Changing Evidence Accumulation and Sharing Behaviour Through Incentives

Laura Katharina Globig

A dissertation submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
of
University College London

Department of Experimental Psychology
University College London

February 26th, 2024

Declaration

I, Laura K. Globig, 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

Behaviour is driven by external incentives (e.g., money, social feedback) and internal incentives (e.g., emotions). We monitor the stock market for profitable investments, we share posts online to receive positive feedback, and we accumulate evidence for the success of a medical procedure to ease pre-surgery anxiety. This thesis explores how incentives alter evidence accumulation and sharing behaviour.

People accumulate evidence to form beliefs which elicit positive emotions, even if these beliefs are incorrect. This bias is adaptive, as the benefits for well-being typically outweigh the harm of inaccuracy. Chapter 2 examines if evidence accumulation becomes less biased in threatening environments, where severe harm is probable. Combining a social-threat manipulation with a sequential sampling task and Drift-Diffusion Modelling (DDM), I find that under threat, participants are less biased and require weaker evidence for negative conclusions. This may increase precautionary actions.

Although financial accuracy incentives are thought to reduce biases, the empirical evidence is mixed. Chapter 3 examines why they may fail to reduce biased evidence accumulation. Coupling a perceptual task with DDM, I show that while accuracy incentives increase caution, they modulate a separate element of the accumulation process and thus do not reduce the bias, possibly due to its unconscious nature.

I propose that when accuracy incentives are coupled with feedback, decisions become more accurate. External incentives are pervasive online: people share information to receive ‘likes’. I hypothesize that the incentive structure of social media platforms, whereby social rewards (‘likes’) and punishments (‘dislikes’) are dissociated from accuracy, contributes to the spread of misinformation. Chapter 4 combines simulated social media environments and DDM to show that when feedback is contingent on accuracy, sharing becomes more discerning, reducing misinformation spread.

This thesis unveils the mechanisms through which incentives influence evidence accumulation and sharing behaviour, offering insights for interventions to mitigate biased decision-making.

Impact Statement

In today's digital world we are perpetually faced with opportunities to accumulate evidence and share information. This behaviour is motivated by external incentives, such as money or social feedback, and internal incentives, such as emotions. These incentives sometimes lead to biased decisions. This thesis investigates how altering the incentive structure people face can mitigate biased evidence accumulation and make sharing behaviour more discerning.

When accumulating evidence, individuals are biased towards 'desirable' conclusions that make them feel good and thus prioritize 'desirable' evidence over 'undesirable' evidence. This can have detrimental consequences. For example, wanting to avoid negative emotions may cause you to dismiss critical signs of disease, and thus fail to get help. Here I examine whether and how changing the incentive structure people face — by (1) exposing participants to a threatening environment, or (2) rewarding accurate responses — reduces the bias in evidence accumulation.

I find that under threat weaker evidence is required to reach undesirable conclusions. This can be advantageous as it leads to increased precautionary actions in threatening environments. While the negative effects of stress have been repeatedly underscored this study shows that stress, induced by perceived threat, can be adaptive. These findings may also explain overly pessimistic decisions in those with anxiety and depression and highlight possible target mechanisms for therapeutic interventions.

While rewarding accuracy is commonly thought to reduce biased decisions, the empirical evidence is mixed. I show that accuracy incentives do not reduce biased evidence accumulation and provide a mechanistic explanation for why this is the case. Specifically, I find that accuracy incentives and the bias towards desirable conclusions alter orthogonal aspects of the accumulation process. This suggests that participants are unaware of their own bias. As accuracy incentives are commonly employed to improve decision-making, these findings are relevant to academics, policymakers and industry leaders and may help inform the development of novel interventions against biased decisions.

Building on this research, I propose that when accuracy incentives are coupled with direct feedback, decision-making becomes more accurate. I demonstrate that the spread of misinformation online is facilitated by the existing incentive structure of social media platforms in which existing social rewards (e.g., ‘likes’) and punishments (e.g., ‘dislikes’) are dissociated from the accuracy of the information shared. We share information others will react positively to and avoid sharing information others will react negatively to. However, because this feedback is not tied to accuracy, we sometimes share information even if we suspect it may be false. I find that an intervention which slightly changes this incentive structure, such that feedback is contingent on accuracy increases the proportion of true relative to false information shared, without reducing user engagement. The results offer a framework for an intervention that could be adopted to reduce misinformation spread, which in turn could reduce violence, vaccine hesitancy and political polarization.

Bridging theoretical insights with practical solutions, this research provides a mechanistic account of how changing incentives alters evidence accumulation and sharing behaviour, thereby contributing to the development of theory-based behavioural interventions to improve decision-making.

Acknowledgments

I am grateful to everyone who has played a role in my PhD journey.

I would like to thank my PhD supervisor, Tali, for her guidance and expertise as well as the invaluable opportunities she provided throughout the years. I would also like to thank my second supervisor, Steve, as well as the other members of my thesis committee, Oli, and Irene, for their insightful advice and support.

I am grateful to both past and present members of the Affective Brain Lab and the wider research community at UCL and MIT for their continued camaraderie. Special mentions to Irena for being an essential pillar of support over the years, and to Chris for being both a sounding board and companion in the lab.

To my friends in London, Boston, and further afield: you have accompanied me through trials and triumphs. Your friendship means the world to me.

Finally, my deepest gratitude goes to my family. I am indebted to my grandparents, my parents, and my sister for their unwavering support, which has shaped me into who I am today. A special thanks, of course, also goes to Poppy and Indie, who have been a constant source of joy and comfort.

Notes To Examiners

The findings from Chapter 2 have been published in a peer-reviewed article: Globig, L. K., Witte, K., Feng, G., & Sharot, T. (2021). Under threat, weaker evidence is required to reach undesirable conclusions. *Journal of Neuroscience*, *41*(30), 6502-6510. https://doi.org/10.1523/JNEUROSCI.3194-20.2021

The findings from Chapter 4 have been published in a peer-reviewed article: Globig, L. K., Holtz, N., & Sharot, T. (2023). Changing the incentive structure of social media platforms to halt the spread of misinformation. *Elife*, *12*, e85767. https://doi.org/10.7554/eLife.85767

All chapters have benefited from the guidance and advice of my supervisor Tali Sharot. The 'Factory Task' used in Chapter 2 was designed by Donal Cahill, Filip Gesiarz, and Tali Sharot (Gesiarz et al., 2019). The data presented in Chapter 2 was partially collected by Gloria Feng and Kristin Witte. Nora Holtz assisted with creation of stimuli and task design for Chapter 4.

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction

1.1 Reward-Oriented Evidence Accumulation

In today's *age of information* (Castells, 1996) data is constantly at our fingertips. Recent reports show that every day, internet users spend nearly seven hours online (Statista, 2023), able to effortlessly accumulate an abundance of noisy information (or 'evidence'). This raises a fundamental question: *Why* do we accumulate evidence?

1.1.1 External Incentives

Humans accumulate evidence by continuously sampling noisy information, in order to form beliefs and make decisions (Platt & Glimcher, 1999; Ratcliff, 1978; Usher & McClelland, 2001) which can gain them rewards and help them avoid punishments (= *external outcomes*, **see Figure 1.1**; Gold & Shadlen, 2002, 2007). These outcomes can be tangible, such as monetary gains or losses, and/or intangible, for instance in the form of social feedback, which at times can have tangible consequences. For example, an individual might accumulate evidence to decide whether to invest in emerging technology. To make this decision, they need to arbitrate between two alternatives: 'to invest' or 'to withhold'. In doing so, they may be motivated by financial profit and by others' reactions to their decision. When the evidence in favour of one alternative relative to the other is large enough, they make their decision (see also Ratcliff & McKoon, 2008). For instance, if the individual finds strong evidence that an investment will yield substantial financial and/or social returns, they form the belief that the investment is profitable, decide to invest, and may later reap the rewards.

1.1.2 Internal Incentives

Evidence accumulation is not solely motivated by external incentives. Rather, information also carries intrinsic utility (Blanchard et al., 2015; Bromberg-Martin et al., 2024; Bromberg-Martin & Hikosaka, 2009, 2011; Charpentier et al., 2018; Eliaz & Schotter, 2007; Grant et al., 1998) even if it provides no instrumental benefit. It can deepen our sense of understanding the world around us (Dörnemann et al., 2022; Johnston & Davey, 1997; Kobayashi & Hsu, 2019; Sharot & Sunstein, 2020). For example, imagine someone gathers information about black holes. While this information is unlikely to motivate decisions which lead to financial or social rewards, it increases the individual's understanding of the world around them. Information can also alter our emotions. For example, learning that we are at high risk of cancer will make us feel anxious, while learning that we received a high score on a test can make us feel happy. Unsurprisingly therefore, people prefer seeking desirable information (Charpentier et al., 2018) - which elicits positive emotions (**see Figure 1.1**).

