance Imaging
GLM	General Linear Model
IFG	Inferior Frontal Gyrus
LLM	Large Language Model
LMM	Linear Mixed-effects Model
LOOcv	Leave-One-Out Cross-Validation
MTL	Medial Temporal Lobe
RL	Reinforcement Learning
ROI	Region of Interest
RT	Reaction Time
TR	Repetition Time
SD	Standard Deviation
SEM	Standard Error of the Mean
SPM	Statistical Parametric Mapping
UBT	Update Bias Task
vmPFC	Ventromedial Prefrontal Cortex

References

Ahn, W.-Y., Krawitz, A., Kim, W., Busmeyer, J. R., & Brown, J. W. (2011). A Model-Based fMRI Analysis with Hierarchical Bayesian Parameter Estimation. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 95–110. https://doi.org/10.1037/a0020684

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, J., Pennycook, G., & Rand, D. G. (2025). Addressing misperceptions takes more than combating fake news. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.07.002

Apuke, O. D., Omar, B., Tunca, E. A., & Gever, C. V. (2023). The effect of visual multimedia instructions against fake news spread: A quasi-experimental study with Nigerian students. *Journal of Librarianship and Information Science*, *55*(3), 694–703. https://doi.org/10.1177/09610006221096477

Aral, S., & Van Alstyne, M. W. (2010). *The Diversity-Bandwidth Tradeoff* (SSRN Scholarly Paper No. 958158). Social Science Research Network. https://doi.org/10.2139/ssrn.958158

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology. General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

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), 1214–1221. https://doi.org/10.1038/nn1954

Bergerot, C., Barfuss, W., & Romanczuk, P. (2024). Moderate confirmation bias enhances decision-making in groups of reinforcement-learning agents. *PLOS*

Computational Biology, *20*(9), e1012404. https://doi.org/10.1371/journal.pcbi.1012404

Berinsky, A. J. (2017). Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science*, *47*(2), 241–262. Scopus. https://doi.org/10.1017/S0007123415000186

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*(6), 787–793. https://doi.org/10.1038/s41593-018-0152-y

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). *Julia: A Fast Dynamic Language for Technical Computing* (No. arXiv:1209.5145). arXiv. https://doi.org/10.48550/arXiv.1209.5145

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. https://doi.org/10.1016/j.obhdp.2006.07.001

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Borges do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization*, *100*(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *11*(5), 485–494. https://doi.org/10.1101/lm.78804

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology. General*, *149*(4), 746–756. https://doi.org/10.1037/xge0000673

Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054. https://doi.org/10.1016/j.cognition.2019.104054

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (n.d.). *Region of interest analysis using an SPM toolbox*.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. https://doi.org/10.1016/j.neuron.2009.06.009

Bromberg-Martin, E. S., & Monosov, I. E. (2020). Neural circuitry of information seeking. *Current Opinion in Behavioral Sciences*, *35*, 62–70. https://doi.org/10.1016/j.cobeha.2020.07.006

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings. Biological Sciences*, *282*(1810), 20150228. https://doi.org/10.1098/rspb.2015.0228

Browning, M., Behrens, T. E., Jocham, G., O’Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Brulle, R. J., & Roberts, J. T. (2017). Climate Misinformation Campaigns and Public Sociology. *Contexts*, *16*(1), 78–79. https://doi.org/10.1177/1536504217696081

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45–53. https://doi.org/10.1038/s41586-024-07417-w

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carnahan, D., Bergan, D. E., & Lee, S. (2021). Do Corrective Effects Last? Results from a Longitudinal Experiment on Beliefs Toward Immigration in the U.S. *Political Behavior*, *43*(3), 1227–1246. https://doi.org/10.1007/s11109-020-09591-9

Cazé, R. D., & van der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*(6), 711–719. https://doi.org/10.1007/s00422-013-0571-5

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), e2216614120. https://doi.org/10.1073/pnas.2216614120

Chan, M. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour*, *7*(9), 1514–1525. https://doi.org/10.1038/s41562-023-01623-8

Chan, M.-P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chang, E. C. (2001). *Optimism & pessimism: Implications for theory, research, and practice*. American Psychological Association.