It has also been shown that individuals selectively accumulate evidence to form positive beliefs from which they derive internal rewards, such as positive emotions (Gesiarz et al., 2019; Leong et al., 2019). For example, in Gesiarz et al. (2019), participants completed a sequential sampling task, in which they had to determine whether they were in a desirable state, associated with greater financial rewards than losses, or an undesirable state, associated with greater financial losses than rewards. Crucially, they had no control over which state they were in and were financially incentivized to accurately identify whether they were in a desirable or undesirable state. In spite of this, participants were biased towards concluding they were in a desirable state and required weaker evidence to reach a desirable compared to an undesirable conclusion. Computational modelling revealed that this desirability bias was due to participants placing more weight on evidence that aligned with their preferred (positive) belief (i.e., believing that they were in a desirable state), compared to evidence that opposed it. As a result, they were faster to reach a desirable, internally rewarding conclusion and continued to accumulate evidence for longer when it pointed to an undesirable conclusion, perhaps hoping to refute it. This tendency is also observed in the real world. For instance, some missed cancer diagnoses are due to doctors deciding not to conduct additional tests when the initial evidence suggests that the patient may not be ill (Lyratzopoulos et al., 2015). Neurocomputational evidence attributes this bias in evidence accumulation to a selective increase in neural activity for the desirable state, which encompasses both anticipatory activity in the nucleus accumbens (nACC), as well as activity in sensory regions of the brain which track the accumulation of evidence towards the desired conclusion (Calabro et al., 2023; Leong et al., 2019). This suggests that internal incentives not only influence the way in which we accumulate evidence but also alter the neural representation of our current state. Simply put, people accumulate evidence in a way that fosters positive emotions and reduces negative emotions.

1.1.3 A Bias towards Internal Incentives: Adaptation or Flaw?

Prior work suggests that when accumulating evidence, internal incentives may take precedence over external incentives (Gesiarz et al., 2019; Leong et al., 2019). The tasks used in these studies were designed such that a bias towards desirable conclusions was financially suboptimal. Participants preferred forming positive beliefs from which they derived internal rewards, even if these beliefs caused them to incur financial losses. This can have detrimental consequences. For instance, it is thought that the tendency to prioritize internal incentives, such as positive emotions or increased sense of self-efficacy, has contributed to falsely optimistic investment decisions preceding the financial crisis (Shefrin, 2015), as well as a false sense of security which resulted in failure to take preventative measures against natural disasters (Paton, 2003). So why then do humans prioritize positive beliefs, and thus internal incentives over external incentives when they accumulate evidence?

Despite the potential negative repercussions, there are numerous benefits associated with forming positive beliefs even when they are false. It has been suggested that positive beliefs from which individuals derive internal rewards promote mental (Carver & Scheier, 2014; Taylor et al., 2000; Taylor & Brown, 1994), as well as physical well-being (Hernandez et al., 2015; Tindle et al., 2009). Patients who had positive expectations of their future, were less likely to develop coronary heart disease (Tindle et al., 2009), had a lower chance of re-hospitalization after surgery (Scheier et al., 1999), and had higher survival rates after cancer diagnosis (Allison et al., 2003; Novotny et al., 2010). When individuals hold positive beliefs, they feel a greater sense of control and self-efficacy (McFarland & Ross, 1982). This motivates them to adopt measures to protect their well-being, such as engaging in regular physical exercise (Giltay et al., 2007; Steptoe et al., 2006). Positive beliefs have also been shown to reduce stress (Jobin et al., 2014), which, over time, improves wellbeing (Taylor et al., 2000; Taylor & Brown, 1994). By contrast, individuals with major depression tend to be more pessimistic about their future (Cropley & MacLeod, 2003; Strunk et al., 2006; Strunk & Adler, 2009). In light of their positive influence on physical and mental well-being, it has been argued, that positive beliefs are adaptive even when inaccurate (McKay & Dennett, 2009). For instance, if a patient erroneously believes that they have a high probability of fully recovering from a severe injury, this can lead to a more positive attitude and greater effort in the rehabilitation process. As such, accumulating evidence in a way that prioritizes the formation of positive beliefs ('biased evidence accumulation') and thus generates internal rewards, may too be adaptive, even when those beliefs are inaccurate.

When vying for resources, biased individuals may be more persistent and assertive in staking their claims on otherwise unattainable resources, while their more capable, yet less confident rivals might withdraw from the competition (Johnson & Fowler, 2011). Compared to unbiased individuals, those who overestimate their ability to succeed, may also be more likely to apply to challenging jobs, thereby increasing their chances of securing employment. In contrast, when the potential costs of erroneously held positive beliefs clearly exceed the benefits it is advantageous to be unbiased (Johnson & Fowler, 2011). In light of this, it has been suggested that whether individuals prioritize internal incentives over external incentives depends on the environment (Basten et al., 2010; Bromberg-Martin & Sharot, 2020; Garrett et al., 2018; Sharot & Garrett, 2016). The same may be true for evidence accumulation (**see Figure 1.1**). In environments where the benefits of being biased towards positive beliefs outweigh the costs, as is the case in safe environments, individuals place less weight on negative, undesirable evidence, which may be advantageous for their physical and mental wellbeing. By contrast, in environments where the costs of being biased towards positive beliefs clearly outweigh the benefits, such as threatening environments, they shift their attention towards negative evidence and prioritize external incentives. Depending on the outcomes they obtain, they can then adjust their behaviour to maximize overall utility (Bond et al., 2023).

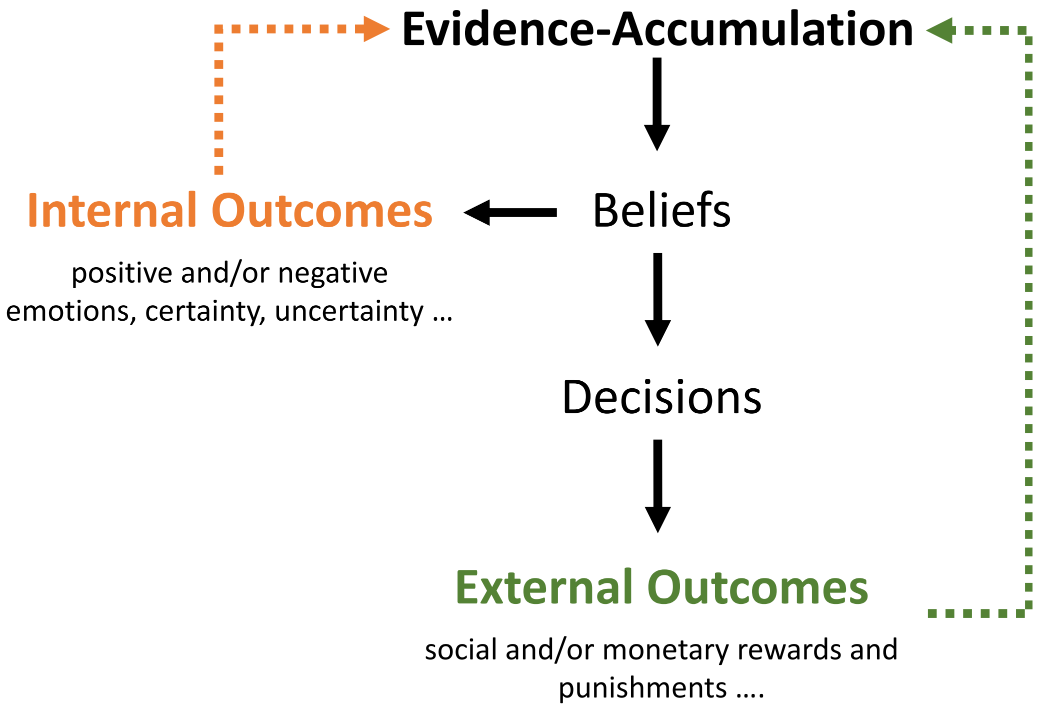
Evidence-Accumulation
Internal Outcomes
positive and/or negative emotions, certainty, uncertainty …
Beliefs
Decisions
External Outcomes
social and/or monetary rewards and punishments ….

Figure 1.1. Reward-Oriented Evidence Accumulation. Humans accumulate evidence to form beliefs which motivate decisions and actions towards desirable external outcomes (e.g., monetary rewards). However, they are also motivated to accumulate evidence to form beliefs which give rise to desirable internal outcomes (e.g., positive emotions), sometimes even at the expense of external outcomes. Informed by the outcomes they obtain; they adaptively adjust their behaviour to maximize overall utility in a given environment.

1.1.4 Changing the Incentives to Accumulate Evidence

Thus far, it has been shown that internal incentives influence evidence accumulation, such that participants are biased towards desirable (= internally rewarding) conclusions (Gesiarz et al., 2019; Leong et al., 2019). Prior research has also underscored the influence of contextual features on value-based decision-making (for a review see Engelmann & Hein, 2013). If individuals adaptively prioritize internal and external incentives based on their environment, evidence accumulation should be less biased when false beliefs resulting in false conclusions are costly. Based on this assumption, this thesis investigates whether incentives for evidence accumulation can be altered to make evidence accumulation less biased. First, I will test the hypothesis that exposing participants to an environment in which the costs of false beliefs are high, they are less biased towards desirable conclusions. Then, I will investigate whether increasing the external incentives to provide correct responses can mitigate the influence of internal incentives on evidence accumulation.

Perceived Threat

In threatening environments, the costs of false beliefs can be especially high (Dunning, 2009; Haselton & Nettle, 2005; Johnson & Fowler, 2011). These environments elicit a physiological stress response which signals a high risk-situation (Nesse et al., 2016). In these situations, selective sampling of positive, desirable evidence could be detrimental, as the potential for adverse outcomes is high. Rather, it is adaptive to err on the side of caution. Imagine, you are walking through a high-crime area at night, and you see someone walking closely behind you. It might just be a coincidence, but it could also be someone trying to accost you. Assuming the latter, you will likely change your route and as a result avert potential danger. It then follows that under threat individuals prioritize the accumulation of negative, undesirable evidence to pre-emptively mitigate potential losses, thereby reducing the bias in evidence accumulation towards desirable conclusions. This hypothesis aligns with prior research. When individuals are anxious, anticipated adverse consequences override the importance of possible desirable outcomes in risky decision-making (Engelmann et al., 2015). In line with this, Robinson et al., (2011) observed a threat-induced bias towards aversive stimuli in an emotional Stroop task in which participants had to identify appetitive and aversive faces. Stress reportedly increases attention to negative stimuli (Robinson et al., 2012; Robinson, Vytal, et al., 2013), thereby reducing preferential belief updating for desirable information (Garrett et al., 2018), and improving response inhibition (Robinson, Krimsky, et al., 2013). However, until now, the hypothesis that perceived threat mitigates biased evidence accumulation has not been tested.

In Chapter 2, I therefore examine how perceived threat impacts evidence accumulation. I combine a social-threat manipulation with a sequential sampling task in which participants have to determine whether they are in a desirable state, associated with greater financial rewards than losses, or an undesirable state, associated with greater financial losses than rewards. Using computational modelling, I then examine if and how acute stress alters the process by which evidence is accumulated when participants are motivated to reach desirable, internally rewarding conclusions.

Accuracy Incentives

If evidence accumulation becomes less biased in environments, in which false beliefs are costly, it is plausible, that directly increasing the external incentives for accurate beliefs might have a similar effect. The rationale for this is that when external incentives are tied to accuracy, the costs of false beliefs increase. One way this may be achieved, is by financially incentivizing people to make more accurate decisions. Historically, many companies have employed financial bonuses not only to improve overall performance but also as a targeted approach to diminish biases. For instance, Goldman Sachs implemented financial incentives in an attempt to reduce bias in hiring practices (AFR, 2015). This strategy is predicated on the understanding that biases and heuristics, can be moderated through deliberate and effortful thinking when performance-related rewards are at stake (Bonner & Sprinkle, 2002; Botvinick & Braver, 2015; Smith & Walker, 1993).

Yet, despite the intuitive appeal of this approach, prior research using financial accuracy incentives presents a conflicting picture (Dale et al., 2007; Engelmann et al., 2019; Epley & Gilovich, 2005; Lefebvre et al., 2011; Meub et al., 2013, 2013; Wright & Anderson, 1989, 1989; Zhang & Rand, 2023). On the one hand, accuracy incentives foster deliberation and (slow) rational thinking thereby reducing the reliance on cognitive shortcuts (Dale et al., 2007; Epley & Gilovich, 2005; Lefebvre et al., 2011; Meub et al., 2013; Wright & Anderson, 1989). On the other hand, there is some evidence to suggest that accuracy incentives are ineffective against motivational biases (Engelmann et al., 2019; Enke et al., 2023; Zhang & Rand, 2023; but see Prior et al., 2015; Rathje et al., 2023), which occur because individuals *want* to hold certain beliefs, such as wanting to believe they will be promoted (Montibeller & Von Winterfeldt, 2015).

In Chapter 3, I test a possible explanation for why financial accuracy incentives may fail to mitigate motivational biases. I hypothesize that financial incentives motivate participants to invest more cognitive effort. While this may aid in reducing the use of heuristics, which are the result of fast, careless processing; increased effort will fail to mitigate biases, when the cause of the error is a bias that is beyond participants' awareness, such as a bias in how evidence itself is processed (for example as when people put greater weight on desirable than undesirable evidence as shown in Gesiarz et al., 2019; Globig et al., 2021). To test this hypothesis, I examine the effect of varying accuracy incentives in a perceptual evidence accumulation task, in which participants have to determine whether they are in a desirable state, associated with financial rewards, or an undesirable state, associated with no rewards. Using computational modelling, I then examine if and how accuracy incentives alter the process by which evidence is accumulated when participants are motivated to reach a desirable, internally rewarding conclusion. The results will be particularly relevant in areas such as finance, politics, and healthcare, where biased evidence accumulation can have detrimental consequences and financial incentives are frequently used to alter behaviour (AFR, 2015; Fainman & Kucukyazici, 2020; Hasnain & Pierskalla Henryk, 2012).

Together, the results from Chapter 2 and 3 will enhance our understanding of internal and external incentives for evidence accumulation. The results can inform the development of interventions to prevent biased accumulation in situations in which it can have detrimental consequences.

1.2 Reward-Oriented Information-Sharing

As the reach of digital information technologies widens, it has become increasingly easy not only to accumulate evidence, but also to share information with others. Whether it is telling a friend about a new restaurant or sharing information about a disease outbreak, the social transmission of information has become a ubiquitous feature of everyday life. While this mounting ease of sharing has facilitated globalization and social connection, it has also given rise to an unprecedented explosion of misinformation (Lazer et al., 2018). This has had drastic consequences such as increased polarization, racism, and resistance to climate action and vaccines (Barreto et al., 2021; Rapp & Salovich, 2018; Tsfati et al., 2020; Van Bavel et al., 2021).

Unsurprisingly therefore, researchers, policymakers and industry leaders alike are working on mitigating the spread of misinformation (Saltz et al., 2021; Traberg et al., 2022). As yet, however, existing interventions to halt the spread, such as flagging misleading content, have only had limited impact (Chan et al., 2017; Grady et al., 2021; Lees et al., 2022). Here, I propose that to effectively improve sharing behaviour, a deep understanding of the underlying incentives driving sharing decisions is necessary. This type of 'from-theory-to-practice' approach has long been adopted in the medical sector, where researchers devote significant amounts of time and resources to investigating the pathways of health and disease in order to develop and optimize treatments. Such theory-informed behavioural change interventions have been shown to be more successful in promoting health-related behaviour than non-theory-based interventions (Webb et al., 2010). Yet, despite frequently drawing analogies between infectious diseases and the spread of misinformation (Bonnevie et al., 2021; van der Linden, 2022; Zarocostas, 2020), over 70% of existing interventions are not informed by basic theory (Ziemer & Rothmund, 2024).

1.2.1 Incentives for Information-Sharing

Classical reinforcement theory stipulates that humans seek actions that result in the greatest rewards and avoid those that lead to punishments (Skinner, 1966). Accordingly, they accumulate evidence to make decisions which result in positive outcomes. One such value-based decision is the decision to share information (Globig et al., 2023; Lin et al., 2022; Scholz et al., 2020).

External Incentives

Prior research shows that people share information with others to optimize external outcomes (Lindström et al., 2021; Scholz et al., 2017, 2020; Scissors et al., 2016). Often, these outcomes are intangible, in the form of social feedback (Barasch, 2020). This feedback can either be positive ('social rewards'), through social approval, reputational gains, or increased sense of social connection (Delgado et al., 2023) or negative ('social punishments'), through criticisms, reputational losses, or isolation (Brudner et al., 2023). Since the brain processes social feedback analogously to monetary outcomes (Bhanji & Delgado, 2014; Gu et al., 2020; Meshi et al., 2013; Ruff & Fehr, 2014; Wake & Izuma, 2017), and the receipt of social rewards activates reward-regions of the brain including the ventral striatum and the ventromedial prefrontal cortex (VMPFC, Davey et al., 2010; Klucharev et al., 2009; Morelli et al., 2014), it is no surprise that social feedback has been shown to be just as, if not more, effective in directing human action (for a review see Tamir & Hughes, 2018). For instance, humans are more likely to share information with others if they believe it will elicit positive feedback, and choose to abstain from doing so if they fear it will elicit negative feedback (Brudner et al., 2023). Social media platforms have capitalized on this sensitivity to incentives, by implementing social rewards, such as 'likes' and 'reposts' to maximize user engagement. More recently, they have also implemented monetization schemes, which financially reward some users whose sharing behaviour generates particularly high engagement (Alizadeh et al., 2023). In these cases, intangible rewards, in the form of social feedback, bring about tangible rewards in the form of monetary gains. Thus, social media users are motivated to share information with others in order to maximize the rewards they receive. The more responsive they perceive their social network to be, the more they share (Walsh et al., 2020). Over time, users learn what elicits positive or negative reactions and thus adjust their sharing behaviour to increase the likelihood of positive feedback (Brady et al., 2021; Lindström et al., 2021; Scissors et al., 2016). As such, information-sharing is motivated by external incentives (**see Figure 1.2**).

Internal Incentives

On top of these external incentives, it has also been suggested that sharing may be rewarding in itself (Baek et al., 2017; Tamir et al., 2015; Tamir & Mitchell, 2012). For example, consider sharing your experience about a recent unsuccessful job interview. Disclosing this experience may elicit reactions of social support, but it can also help you regulate your emotions by allowing you to reflect on the interview (Berger, 2014; Rimé, 2009). This may help you make sense of the experience. For instance, through self-reflection you may realize that the unsuccessful interview was not a result of your performance, but rather because the job did not match your interests, thereby bolstering your perceived self-efficacy and mitigating negative emotions (Niederhoffer & Pennebaker, 2009). Indeed, a recent study shows that information-sharing processes involve the activation of self-related-processing regions of the medial prefrontal cortex (MPFC) and the posterior cingulate cortex (PCC, Baek et al., 2017), suggesting that self-reflection is a key motivation for information-sharing. It then follows that sharing decisions are likely motivated not only by external incentives but also by internal incentives (Chen et al., 2019; Fu et al., 2017; Rode, 2016, **see Figure 1.2**).