Chen, T., Huang, J., Cui, J., Li, Z., Wang, Y., Irish, M., & Chan, R. C. K. (2022). Functional Coupling between the Fronto-Parietal Network and Default Mode Network Is Associated with Balanced Time Perspective. *Brain Sciences*, *12*(9), 1201. https://doi.org/10.3390/brainsci12091201

Cheng, C. X. (2018). Confirmation Bias in Investments. *International Journal of Economics and Finance*, *11*(2), 50. https://doi.org/10.5539/ijef.v11n2p50

Cheng, Z., Moser, A. D., Jones, M., & Kaiser, R. H. (2024). Reinforcement learning and working memory in mood disorders: A computational analysis in a developmental transdiagnostic sample. *Journal of Affective Disorders*, *344*, 423–431. https://doi.org/10.1016/j.jad.2023.10.084

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *29*(39), 12315–12320. https://doi.org/10.1523/JNEUROSCI.2575-09.2009

Chierchia, G., Soukupová, M., Kilford, E. J., Griffin, C., Leung, J., Palminteri, S., & Blakemore, S.-J. (2023). Confirmatory reinforcement learning changes with age during adolescence. *Developmental Science*, *26*(3), e13330. https://doi.org/10.1111/desc.13330

Christner, C., Merz ,Pascal, Barkela ,Berend, Jungkunst ,Hermann, & and von Sikorski, C. (2024). Combatting Climate Disinformation: Comparing the Effectiveness of Correction Placement and Type. *Environmental Communication*, *18*(6), 729–742. https://doi.org/10.1080/17524032.2024.2316757

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2023a). *The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter* (No. arXiv:2307.07960). arXiv. https://doi.org/10.48550/arXiv.2307.07960

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2023b). *The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter*.

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working Memory Load Strengthens Reward Prediction Errors. *Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Compton, J. (2013). Inoculation theory. In *The SAGE handbook of persuasion: Developments in theory and practice, 2nd ed* (pp. 220–236). Sage Publications, Inc.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science (New York, N.Y.)*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814 da Silva Pinho, A., Céspedes Izquierdo, V., Lindström, B., & van den Bos, W. (2024). Youths' sensitivity to social media feedback: A computational account. *Science Advances*, *10*(43), eadp8775. https://doi.org/10.1126/sciadv.adp8775

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLOS Computational Biology*, *10*(1), e1003441. https://doi.org/10.1371/journal.pcbi.1003441

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Daw, N. D., & Tobler, P. N. (2013). Value learning through reinforcement: The basics of dopamine and reinforcement learning. In P. W. Glimcher & E. Fehr (Eds), *Neuroeconomics (2nd edition)* (pp. 283–298). Academic Press.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *14*(2), 238–257. https://doi.org/10.1177/1088868309352251

Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *eLife*, *8*, e43499. https://doi.org/10.7554/eLife.43499

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014a). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLOS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014b). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLOS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Donchin, E., & Coles, M. G. H. (1988). Is the P300 Component a Manifestation of Context Updating? *Behavioral and Brain Sciences*, *11*(3), 357–374. https://doi.org/10.1017/s0140525x00058027

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The Psychology of Conspiracy Theories. *Current Directions in Psychological Science*, *26*(6), 538–542. https://doi.org/10.1177/0963721417718261

Drerup, C., Garcia-Pelegrin, E., Wilkins, C., Herbert-Read, J. E., & Clayton, N. S. (2025). Tactical deception in cephalopods: A new framework for understanding cognition. *Trends in Ecology & Evolution*, *40*(8), 740–748. https://doi.org/10.1016/j.tree.2025.04.016

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Eil, D., & Rao, J. M. (2011a). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. https://doi.org/10.1257/mic.3.2.114

Eil, D., & Rao, J. M. (2011b). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. https://doi.org/10.1257/mic.3.2.114

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24. https://doi.org/10.1016/j.tics.2015.07.010

Fallon, S. J., Smulders, K., Esselink, R. A., van de Warrenburg, B. P., Bloem, B. R., & Cools, R. (2015). Differential optimal dopamine levels for set-shifting and working memory in Parkinson's disease. *Neuropsychologia*, *77*, 42–51. https://doi.org/10.1016/j.neuropsychologia.2015.07.031

Farashahi, S., Donahue, C. H., Hayden, B. Y., Lee, D., & Soltani, A. (2019). Flexible combination of reward information across primates. *Nature Human Behaviour*, *3*(11), 1215–1224. https://doi.org/10.1038/s41562-019-0714-3

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior* (pp. xxii, 461). Cambridge University Press. https://doi.org/10.1017/9781316272503

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Filipowicz, A. L., Glaze, C. M., Kable, J. W., & Gold, J. I. (2020). Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *eLife*, *9*, e57872. https://doi.org/10.7554/eLife.57872

Flaxman, S., Goel, S., & Rao, J. M. (2016). *Filter Bubbles, Echo Chambers, and Online News Consumption* (SSRN Scholarly Paper No. 2363701). Social Science Research Network. https://doi.org/10.2139/ssrn.2363701

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, *75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624. https://doi.org/10.1038/s41593-018-0104-6

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. https://doi.org/10.1002/hbm.24047

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302. https://doi.org/10.1016/s1053-8119(03)00202-7