1.2.2 Sharing as a Value-Based Decision

Building on this work, sharing has been characterized as a value-based decision, that involves both self-related and other-related considerations (Scholz et al., 2020, 2023). Sharing decisions elicit activity in regions associated with positive valuation, including the ventral striatum and VMPFC (Berger, 2014; De Angelis et al., 2012; Lee & Ma, 2012; Wien & Olsen, 2014). Activity in these regions scales with how enthusiastically messages are shared (Falk et al., 2012). Echoing these findings, this thesis conceptualizes sharing decisions, as the result of reward-oriented evidence accumulation, in which sharers adjudicate between, sometimes competing, internal and external incentives (**see Figure 1.2**), weighing the costs against the benefits throughout. In some cases, they favour internal incentives, as evidenced by research reporting that participants choose to forgo monetary rewards in order to disclose information about themselves (Tamir & Mitchell, 2012). In others, the prospect of external rewards, such as social capital may override internal rewards (Fu et al., 2017). This is especially pertinent in online contexts, where shared information can reach a larger audience and thus elicits a greater volume of feedback, i.e., social incentives (Bodaghi & Oliveira, 2020). Over time users learn what type of information elicits the greatest amount of rewards and thus adaptively adjust their behaviour to maximize overall utility (Brady et al., 2021; Lindström et al., 2021; Scissors et al., 2016). Sharing that elicits rewards is reinforced, sharing that elicits punishments is suppressed. As, external rewards tend to be easier to quantify than internal rewards and given that social incentives are already a core feature of existing social media platforms, the reinforcing feedback loop between sharing and social incentives provides a bullseye for measures designed to improve sharing behaviour online.

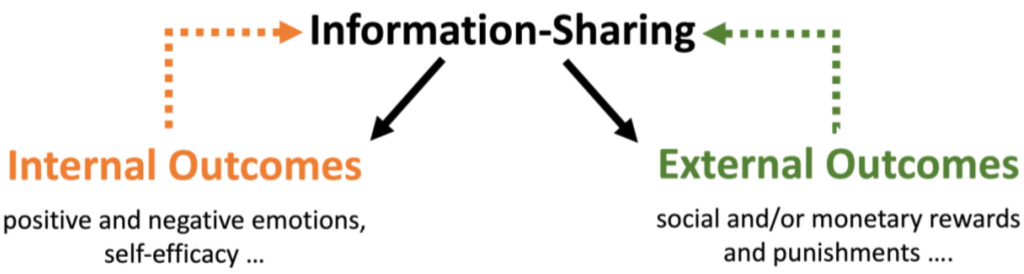
Information-Sharing
Internal Outcomes
positive and negative emotions, self-efficacy ...
External Outcomes
social and/or monetary rewards and punishments ....

Figure 1.2 Reward-Oriented Information-Sharing. Individuals share information to obtain external rewards (e.g., positive social feedback and/or monetary rewards). However, they are also motivated to share information which gives rise to internal rewards (e.g., positive emotions, self-efficacy). Informed by the outcomes they obtain and weighing the costs against the benefits throughout, they adaptively adjust their behaviour to maximize overall utility.

1.2.3 Shaping Sharing Decisions with Social Feedback

The reinforcement-model like pattern of online behaviour, in which users respond to social rewards and punishments (Brady et al., 2021; Lindström et al., 2021; Rosenthal-von der Pütten et al., 2019; Scissors et al., 2016) may also explain the apparent disconnect between what people believe themselves and what they share online (Pennycook et al., 2021; Ren et al., 2021). For instance, users may choose to abstain from sharing content they believe will elicit negative reactions and instead post content which they believe will maximize engagement (Brady et al., 2021; Lindström et al., 2021; Scissors et al., 2016). Notably, it has been shown that misinformation often generates more engagement than reliable posts (Lazer et al., 2018). As such, users have little reason to take into account the veracity of information when deciding what type of content to share. Instead, they are incentivized to share the type of content that they anticipate will elicit the largest amount of positive engagement, which sometimes translates to monetary rewards. Thus, both social (e.g., 'likes', 'shares') and monetary incentives (e.g., monetization schemes) contribute to spread of misinformation online (Alizadeh et al., 2023).

With time, this not only vastly accelerates the spread of misinformation, but may also alter users' beliefs. Empirical evidence suggests that when people learn that others share their belief, their confidence in those beliefs increases (Kappes et al., 2020). Moreover, it has been shown that even a single repetition of misinformation leads to it being perceived as more accurate, irrespective of whether it aligns with the user's ideology (Murray et al., 2020; Pennycook et al., 2018). This 'illusory truth effect' (Hasher et al., 1977) then in turn influences sharing, suggesting a bi-directional relationship of sharing and beliefs (Van Bavel et al., 2021; Vellani et al., 2023). Thus, the current incentive structure of social media platforms, in which incentives, such as rewards in the form of 'likes' and punishments in the form of 'dislikes' are dissociated from the veracity of the information shared, facilitates the dissemination of misinformation, and may also fuel the spread of erroneous beliefs (see also Epstein et al., 2023). Users strive to maximize external rewards, without being motivated to consider whether the information they are sharing is accurate.

In Chapter 4, I suggest that to mitigate the spread of misinformation, a modified incentive structure is needed in which rewards and punishments are directly contingent on the veracity of the information shared. To test this assumption, I examine the impact of slightly altering the engagement options offered to users. Specifically, I add an option to react to posts using ‘trust’ and ‘distrust’ buttons, which are, by definition, related to veracity. I then examine how this change affects discernment in sharing behaviour between true and false information and if it can help in correcting false beliefs. Using computational modelling I dissect the mechanisms underlying this change.

1.3 Drift-Diffusion Modelling: A Window into the Mechanisms of Decision-Making

A common issue of behavioural studies is the difficulty to obtain insights into the mechanisms underlying observed behaviour (Roberts & Hutcherson, 2019). In many studies, directly observable variables are limited to choices and response times. Notably, these variables are aggregate outcomes and may mask the underlying processes (Clithero, 2018; Stafford et al., 2020). For example, slower response times can be the result of more tenuous cognitive processing, and/or a higher threshold for decision certainty, reflecting a more cautious decision-making process (Forstmann et al., 2016; Voss et al., 2013). Unsurprisingly therefore, a growing number of researchers are shifting focus towards computational models (Calder et al., 2018). By formalizing observed behaviours as mathematical frameworks, these models allow researchers to examine the underlying mechanisms of decision-making (Wilson & Collins, 2019).

For instance, Drift-Diffusion Models (DDM), a subtype of sequential sampling models, describe the process by which individuals accumulate evidence to make a decision (Ratcliff et al., 2016), for example to share information (Globig et al., 2023; Lin et al., 2023). They model the decision process between two (or more) alternatives as the noisy accumulation of evidence over time (Ratcliff, 1978; Ratcliff & Rouder, 1998; Voss et al., 2013, see Roxin, 2019 for multi-alternative DDMs). In doing so, they assume evidence accumulation is driven by random fluctuations, i.e., diffusion, to mimic the gradual integration of sensory information in the brain. More specifically, DDMs assume that agents accumulate evidence at a given rate until the evidence for one alternative relative to the other alternative(s) is large enough to reach a pre-determined decision threshold and a decision is made (Forstmann et al., 2016; Ratcliff et al., 2016). This evidence can be external, e.g., from the stimuli observed, or internal, based on internal representations of the alternatives (Krajbich, 2019).

Conventionally, DDMs include the following parameters: (1) the drift rate (v) – which is the rate at which evidence is accumulated; (2) the distance between decision thresholds (α) — which captures the amount of evidence required to form a decision; (3) the starting point (z) of the accumulation process; and (4) the amount of non-decision time (t0)—which includes stimulus encoding as well as response preparation and execution. Each of these parameters can be fixed based on prior assumptions or can be allowed to vary freely, for example as a function of different variables, such as trial condition or trial difficulty. These parameters have been found to correspond with specific neural and physiological patterns (Basten et al., 2010; Cavanagh, Wiecki, et al., 2011; Krajbich et al., 2010; Mulder, 2014). This underscores the potential of DDMs to successfully capture and provide insights into neural processes.

Taken together, DDMs therefore offer a viable, cost-efficient, and complementary solution to gather insights into decision-making processes (Berlinghieri et al., 2023). They can help identify potential target mechanisms for measures designed to reduce biases (Arkes, 1991; Krajbich, 2022). In this thesis I capitalize on the potential of DDM to dissect the mechanisms underlying the influence of incentives on evidence accumulation and sharing decisions.

1.4 Summary

When navigating today's information ecosystem, individuals are perpetually faced with opportunities to accumulate evidence and share information. This behaviour is motivated by (1) external incentives, such as financial or social rewards and punishments (Gold & Shadlen, 2002; Rosenthal-von der Pütten et al., 2019), and (2) internal incentives, such as positive and negative emotions

(Baek et al., 2017; Gesiarz et al., 2019; Leong et al., 2019; Tamir et al., 2015; Tamir & Mitchell, 2012). Sometimes, these incentives can steer individuals towards decisions that yield adverse consequences, such as overly optimistic investment decisions (Barber & Odean, 1999; Shefrin, 2015), or sharing of misinformation (Pennycook et al., 2021). While prior research has sought to understand how different motivations influence evidence accumulation (Gesiarz et al., 2019; Hausmann & Läge, 2008; Kelly & O'Connell, 2013; Leong et al., 2019) and information-sharing decisions (Baek et al., 2017; Rode, 2016; Scholz et al., 2020), how and when changing these incentives can reduce biased evidence accumulation and increase the sharing of accurate information has been largely unexplored and is the focus of this thesis.

Individuals accumulate evidence to arrive at desirable, internally rewarding conclusions, even at the expense of external outcomes (Gesiarz et al., 2019; Leong et al., 2019). By prioritizing internal incentives, they discount undesirable, negative evidence. As a result, they are biased towards internally rewarding conclusions (Gesiarz et al., 2019; Leong et al., 2019). This bias can contribute to negligent decision-making, as for instance observed during the financial collapse of 2008 (Shefrin, 2015). To prevent such negative ramifications, it is imperative to understand how and when biased evidence accumulation can be mitigated.

It has been suggested that individuals adaptively prioritize internal and external incentives depending on their environment (Dunning, 2009; Haselton & Nettle, 2005; Johnson & Fowler, 2011). In Chapter 2, I test the hypothesis that when individuals find themselves in a high-threat environment, in which the costs of false beliefs are high, they are less biased towards desirable conclusions. To that end, I combine a social-threat-manipulation with a sequential sampling task and DDM to understand how threat alters evidence accumulation.

Intuitively, we assume that financially incentivizing individuals to provide correct responses will enhance the accuracy of their decisions. In Chapter 3, I test an explanation for why financial accuracy incentives may fail to mitigate the influence of internal incentives on evidence accumulation. I tease apart the underlying mechanisms by combining an incentivized perceptual evidence accumulation task with DDM.

Together, these findings will enhance our understanding of the role of internal and external incentives in evidence accumulation and can aid in the development of interventions to prevent the potential negative ramifications of biased evidence accumulation.

One example of a decision based on evidence accumulation is the decision to share information with others. Over five billion people globally are now active on social media platforms (Statista, 2022), making them a vital source of information for many (Pew Research Center, 2021). Ensuring this information is reliable is therefore of utmost importance. Much of the success of social media platforms has been attributed to the human need for rewards (Rosenthal-von der Pütten et al., 2019). Over time, users learn what type of information elicits the greatest amount of positive feedback (e.g., 'likes' and 'shares') and adjust their sharing behaviour accordingly (Brady et al., 2021; Lindström et al., 2021; Scissors et al., 2016). But as engagement is not tied to the veracity of information, and misinformation often elicits more engagement than reliable information, users are not motivated to discern between the two. I thus argue that the spread of misinformation is facilitated by the existing incentive structure of social media platforms. In Chapter 4, I explore whether a slight change to this structure, such that social rewards and punishments are contingent on the veracity of information can overturn the adverse effects of incentives on sharing behaviour and thereby reduce the spread of misinformation online. I provide a mechanistic account of how this change affects discernment between true and false posts, by modelling sharing decisions as a drift-diffusion process.

In conclusion, my research will shed new light on how incentives to accumulate evidence and share information can be modulated to enhance discernment in decision-making. The results not only expand our understanding of how incentives impact evidence accumulation and sharing but also highlight the need for innovative approaches in policy and decision-making strategies. Ultimately, this work paves the way for more informed and nuanced applications of behavioural interventions in various domains ranging from public policy to individual decision-making processes.

Chapter 2: Under Threat Weaker Evidence is Required to Reach Undesirable Conclusions

2.1 Abstract

Critical decisions, such as in domains ranging from medicine to finance, are often made under threatening circumstances that elicit stress and anxiety. The negative effects of such reactions on learning and decision-making have been repeatedly underscored. In contrast, here we show that perceived threat alters the process by which evidence is accumulated in a way that may be adaptive. Participants (N = 91) completed a sequential sampling task in which they were incentivized to accurately determine whether they were in a desirable state, which was associated with greater rewards than losses, or an undesirable state, which was associated with greater losses than rewards. Prior to the task participants in the 'threat group' experienced a social-threat manipulation. Results show that perceived threat led to a reduction in the strength of evidence required to reach an undesirable conclusion. Computational modelling revealed this was due to an increase in the relative rate by which negative evidence was accumulated. The effect of the threat manipulation was global, as the alteration to evidence accumulation was observed for evidence which was not directly related to the cause of the threat. Requiring weaker evidence to reach undesirable conclusions in threatening environments may be adaptive as it can lead to increased precautionary action.

2.2 Introduction

Many important decisions are made when people feel stressed and anxious (Beilock, 2010). Consider a doctor in the operating theatre who needs to decide on the best course of action, a soldier on the battlefield who must decide whether to attack, or a driver stuck in traffic selecting which route to take. Whether calm or stressed, to make good decisions people need to accumulate evidence over time (Forstmann et al., 2016; Platt & Glimcher, 1999; Ratcliff, 1978; Usher & McClelland, 2001). For example, a doctor may decide to consult multiple colleagues before deciding to amputate. Because evidence can be unlimited, an agent needs to determine when the available data is strong enough to reach a conclusion (Gluth et al., 2012, 2013). Here, we examine how perceived threat impacts the process by which evidence is accumulated to reach a conclusion.

A feature of threatening environments is that the potential for adverse outcomes is high. In these instances, it is adaptive to err on the side of caution. For example, imagine you are walking through a dark alley and hear a ‘pop’. The sound may be a gunshot or perhaps the uncorking of a champagne bottle. Interpreting the sound as the former will cause you to escape and mitigate potential risk. Thus, under perceived threat it may be adaptive to interpret a stimulus as undesirable even if the strength of the evidence supporting this conclusion is only limited. The psychophysiological reaction induced by threat can provide a global, rather than specific, danger signal. We thus hypothesized that the effects of threat on evidence accumulation may be observed even when the source of the threat is unrelated to the decision at hand (e.g., a psychophysiological reaction triggered by a professional conflict may impact how the ‘pop’ is interpreted).

Computationally, this process may occur in at least two ways. First, under perceived threat people may be predisposed towards undesirable conclusions before attaining any evidence (e.g., you may believe the road you are walking down is dangerous before observing any evidence to that effect). A second, not mutually exclusive possibility, is that under perceived threat an undesirable piece of evidence (e.g., an anxious looking man walking down the road) drives beliefs towards an undesirable conclusion (‘this road is dangerous’), more so than a desirable piece of evidence (e.g., people are walking past you relaxed and happy) towards a desirable conclusion (‘this road is safe’). These two distinct mechanisms will result in the same observable behaviour. Specifically, weaker evidence will be needed to support undesirable conclusions under perceived threat.

To tease apart these mechanisms, we used a sequential sampling model to model noisy evidence accumulation towards either of two decision thresholds (Ratcliff, 1978; Ratcliff & Rouder, 1998; Voss et al., 2013). The model allows estimation of both (1) starting point, and (2) rate of evidence accumulation, reflecting the quality of information processing. We can then measure whether either of these factors are influenced by the desirability of a decision and how this is influenced by perceived threat.

We exposed participants to an acute threat manipulation in the lab (Garrett et al., 2018) or a control condition, and then asked them to complete an evidence accumulation task (Gesiarz et al., 2019) that was unrelated to the cause of the threat. In the task, participants witnessed various stimuli that were contingent upon which one of two hidden states they were in. One state was associated with greater rewards than losses (desirable state) and the other with greater losses than rewards (undesirable state). Participants had no control over which state they were in; their task was simply to determine the state, gaining additional rewards for accurate conclusions and losing rewards for inaccurate conclusions. Thus, it was in participants' interest to be as accurate as possible. They were allowed to accumulate as much evidence as they wished before making a decision. Here, we examined if and how perceived threat impacts the accumulation of evidence towards a decision.

2.3 Methods

Experimental Design.

Participants. A total of 91 individuals participated in this study at two sites: University College London (UCL, N = 51) and Massachusetts Institute of Technology (MIT, N = 40). They were recruited via the participant pools of UCL and MIT. All analyses were repeated separately for participants tested in the two different locations (MIT, UCL). There were no differences between locations in any of our results including model-free analysis, psychometric equations and DDM.

Participants gave written, informed consent and were remunerated £7.50/$15 for their participation plus an unspecified performance-related bonus. Ethical approval was provided by the Research Ethics Committees at UCL and MIT. Both experiments were performed in accordance with the principles expressed in the Declaration of Helsinki. One participant who terminated the experiment early and another who failed all comprehension checks were excluded from the analysis. In addition, we followed the exclusion criteria previously published for this task (Gesiarz et al., 2019): we excluded two participants whose accuracy was below chance (50%) and four who provided responses based only on the first stimulus in over half the trials. Thus, data of 83 participants was included in the analysis (M_{age} = 30.29, SD_{age} ± 12.20; female = 37, male = 46, UCL = 43, MIT = 40). Each participant was randomly assigned to either the threat manipulation group (N = 40, M_{age} = 28.98, SD_{age} ± 11; female = 14, male = 26, UCL = 21 and MIT = 19) or the control group (N = 43, M_{age} = 31.51, SD_{age} ± 13.23; female = 23, male = 20; UCL = 22, MIT = 21).