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234. https://doi.org/10.1016/j.neuroimage.2006.08.035

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, *53*(1), 139–152. https://doi.org/10.3758/s13428-020-01408-1

Fu, Q., Moerbeek, M., & Hoijtink, H. (2022). Sample size determination for Bayesian ANOVAs with informative hypotheses. *Frontiers in Psychology*, *13*, 947768. https://doi.org/10.3389/fpsyg.2022.947768

Gagne, C., Zika, O., Dayan, P., & Bishop, S. J. (2020). Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife*, *9*, e61387. https://doi.org/10.7554/eLife.61387

Garrett, N., & Daw, N. D. (2020). Biased belief updating and suboptimal choice in foraging decisions. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-16964-5

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating Beliefs under Perceived Threat. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Garrett, N., & Sharot, T. (2014). How Robust Is the Optimistic Update Bias for Estimating Self-Risk and Population Base Rates? *PLoS ONE*, *9*(6), e98848. https://doi.org/10.1371/journal.pone.0098848

Garrett, N., & Sharot, T. (2017a). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*, *50*, 12–22. https://doi.org/10.1016/j.concog.2016.10.013

Garrett, N., & Sharot, T. (2017b). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition: An International Journal*, *50*, 12–22. https://doi.org/10.1016/j.concog.2016.10.013

Garrett, N., & Sharot, T. (2023). There is no belief update bias for neutral events: Failure to replicate Burton et al. (2022). *Journal of Cognitive Psychology (Hove, England)*, *35*(8), 876–886. https://doi.org/10.1080/20445911.2023.2245112

Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience*, *8*(AUG). Scopus. https://doi.org/10.3389/fnhum.2014.00639

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(2), 483–495. https://doi.org/10.1093/cercor/bhn098

Glickman, M., & Sharot, T. (2024). AI-induced hyper-learning in humans. *Current Opinion in Psychology*, *60*, 101900. https://doi.org/10.1016/j.copsyc.2024.101900

Globig, L. K., & Sharot, T. (2024). Considering information-sharing motives to reduce misinformation. *Current Opinion in Psychology*, *59*, 101852. https://doi.org/10.1016/j.copsyc.2024.101852

Glockner, M., Hou, Y., & Gurevych, I. (2022). Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5916–5936). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.397

Glover, G. H. (2011). Overview of Functional Magnetic Resonance Imaging. *Neurosurgery Clinics of North America*, *22*(2), 133–139. https://doi.org/10.1016/j.nec.2010.11.001

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS ONE*, *17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, *62*(1), 154–166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Haber, S. N. (2011). *Neural Circuits of Reward and Decision Making: Integrative Networks across Corticobasal Ganglia Loops*. https://doi.org/10.7551/mitpress/8791.003.0004

Han, J., Cha, M., & Lee, W. (2020). Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-39

Hanley, H. W. A., & Durumeric, Z. (2024). *Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites* (No. arXiv:2305.09820). arXiv. https://doi.org/10.48550/arXiv.2305.09820

Hare, T. A., O’Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *The Journal of Neuroscience*, *28*(22), 5623–5630. https://doi.org/10.1523/JNEUROSCI.1309-08.2008

Harris, C., Aguirre, C., Kolli, S., Das, K., Izquierdo, A., & Soltani, A. (2021). *Unique features of stimulus-based probabilistic reversal learning* (p. 2020.09.24.310771). bioRxiv. https://doi.org/10.1101/2020.09.24.310771

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. https://doi.org/10.1037/a0015701

Haselton, M. G., & Nettle, D. (2006). The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review*, *10*(1), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology* (pp. 724–746). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470939376.ch25

Hayes, W. M., Yax, N., & Palminteri, S. (2025). Relative Value Encoding in Large Language Models: A Multi-Task, Multi-Model Investigation. *Open Mind*, *9*, 709–725. https://doi.org/10.1162/opmi_a_00209

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In *An introduction to model-based cognitive neuroscience* (pp. 25–48). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-2236-9

Hofmans, L., & van den Bos, W. (2025). Neural correlates of Bayesian social belief updating in the medial prefrontal cortex. *Cerebral Cortex (New York, NY)*, *35*(8), bhaf251. https://doi.org/10.1093/cercor/bhaf251

Holroyd, C., & Coles, M. (2002). The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity. *Psychological Review*, *109*, 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Hoxha, I., Sperber, L., & Palminteri, S. (2024). *Evolving choice hysteresis in reinforcement learning: Comparing the adaptive value of positivity bias and gradual perseveration*. OSF. https://doi.org/10.31234/osf.io/zprxe

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, *15*(3), 470–476, S1-3. https://doi.org/10.1038/nn.3017

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, *7*(4), e1002028. https://doi.org/10.1371/journal.pcbi.1002028

Isaac, M., & Schleifer, T. (2025, January 7). Meta Says It Will End Its Fact-Checking Program on Social Media Posts. *The New York Times*. https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking

Iyengar, S., & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, *59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Jocham, G., Furlong, P. M., Kröger, I. L., Kahn, M. C., Hunt, L. T., & Behrens, T. E. J. (2014). Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. *Neuroimage*, *100*, 498–506. https://doi.org/10.1016/j.neuroimage.2014.06.005

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Johnson-Laird, P. N. (2012). Mental models and consistency. In *Cognitive consistency: A fundamental principle in social cognition* (pp. 225–244). The Guilford Press.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, *65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. https://doi.org/10.1038/nn2007

Kandroodi, M. R., Vahabie, A.-H., Ahmadi, S., Araabi, B. N., & Ahmadabadi, M. N. (2021). *Optimal Reinforcement Learning with Asymmetric Updating in Volatile Environments: A Simulation Study* (p. 2021.02.15.431283). bioRxiv. https://doi.org/10.1101/2021.02.15.431283

Kao, C.-H., Feng, G. W., Hur, J. K., Jarvis, H., & Rutledge, R. B. (2023). Computational models of subjective feelings in psychiatry. *Neuroscience and Biobehavioral Reviews*, *145*, 105008. https://doi.org/10.1016/j.neubiorev.2022.105008

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, *23*(5), 1301–1326. https://doi.org/10.1177/1461444820959296

Kappes, A., Faber, N. S., Kahane, G., Savulescu, J., & Crockett, M. J. (2018). Concern for Others Leads to Vicarious Optimism. *Psychological Science*, *29*(3), 379–389. https://doi.org/10.1177/0956797617737129

Karlsen, R., Beyer, A., & Steen-Johnsen, K. (2020). Do High-Choice Media Environments Facilitate News Avoidance? A Longitudinal Study 1997–2016. *Journal of Broadcasting & Electronic Media*, *64*(5), 794–814. https://doi.org/10.1080/08838151.2020.1835428

Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, *87*, 31–45. https://doi.org/10.1016/j.jmp.2018.09.002

Kawabata, H., & Zeki, S. (2004). Neural Correlates of Beauty. *Journal of Neurophysiology*, *91*(4), 1699–1705. https://doi.org/10.1152/jn.00696.2003

Khalid, I., Morlaas, O., Bottemanne, H., Thonon, L., Costa, T. D., Fossati, P., & Schmidt, L. (2024). Effects of experiencing the COVID-19 pandemic on optimistically biased belief updating. *eLife*, *13*. https://doi.org/10.7554/eLife.101157.1

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037/0033-295X.94.2.211

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*, 557–577. https://doi.org/10.1017/S0048577201990559

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. https://doi.org/10.1038/nature02169

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587. Scopus. https://doi.org/10.1017/S003329171500207X

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012a). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. Scopus. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012b). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. Scopus. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Kube, T., & Rozenkrantz, L. (2021). When Beliefs Face Reality: An Integrative Review of Belief Updating in Mental Health and Illness. *Perspectives on Psychological Science*, *16*(2), 247–274. https://doi.org/10.1177/1745691620931496

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. *Frontiers in Psychology*, *8*, 1087. https://doi.org/10.3389/fpsyg.2017.01087

Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. *Journal of Neuroscience*, *38*(37), 7996–8010. https://doi.org/10.1523/JNEUROSCI.0266-18.2018

Kuzmanovic, B., Rigoux, L., & Vogeley, K. (2019a). Brief Report: Reduced Optimism Bias in Self-Referential Belief Updating in High-Functioning Autism. *Journal of Autism and Developmental Disorders*, *49*(7), 2990–2998. https://doi.org/10.1007/s10803-016-2940-0

Kuzmanovic, B., Rigoux, L., & Vogeley, K. (2019b). Brief Report: Reduced Optimism Bias in Self-Referential Belief Updating in High-Functioning Autism. *Journal of Autism* *and Developmental Disorders*, *49*(7), 2990–2998. Scopus. https://doi.org/10.1007/s10803-016-2940-0

Kuzmanovic, B., Rigoux, L., & Vogeley, K. (2019c). Brief Report: Reduced Optimism Bias in Self-Referential Belief Updating in High-Functioning Autism. *Journal of Autism and Developmental Disorders*, *49*(7), 2990–2998. https://doi.org/10.1007/s10803-016-2940-0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., Reddy, C. B., Kepecs, A., Harris, K. D., & Carandini, M. (2020). Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and Reward Value. *Neuron*, *105*(4), 700-711.e6. https://doi.org/10.1016/j.neuron.2019.11.018