Manipulation Procedure and Manipulation Check. We followed the exact same threat manipulation as in Garrett et al., (2018). Participants assigned to the 'threat group' were informed that at the end of the experiment they would be required to deliver a speech on a surprise topic, which would be recorded on video and judged live by a panel of staff members. They were shown an adjacent room where chairs and tables were already organized for the panel. This manipulation is a variation of the Trier Social Stress Test (Birkett, 2011) with the key difference being that participants in this task were threatened by the possibility of a stressful social event and completed the main task under anticipation of the threat, but the threat was never executed. Having participants believe the threatening event would take place at the end of the task, rather than before, increased the likelihood that participants' anxiety levels remained high throughout the task. In addition, participants were presented with six difficult mathematical problems that they were asked to try and solve in 30 seconds (s). This exact manipulation procedure was previously executed in our lab, and has been shown to significantly heighten cortisol levels, skin conductance and self-reported state anxiety (Garrett et al., 2018). Our lab has also shown that the manipulation-induced changes in self-reported state anxiety (measured using the Spielberger State Trait Anxiety Inventory, STAI, Marteau & Bekker, 1992) correlated across participants with physiological indicators of stress (Garrett et al., 2018).

Participants assigned to the ‘control group’ were informed that at the end of the experiment they would be required to write a short essay on a surprise topic, which would not be judged. They were then presented with six elementary mathematical problems to solve in 30s. This control manipulation has been shown not to heighten cortisol levels, skin conductance and self-reported state anxiety (Garrett et al., 2018). As a manipulation check, before and after the induction procedure, we asked participants to complete the STAI (Marteau & Bekker, 1992) as a measure of anxiety.

Behavioural Task. After completing the threat/control manipulation, participants played 80 trials of the ‘Factory Task’ (Gesiarz et al., 2019, **see Figure 2.1**). On each trial participants witnessed an animated sequence of televisions (TVs) and telephones passing along a conveyor belt. There were two types of trials: telephone factory trials and TV factory trials. In telephone factory trials, the probability of each item in the animated sequence being a telephone was 0.6, and the probability of each item in the animated sequence being a TV was 0.4. For TV factory trials the proportions were reversed. The trial type was randomly determined with replacement on every trial with an equal probability for each trial type. Participants were tasked with determining whether they were in a telephone factory trial or a TV factory trial. Since the trial type was not directly observable, their means of doing this was through reverse inference over the sequence of objects they were seeing. Participants were free to respond as soon as they wished after initiating the trial and the sequence would continue until they made their decision.

Participants began the game with an endowment of 5000 points. Each 100 points was worth 1 pence/1 cent. One of the two factory types was randomly assigned per participant to be the ‘desirable’ factory type and the other to be the ‘undesirable’ factory type. Participants were informed that each time they visited the desirable factory (desirable state), they would win points, and each time they visited the undesirable factory (undesirable state), they would lose points. We did not specify the exact number of points they would win or lose. Crucially, this bonus was entirely outside of the participants’ control, i.e., it was not affected by the conclusions participants made. Separately, participants were informed that they would earn an unspecified number of points for making a correct conclusion and lose an unspecified number of points for making an incorrect conclusion. We informed participants that the magnitudes of each unspecified bonus/loss were independent of each other, potentially unequal and varied randomly on each trial.

The task was the same as published previously (Gesiarz et al., 2019), except that we jittered the presentation time of the stimuli, so that participants were less likely to have a clear expectation of when the next stimulus would be observed. Due to a technical error this jitter was slightly different across sites (average stimuli presentation time at UCL: M = 657.28 milliseconds (ms), SD ± 1060.73ms; MIT: M = 373.65ms, SD ± 49.69ms). The lag between stimuli was ~150ms.

Trials in which participants made their decision before observing the second object were removed. In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials. Following Gesiarz et al., (2019), we excluded four participants for this reason, as well as a further 72 responses made before seeing the second item.

Training. Prior to playing the task, participants received extensive instructions and were required to answer multiple-choice comprehension check questions on the key points of the task, with the question repeated until they either chose correctly or failed three times, upon which the correct answer was displayed. The comprehension check questions addressed the following key points of how the game worked: that telephone factories mostly produced telephones, but sometimes produced TVs; that the bonus for visiting desirable factories was independent of the decision they made; which factory was their desirable factory; and that trial types (i.e., if they were in a TV or telephone factory) were randomly determined and it was not guaranteed that they would see exactly the same amount of each type of factory. Participants then played a practice session of 20 trials, where the trial type was visibly displayed to them (i.e., if they were in a TV or telephone factory), so they could have prior experience of the outcome contingencies and the trial type distribution.

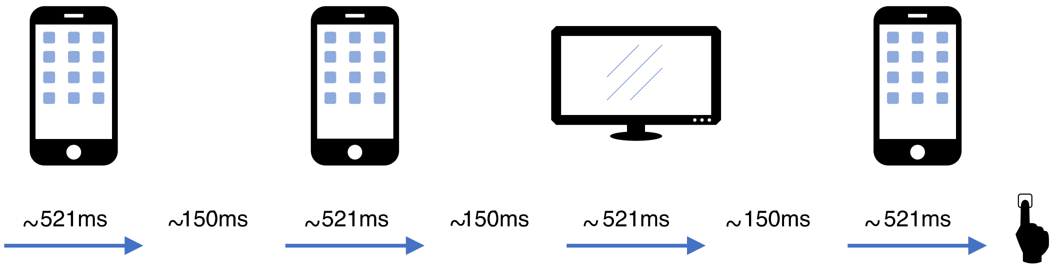

Figure 2.1 The 'Factory Task'. In each trial, participants saw an animated sequence of TVs and telephones passing along a conveyor belt. Their task was to accurately determine whether they were in (a) a telephone factory, i.e., a factory that produces telephones most of the time or (b) a TV factory, i.e., a factory that produces TVs most of the time. They were incentivized for accuracy and could enter their decision whenever they liked. Each participant had 'invested' in one factory. On trials where they happened to be in that (desirable) factory they gained points, on trials in which they happened to be in the other (undesirable) factory they lost points. Notably, this bonus was beyond participants' control and was not affected by the actual decision made. Stimulus presentation time was jittered, so that participants were less likely to have a clear expectation of when the next stimulus would be observed. Stimulus presentation time on average was ~521ms. The lag between stimuli was on average ~150ms.

Statistical Analysis.

Manipulation Check. An independent two-tailed t-test was computed to assess the difference in percentage change in STAI ((post STAI - pre STAI)/pre STAI) between the threat and control group. One-sample t-tests were computed to assess percentage change against zero within each group. All statistical tests conducted in the present article are two-sided. Analysis was conducted using IBM SPSS 27 and R Studio (Version 1.3.1056).

Psychometric Function. We followed the same analysis as in Gesiarz et al., (2019) to relate participants' responses to the strength of evidence they observed. We fitted a psychometric function, using a generalized mixed effects equivalent of a logistic regression, with fixed and random effects for all independent variables. We fitted these functions separately for participants for whom the TV factory was desirable and for whom the TV factory was undesirable, and separately for each group (control, threat).

$$\mathrm{P(TV)} \ = \frac{1}{1 + \mathrm{e}^{-(\beta_1 \mathrm{X} - \beta_0)}}$$

Where *P(TV)* is the probability of a participant indicating they were in a TV factory; *X* is the proportion of TV stimuli out of all stimuli observed in a trial. This variable was centred, thus ranging from 0.5 when all samples were TVs to -0.5 when all samples were phones; *β_0* is the indifference point – reflecting the proportion of TVs required to respond TV 50% of the time. If $\beta_0 = 0$, participants would indicate they were in a TV factory half the time when half the samples were TVs. When β_0 is low, the function will move left and vice versa. β1 is the slope, reflecting by how much the probability of a participant indicating they were in a TV factory increases when the proportion of TVs increases by one unit.

Pieces of evidence gathered. We examined whether the total number of pieces of evidence (TVs + telephones) differed when participants reached an undesirable or desirable conclusion (within-subject variable) and/or depending on group (between-subject variable). Number of pieces of evidence gathered before responding was entered into a mixed 2 (group: control/threat) by 2 (valence of conclusion: desirable/undesirable) ANOVA. We also allowed for an interaction of group and valence of conclusion.

Drift-Diffusion Modelling. Our aim in modelling our task using the Drift-Diffusion framework was to assess how perceived threat impacted the evidence accumulation process. In particular, we wanted to assess (1) whether the evidence accumulation process in the threat and control groups was best represented by the same model or a different model, and (2) whether perceived threat impacted the parameters of the evidence accumulation process in our data.

We implemented and compared four different specifications of a DDM (**see Table 2.1**). The models included the following parameters: (1) t0—amount of non-decision time; (2) α—distance between decision thresholds; (3) z—starting point of the accumulation process; and (4) v—drift rate - the rate of evidence accumulation. Crucially, in models 1 and 3 the starting point was fixed to 0.5, while in models 2 and 4, we allowed the starting point to vary towards one threshold (its value could vary between 0 and 1, thus allowing a valence-dependent starting point bias). In models 1 and 2 with an unbiased drift rate, the parameter was symmetric for desirable and undesirable factories (v and -v). In models 3 and 4 we allowed the drift rate to vary (which we call a valence-dependent drift rate bias) depending upon whether the participant was visiting a desirable factory or an undesirable factory (thus allowing a process bias). In these models we included a term reflecting the difference between drift rates for desirable and undesirable factories (β1factory desirability). 'Factory desirability'—is the true factory visited coded as 1 for desirable factories and 0 for undesirable factories. Positive values indicated a bias towards desirable conclusions, and negative values indicated a bias towards undesirable conclusions. β0 is a constant for the drift rate.

Number	Model	Starting Point (z)	Drift Rate (v)
1.	Valence independent	z = 0.5	v
2.	Valence-dependent starting point	0<z<1	v
3.	Valence-dependent drift rate	z = 0.5	v = β0+β1factorydesirability

4.	Valence-dependent drift rate and starting point	$0<z<1$	$v = \beta0+\beta1 factorydesirability$

Table 2.1. DDM Specification. For each group we ran four models which differed in whether we allowed the starting point to vary (model 2 & 4), whether we included a valence-dependent drift rate bias (model 3 & 4), or neither (model 1).

We used the HDDM software toolbox (Wiecki et al., 2013) to estimate the parameters of our models. The HDDM package employs hierarchical Bayesian parameter estimation, using Markov Chain Monte Carlo (MCMC) methods to sample the posterior probability density distributions for the estimated parameter values. We estimated both group-level parameters as well as parameters for each individual participant. Parameters for individual participants were assumed to be randomly drawn from a group-level distribution. Participants' parameters both contributed to and were constrained by the estimates of group-level parameters. In fitting the models, we used priors that assigned equal probability to all possible values of the parameters. Models were fit to log-transformed response times (RTs) as done previously (Gesiarz et al., 2019), because RTs were non-normally distributed and had a heavy positive skew. Also, since our 'error' RT distribution included relatively fast errors we included an inter-trial starting point parameter (sz) for both models to improve model fit (Ratcliff & Rouder, 1998). We sampled 20,000 times from the posteriors, discarding the first 5,000 as burn in and thinning set at 5. MCMC are guaranteed to reliably approximate the target posterior density as the number of samples approaches infinity. To test if the MCMC converged within the allotted time, we used Gelman-Rubin statistic (Rubin & Gelman, 1992) on 5 chains of our sampling procedure. The Gelman–Rubin diagnostic evaluates MCMC convergence by analysing the difference between multiple Markov chains. The convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter. In each case, the Gelman-Rubin statistic was close to one (<1.1), suggesting that MCMC were able to converge. To assess if the parameters describing the bias in prior and drift rate are significantly different in the control and threat group, we compared

95% Confidence Intervals (CIs) of the parameters' values. Specifically, for each parameter in each group we calculated the 95% CIs. If the 95% CIs for a parameter between groups did not overlap, we consider there to be a significant difference. We also calculated the difference in the posterior distributions and reported the 95% Highest Density Interval (HDI) of the difference. If this HDI did not overlap zero, we considered there to be a meaningful difference between the two groups. HDI testing was conducted in R using *HDInterval* (Meredith & Kruschke, 2016).

In addition, model fits were compared using the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002), which is a generalization of the Akaike Information Criterion (AIC) for hierarchical models. The DIC is commonly used when the posterior distributions of the models have been obtained by MCMC simulation (Gamerman & Lopes, 2006). It allows one to assess the goodness of fit, while penalizing for model complexity.

To validate the winning model, we used each group's parameters obtained from participants' data to simulate log-transformed RTs and responses separately for the threat and control group. We used the exact number of participants, total number of trials and trial structure as in the experiment. Simulated data was then used to (1) perform model recovery analysis and (2) to compare the pattern of participants' response to the pattern of simulated responses, separately for each group. We sampled 2,000 times from the posteriors, discarding the first 500 as burn in. Simulation and model recovery analysis were performed using the HDDM software toolbox (Wiecki et al., 2013).

Proportion of correctly identified factories. We examined whether group (between-subject variable) and valence of factory visited (within-subject variable) affected the proportion of correctly identified factories as desirable or undesirable. Proportions were calculated for each participant and then entered into a mixed 2 (group: control/threat) by 2 (valence of factory: desirable/undesirable) ANOVA. We also allowed for an interaction of group and valence of factory. We compared the pattern of results obtained from participants' real data to those obtained from the simulated data.

2.4 Results

Threat manipulation was successful. The manipulation was successful in inducing perceived threat. Participants in the threat group reported a significantly larger increase in anxiety as a result of the manipulation (increase in STAI score after the manipulation relative to before M = 40.815%, SE = 5.928, t(39) = 6.885, p < 0.001, Cohen's d = 1.089), compared to those in the control group, who in fact showed a reduction in anxiety (M = -5.742%, SE = 1.257, t(42) = 4.568, p < 0.001, Cohen's d = 0.697, difference between the two groups: t(81) = 7.943, p < 0.001, Cohen's d = 1.715, **Figure 2.2**) an effect often observed in control participants, who tend to relax as they learn more about the task at hand (Garrett et al., 2018).

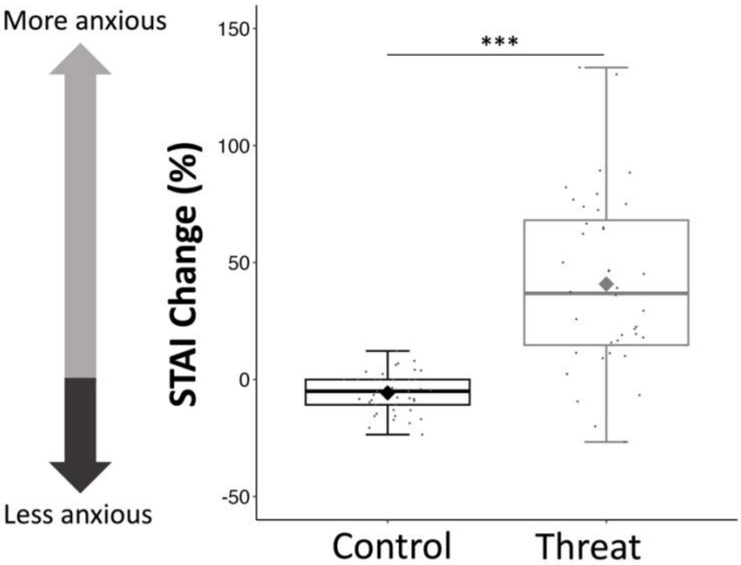

Figure 2.2. Threat manipulation was successful. Participants in the threat group became significantly more anxious after the manipulation than in the control group. Data are plotted as boxplots for each condition, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 x interquartile range. Diamond shape indicates

the mean percentage change in STAI per experimental group. Individuals' percentage STAI change are shown separately as grey dots. ***p < 0.001

Under threat participants required weaker evidence to conclude they are in an undesirable factory. We first examined whether perceived threat alters the strength of evidence participants required to reach desirable and undesirable conclusions. To that end, we fit a psychometric function to the data which relates the percentage of TVs observed on a trial (i.e., the strength of the evidence to judge a factory as a TV factory) to participants' conclusion about whether they were visiting a TV or telephone factory. This was done separately for participants for whom the TV factory was desirable and for whom it was undesirable in the threat and control group.

As observed in **Figure 2.3a**, under perceived threat the psychometric function of participants for whom the TV factory was undesirable (solid orange line) was shifted left compared to controls (dotted orange line). This means that compared to control, under threat participants required a smaller proportion of TVs to be observed before reaching the conclusion that they were in a TV factory when the TV factory was undesirable (indifference parameter was higher for the threat group: β0 = 0.113, 95% CI [-0.185, 0.411], than controls: β0 = -0.582, 95% CI [-0.964, -0.20], Cohen's d = 0.61, **Figure 2.3a**). No such difference is observed when the TV factory is desirable. Participants in both groups required an equal proportion of TVs to be observed before reaching the conclusion that they were in a TV factory. This can be seen in **Figure 2.3b** where the psychometric functions for threat and control participants overlap (indifference parameter was not different for the threat group: β0 = 0.277, 95% CI [-0.068, 0.622] and control group: β0 = 0.23, 95% CI [-0.079, 0.53], Cohen's d = 0.05, **Figure 2.3b**).

While participants in the control group required weaker evidence to conclude they were in a desirable factory than an undesirable factory (replicating previous findings from Gesiarz et al., 2019), this difference was abolished under perceived threat. This can be observed where the psychometric function of control participants for whom the TV factory was desirable (dotted blue line, **Figure 2.3c**) is shifted to the left of control participants for whom the TV factory was undesirable (dotted orange line, **Figure 2.3c**; indifference parameter was greater for desirable factory: β0 = 0.23, 95% CI [-0.079, 0.53], than undesirable: β0 = -0.582, 95% CI [-0.964, -0.20], Cohen’s d = 0.694, **Figure 2.3c**), while for participants in the threat group they overlap (indifference parameter when the TV factory was desirable β0 = 0.277, 95% CI [-0.068, 0.622] and undesirable β0 = 0.113, 95% CI [-0.185, 0.411], Cohen’s d = 0.16, **Figure 2.3d**).

As expected, both in the threat and control group the greater the proportion of TVs in a trial the more likely participants were to judge the factory as a TV factory (control: TV factory desirable: β1 = 25.55, 95% CI [23.20, 27.90], TV factory undesirable: β1 = 24.943 [15.623, 34.262], Cohen’s d = 0.023, **Figure 2.3c**; threat: TV factory desirable: β1 = 27.79, 95% CI [19.169, 36.411], TV factory undesirable: β1 = 23.043, 95% CI [17.308, 28.778], Cohen’s d = 0.2, **Figure 2.3d**).