Lancet, T. E. of T. (2010). Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *375*(9713), 445. https://doi.org/10.1016/S0140-6736(10)60175-4

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. https://doi.org/10.1016/j.neuron.2009.09.040

Lee, N. M. (2018). Fake news, phishing, and fraud: A call for research on digital media literacy education beyond the classroom. *Communication Education*, *67*(4), 460–466. https://doi.org/10.1080/03634523.2018.1503313

Leeuw, J. R. de, Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. https://doi.org/10.21105/joss.05351

Lefebvre, G., Deroy, O., & Bahrami, B. (2024). The roots of polarization in the individual reward system. *Proceedings. Biological Sciences*, *291*(2017), 20232011. https://doi.org/10.1098/rspb.2023.2011

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9. https://doi.org/10.1038/s41562-017-0067

Lefebvre, G., Summerfield, C., & Bogacz, R. (2022). A Normative Account of Confirmation Bias During Reinforcement Learning. *Neural Computation*, *34*(2), 307–337. https://doi.org/10.1162/neco_a_01455

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

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. https://doi.org/10.1177/1529100612451018

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: A model and a method. *Journal of Experimental Psychology. Human Perception and Performance*, *10*(2), 276–291. https://doi.org/10.1037//0096-1523.10.2.276

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), Article 3. https://doi.org/10.1038/s41562-021-01056-1

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, *58*, 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

Ma, Y., Li, S., Wang, C., Liu, Y., Li, W., Yan, X., Chen, Q., & Han, S. (2016). Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback. *Proceedings of the National Academy of Sciences*, *113*(33), 9256–9261. https://doi.org/10.1073/pnas.1604285113

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian

Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Martín-Luengo, B., Zinchenko, O., Dolgoarshinnaia, A., & Leminen, A. (2021). Retrospective confidence judgments: Meta-analysis of functional magnetic resonance imaging studies. *Human Brain Mapping*, *42*(10), 3005–3022. https://doi.org/10.1002/hbm.25397

Maslow, A. H. (1950). Self-actualizing people: A study of psychological health. *Personality*, *Symposium 1*, 11–34.

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*(2), 293–296. https://doi.org/10.1016/j.schres.2005.06.020

McClung Lee, A. (1949). LAZARSFELD, PAUL F., BERNARD BERELSON, and HAZEL GAUDET. The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign. (Second edition.) Pp. xxxiii, 178. New York: Columbia University Press, 1948. $2.75. *The ANNALS of the American Academy of Political and Social Science*, *261*(1), 194–194. https://doi.org/10.1177/000271624926100137

McCoy, B., Jahfari, S., Engels, G., Knapen, T., & Theeuwes, J. (2019). Dopaminergic medication reduces striatal sensitivity to negative outcomes in Parkinson's disease. *Brain*, *142*(11), 3605–3620. https://doi.org/10.1093/brain/awz276

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991–996. https://doi.org/10.1126/science.adl2829

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia output and cognition: Evidence from anatomical, behavioral, and clinical studies. *Brain and Cognition*, *42*(2), 183–200. https://doi.org/10.1006/brcg.1999.1099

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(Volume 24, 2001), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mitchell, R. (1986). Deception, perspectives on human and nonhuman deceit. *State University of New York Press eBooks*. https://www.academia.edu/120413591/Deception_perspectives_on_human_and_nonhuman_deceit

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Morris, B. E., Carl. (1977, May 1). *Stein's Paradox in Statistics*. Scientific American. https://www.scientificamerican.com/article/steins-paradox-in-statistics/

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal–Subcortical Circuit and Asymmetric Belief Updating. *The Journal of Neuroscience*, *35*(42), 14077–14085. https://doi.org/10.1523/JNEUROSCI.1120-15.2015

Moutsiana, C., Garrett, N., Clarke, R. C., Lotto, R. B., Blakemore, S.-J., & Sharot, T. (2013). Human development of the ability to learn from bad news. *Proceedings* *of the National Academy of Sciences*, *110*(41), 16396–16401. https://doi.org/10.1073/pnas.1305631110

Nadler, J., Baumgartner, S., & Washington, M. (2021). MTurk for working samples: Evaluation of data quality 2014 – 2020. *North American Journal of Psychology*, *23*(4), 741–752.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil–linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M., & Frank, M. J. (2021). All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain*, *144*(3), 1013–1029. https://doi.org/10.1093/brain/awaa453

Neumann, J. V., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nussenbaum, K., Kahn, A., Zhang, A., Daw, N., & Hartley, C. (2025). *Shifts in learning dynamics drive developmental improvements in the acquisition of structured knowledge*. OSF. https://doi.org/10.31234/osf.io/amvth_v1

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

O’Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53. https://doi.org/10.1196/annals.1390.022

O’Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, *41*(2), 147–155. https://doi.org/10.1016/S0028-3932(02)00145-8

Oganian, Y., Heekeren, H. R., & Korn, C. W. (2019a). Low foreign language proficiency reduces optimism about the personal future. *Quarterly Journal of Experimental Psychology*, *72*(1), 60–75. https://doi.org/10.1177/1747021818774789

Oganian, Y., Heekeren, H. R., & Korn, C. W. (2019b). Low foreign language proficiency reduces optimism about the personal future. *Quarterly Journal of Experimental Psychology*, *72*(1), 60–75. https://doi.org/10.1177/1747021818774789

Ohta, H., Satori, K., Takarada, Y., Arake, M., Ishizuka, T., Morimoto, Y., & Takahashi, T. (2021). The asymmetric learning rates of murine exploratory behavior in sparse reward environments. *Neural Networks*, *143*, 218–229. https://doi.org/10.1016/j.neunet.2021.05.030

Ossola, P., Garrett, N., Sharot, T., & Marchesi, C. (2020). Belief updating in bipolar disorder predicts time of recurrence. *eLife*, *9*, e58891. https://doi.org/10.7554/eLife.58891

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022).

Training language models to follow instructions with human feedback (No. arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in Orbitofrontal Cortex Encode Economic Value. *Nature*, *441*(7090), 223–226. https://doi.org/10.1038/nature04676

Palminteri, S. (2023). Choice-confirmation bias and gradual perseveration in human reinforcement learning. *Behavioral Neuroscience*, *137*(1), 78–88. https://doi.org/10.1037/bne0000541

Palminteri, S. (2025a). Human reinforcement learning processes and biases: Computational characterization and possible applications to behavioral public policy. *Mind & Society*. https://doi.org/10.1007/s11299-025-00329-w

Palminteri, S. (2025b). Human reinforcement learning processes and biases: Computational characterization and possible applications to behavioral public policy. *Mind & Society*. https://doi.org/10.1007/s11299-025-00329-w

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621. https://doi.org/10.1016/j.tics.2022.04.005

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Park, J., Konana, P., Gu, B., Kumar, A., & Raghunathan, R. (2010). *Confirmation Bias, Overconfidence, and Investment Performance: Evidence from Stock Message Boards* (SSRN Scholarly Paper No. 1639470). https://doi.org/10.2139/ssrn.1639470

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology. General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-30073-5

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations* (No. arXiv:2212.09251). arXiv. https://doi.org/10.48550/arXiv.2212.09251

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. https://doi.org/10.1037/h0049234

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*, *12*(1), 5966. https://doi.org/10.1038/s41598-022-10070-w

Pike, A. C., & Robinson, O. J. (2022). Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *79*(4), 313–322. https://doi.org/10.1001/jamapsychiatry.2022.0051

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/s1364-6613(02)01964-2

Pittman, M., & Haley, E. (2023). Cognitive Load and Social Media Advertising. *Journal of Interactive Advertising*. https://www.tandfonline.com/doi/abs/10.1080/15252019.2022.2144780

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, *27*(37), 9984–9988. https://doi.org/10.1523/JNEUROSCI.2131-07.2007

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, *2*(1), 67–70. https://doi.org/10.1093/scan/nsm006

Polich, J. (2007). Updating P300: An Integrative Theory of P3a and P3b. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), e2104235118. https://doi.org/10.1073/pnas.2104235118

Pulcu, E., & Browning, M. (2019). The Misestimation of Uncertainty in Affective Disorders. *Trends in Cognitive Sciences*, *23*(10), 865–875. https://doi.org/10.1016/j.tics.2019.07.007

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Rapoza, K. (n.d.). *Can 'Fake News' Impact The Stock Market?* Forbes. Retrieved 5 August 2025, from https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/

Rathje, S., & Van Bavel, J. J. (2025). The psychology of virality. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.06.014

Rathje, S., Ye, M., Globig, L., Pillai, R., de Mello, V., & Van Bavel, J. (2025). *Sycophantic AI increases attitude extremity and overconfidence*. OSF. https://doi.org/10.31234/osf.io/vmyek_v1

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory: Vol. Vol. 2*.

Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1822), 20200131. https://doi.org/10.1098/rstb.2020.0131

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 2634. https://doi.org/10.1038/s41467-020-16278-6

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), 201199. https://doi.org/10.1098/rsos.201199

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65. https://doi.org/10.1057/s41599-019-0279-9

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*(5), 880–892. https://doi.org/10.1037/0022-3514.32.5.880

Ruohonen, J. (2024). A comparative study of online disinformation and offline protests. *SN Social Sciences*, *4*(12), 232. https://doi.org/10.1007/s43545-024-01029-x

Russell, P. N. S. J. (2015). *Artificial Intelligence: A Modern Approach, 3Rd Edition*. PEARSON INDIA.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose Opinions Do Language Models Reflect?* (No. arXiv:2303.17548). arXiv. https://doi.org/10.48550/arXiv.2303.17548

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062

Sazhin, D., Dachs, A., & Smith, D. V. (2025). Meta-Analysis Reveals That Explore-Exploit Decisions are Dissociable by Activation in the Dorsal Lateral Prefrontal Cortex, Anterior Insula, and the Dorsal Anterior Cingulate Cortex. *bioRxiv: The Preprint Server for Biology*, 2023.10.21.563317. https://doi.org/10.1101/2023.10.21.563317

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice.

Psychonomic Bulletin & Review, *22*(2), 391–407. https://doi.org/10.3758/s13423-014-0684-4

Schmid-Petri, H., & Bürger, M. (2022). The effect of misinformation and inoculation: Replication of an experiment on the effect of false experts in the context of climate change communication. *Public Understanding of Science (Bristol, England)*, *31*(2), 152–167. https://doi.org/10.1177/09636625211024550

Schubert, J. A., Jagadish, A. K., Binz, M., & Schulz, E. (2024). In-Context Learning Agents Are Asymmetric Belief Updaters. *Proceedings of the 41st International Conference on Machine Learning*, 43928–43946. https://proceedings.mlr.press/v235/schubert24a.html

Schüller, T., Fischer, A. G., Gruendler, T. O. J., Baldermann, J. C., Huys, D., Ullsperger, M., & Kuhn, J. (2020). Decreased transfer of value to action in Tourette syndrome. *Cortex*, *126*, 39–48. https://doi.org/10.1016/j.cortex.2019.12.027

Schulreich, S., & Schwabe, L. (2021). Causal Role of the Dorsolateral Prefrontal Cortex in Belief Updating under Uncertainty. *Cerebral Cortex (New York, N.Y.: 1991)*, *31*(1), 184–200. https://doi.org/10.1093/cercor/bhaa219

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns. In *Advances in Experimental Social Psychology* (Vol. 39, pp. 127–161). Academic Press. https://doi.org/10.1016/S0065-2601(06)39003-X

Sears, D. O., & Freedman, J. L. (1967). Selective Exposure to Information: A Critical Review. *The Public Opinion Quarterly*, *31*(2), 194–213.

Šekrst, K. (2022). Everybody Lies: Deception Levels in Various Domains of Life. *Biosemiotics*, *15*(2), 309–324. https://doi.org/10.1007/s12304-022-09485-9

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Sharot, T., & Garrett, N. (2022). A guideline and cautionary Note: How to use the belief update task correctly. *Methods in Psychology*, *6*, 100091. https://doi.org/10.1016/j.metip.2022.100091

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), Article 11. https://doi.org/10.1038/nn.2949

Sharot, T., Rollwage, M., Sunstein, C. R., & Fleming, S. M. (2023). Why and When Beliefs Change. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *18*(1), 142–151. https://doi.org/10.1177/17456916221082967

Shekhar, M., & Rahnev, D. (2018). Distinguishing the Roles of Dorsolateral and Anterior PFC in Visual Metacognition. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(22), 5078–5087. https://doi.org/10.1523/JNEUROSCI.3484-17.2018

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Soll, J. (2016, December 18). *The Long and Brutal History of Fake News*. POLITICO Magazine. http://politi.co/2FaV5W9

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stein, J., Frey, V., & van de Rijt, A. (2023). Realtime user ratings as a strategy for combatting misinformation: An experimental study. *Scientific Reports*, *13*(1), 1626. https://doi.org/10.1038/s41598-023-28597-x

Steinke, A., Lange, F., & Kopp, B. (2020). Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific Reports*, *10*(1), 15464. https://doi.org/10.1038/s41598-020-72407-7

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

Stuart-Fox, D. (2005). Deception and the origin of honest signals. *Trends in Ecology & Evolution*, *20*(10), 521–523. https://doi.org/10.1016/j.tree.2005.08.004

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific Reports*, *11*(1), 3574. https://doi.org/10.1038/s41598-020-80593-7

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd ed* (pp. xxii, 526). The MIT Press.

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, *28*(19), 3128-3135.e8. https://doi.org/10.1016/j.cub.2018.07.052

Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology. General*, *146*(8), 1143–1149. https://doi.org/10.1037/xge0000298

Tarantola, T., Folke, T., Boldt, A., Pérez, O. D., & Martino, B. D. (2021). *Confirmation bias optimizes reward learning* (p. 2021.02.27.433214). bioRxiv. https://doi.org/10.1101/2021.02.27.433214

Taylor, J. (2023, September 27). X/Twitter scraps feature letting users report misleading information. *The Guardian*. https://www.theguardian.com/technology/2023/sep/27/xtwitter-scraps-function-letting-users-report-misleading-information

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Tenney, E. R., Cleary, H. M. D., & Spellman, B. A. (2009). Unpacking the Doubt in “Beyond a Reasonable Doubt”: Plausible Alternative Stories Increase Not Guilty Verdicts. *Basic and Applied Social Psychology*, *31*(1), 1–8. https://doi.org/10.1080/01973530802659687

Ting, C.-C., Palminteri, S., Lebreton, M., & Engelmann, J. B. (2022). The elusive effects of incidental anxiety on reinforcement-learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 619–642. https://doi.org/10.1037/xlm0001033

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *Annals of the American Academy of Political and Social Science*, *700*(1), 136–151. https://doi.org/10.1177/00027162221087936

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*(1), 35–79. https://doi.org/10.1152/physrev.00041.2012

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*(3), 247–253. https://doi.org/10.1177/0963721419827854

Van Der Linden, S. (2024). Countering misinformation through psychological inoculation. In *Advances in Experimental Social Psychology* (Vol. 69, pp. 1–58). Elsevier. https://doi.org/10.1016/bs.aesp.2023.11.001

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

Van Scoy, L. J., Duda, S. H., Scott, A. M., Baker, A., Costigan, H., Loeffler, M., Sherman, M. S., & Brown, M. D. (2022). A mixed methods study exploring requests for unproven COVID therapies such as ivermectin and healthcare distrust in the rural South. *Preventive Medicine Reports*, *31*, 102104. https://doi.org/10.1016/j.pmedr.2022.102104

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLOS Computational Biology*, *14*(11), e1006632. https://doi.org/10.1371/journal.pcbi.1006632

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vellani, V., Glickman, M., & Sharot, T. (2024). Three diverse motives for information sharing. *Communications Psychology*, *2*(1), 107. https://doi.org/10.1038/s44271-024-00144-y

Vidal-Perez, J., Dolan, R. J., & Moran, R. (2025). Disinformation elicits learning biases. *eLife*, *14*. https://doi.org/10.7554/eLife.106073.1

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*(10), 1337–1353. https://doi.org/10.1080/1369118X.2017.1313883

Vraga, E. K., & Bode, L. (2020). Correction as a Solution for Health Misinformation on Social Media. *American Journal of Public Health*, *110*(Suppl 3), S278–S280. https://doi.org/10.2105/AJPH.2020.305916

Waight, H., Yuan, Y., Roberts, M. E., & Stewart, B. M. (2025). The decade-long growth of government-authored news media in China under Xi Jinping. *Proceedings of the*

National Academy of Sciences, *122*(11), e2408260122. https://doi.org/10.1073/pnas.2408260122

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1870–1884. https://doi.org/10.1016/j.neubiorev.2012.05.008

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wang, W.-C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On Known Unknowns: Fluency and the Neural Mechanisms of Illusory Truth. *Journal of Cognitive Neuroscience*, *28*(5), 739–746. https://doi.org/10.1162/jocn_a_00923

Weinstein, N. D., & Klein, W. M. (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *14*(2), 132–140. https://doi.org/10.1037//0278-6133.14.2.132

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences*, *11*(2), 233–244. https://doi.org/10.1017/S0140525X00049682

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behavioural Brain Research*, *199*(1), 119–128. https://doi.org/10.1016/j.bbr.2008.10.030

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. *Journal* *of Experimental Psychology. General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wineburg, S., & McGrew, S. (2019). Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record*, *121*(11), 1–40. https://doi.org/10.1177/016146811912101102

Wiswall, M., & Zafar, B. (2015a). How Do College Students Respond to Public Information about Earnings? *Journal of Human Capital*, *9*(2), 117–169. https://doi.org/10.1086/681542

Wiswall, M., & Zafar, B. (2015b). How Do College Students Respond to Public Information about Earnings? *Journal of Human Capital*, *9*(2), 117–169. https://doi.org/10.1086/681542

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568. https://doi.org/10.1162/jocn_a_01808

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

Zimmerman, T., Shiroma, K., Fleischmann, K. R., Xie, B., Jia, C., Verma, N., & Lee, M. K. (2023). Misinformation and COVID-19 vaccine hesitancy. *Vaccine*, *41*(1), 136–144. https://doi.org/10.1016/j.vaccine.2022.11.014

Zollo, F., Bessi, A., Vicario, M. D., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, *12*(7), e0181821. https://doi.org/10.1371/journal.pone.0181821

Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, *7*(10), 1667–1681. https://doi.org/10.1038/s41562-023-01640-7