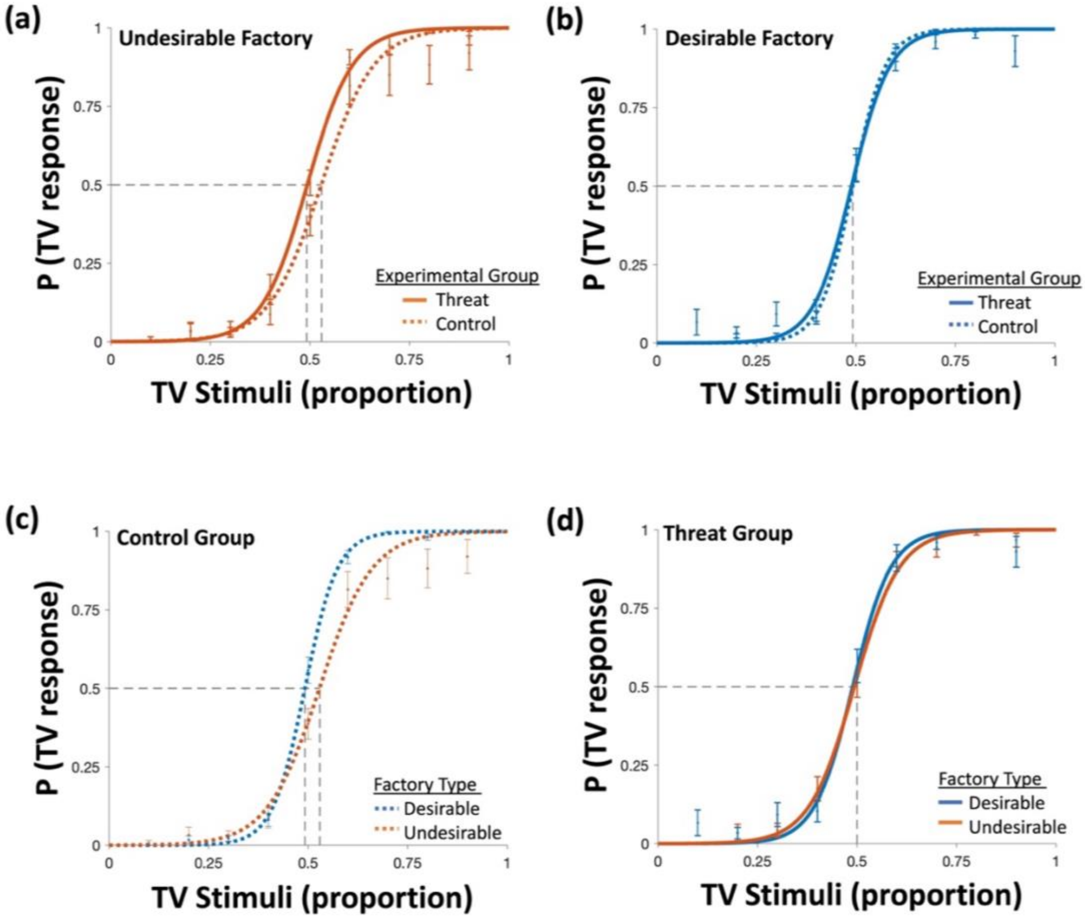

Figure 2.3. Under threat weaker evidence is required to reach undesirable conclusions. Fitted psychometric functions for data of the threat group (solid line) and control group (dotted line). Y axis shows the proportion of times participants indicated they were in a TV factory as a function of the proportion of TV stimuli they observed in a trial prior to making a decision (x axis). In blue is the data of participants for whom the TV factory was the desirable factory. In orange is the data of participants for whom the TV factory was the undesirable factory. **(a)** Compared to the control group, under perceived threat participants required a smaller proportion of TVs to be observed before reaching the conclusion that they were in a TV factory, when the TV factory was undesirable. This can be seen as the solid line (threat group) is shifted left relative to the dotted line (control group). **(b)** No such difference is observed when the TV factory is desirable. **(c)** Participants in the control group required a smaller proportion of TVs to be observed before reaching the conclusion that they were in a TV factory, when the TV factory was desirable than undesirable. This is seen as the blue line (desirable) is shifted left relative to the orange line (undesirable). **(d)** This difference is abolished under perceived threat. Error bars show SE at given level of proportion of TVs observed (error bars for threat group are indicated by ‘x’ at the centre of the error bar). Grey dashed line indicates point of indifference – i.e., how much evidence is needed for participants to say ‘TV’ half the time.

Note, that the total number of pieces of evidence (TVs + telephones) did not differ when participants reached an undesirable or desirable conclusion (F(1,81) = 1.363, p = 0.247, partial η2 = 0.017), nor did it differ as a function of perceived threat (F(1,81) = 0.376, p = 0.542, partial η2 = 0.005), neither was there an interaction between these two factors (F(1,81) = 0.023, p = 0.879, partial η2 = 0.00). Rather, as shown above, it is the proportion of evidence (which signifies the strength of the evidence) needed to reach a conclusion that differed as a function of perceived threat and valence.

Thus far our analysis suggests that perceived threat led to a reduction in the strength of the evidence needed to reach undesirable conclusions, even though the cause of the threat (anticipating a negative social situation) had nothing to do with the task at hand. We next sought to identify the precise computational factor(s) affected by perceived threat during evidence accumulation.

Under threat the drift rate towards undesirable conclusions is greater. Computationally, there are at least two different ways by which perceived threat can lower the strength of evidence needed to reach undesirable conclusions. First, threat may alter the starting point of the accumulation process. That is, if under perceived threat participants are a priori more likely to believe they are in an undesirable state relative to controls, then weaker evidence will be needed to reach that conclusion. Alternatively, perceived threat can enhance the weight given to each piece of undesirable evidence relative to control. This again will lead to weaker evidence needed to reach an undesirable conclusion.

To tease apart these possibilities we modelled the responses as a drift-diffusion process (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013) with the following parameters: (1) t0—amount of non-decision time; (2) α—distance between decision thresholds; (3) z—starting point of the accumulation process; and (4) v—drift rate – i.e., the rate of evidence accumulation (**see Methods for details**). Crucially, in models 1 and 3 the starting point was fixed to 0.5, while in models 2 and 4 we allowed the starting point to vary towards one threshold (thus allowing a starting point bias). In models 3 and 4 we allowed the drift rate to vary (which we call a valence-dependent drift rate bias) depending upon whether the participant was visiting a desirable factory or an undesirable factory (thus allowing a process bias).

The DIC, a generalization of the AIC for hierarchical models, was calculated for each model **(Table 2.2)**. The DIC scores indicated that Model 4 (the valence-dependent model), which included a valence-dependent starting point and drift rate, outperformed all other models for both the threat and the control group. As can be observed in **Figure 2.4,** while for the control group the valence-dependent model was clearly a better fit than the valence-independent model (replicating previous findings from Gesiarz et al., 2019), for the threat group the advantage in terms of fit was modest.

Number	**Model**	**Starting Point (z)**	**Drift Rate (v)**	**DIC (Control)**	**DIC (Threat)**
1.	Valence independent	z = 0.5	v	11373.43	7761.38
2.	Valence-dependent starting point	0<z<1	v	11343.42	7757.88
3.	Valence-dependent drift rate	z = 0.5	v = β0+β1factorydesira-bility	11322.83	7758.58
4.	Valence-dependent drift rate and starting point	0<z<1	v = β0+β1factorydesira-bility	11306.45	7744.82

Table 2.2 DDM Model Fits. For each group we ran four models which differed in whether we allowed the starting point to vary (model 2 & 4), whether we included a valence-dependent drift rate bias (model 3 & 4), or neither (model 1). DIC scores show goodness of fit, with lower numbers indicating better fit.

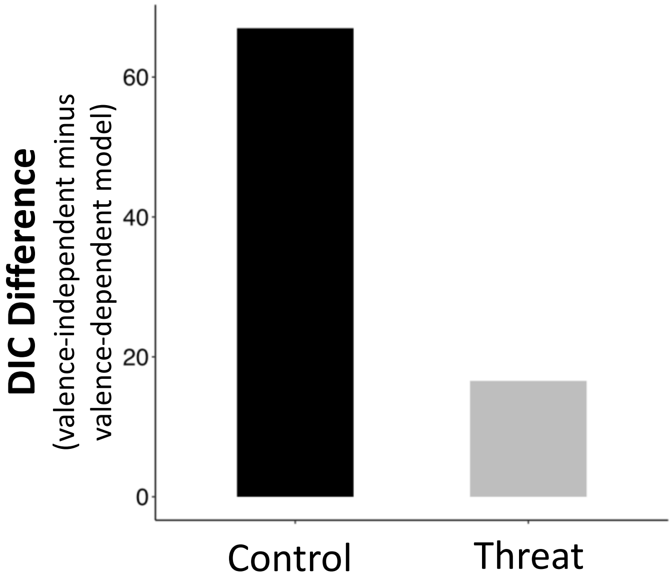

Figure 2.4. Difference in fit between winning valence-dependent model and valence-independent model as a function of perceived threat. The Y axis shows the difference in DIC scores between the valence-independent model and the winning valence-dependent models for the control group (dark grey) and threat group (light grey).

We next examined which of the accumulation parameters were affected by perceived threat. To that end, we calculated 95% CIs of each parameter for each group. If the 95% CIs do not overlap, we infer a significant difference between the two groups.

As observed in **Table 2.3** and **Figure 2.5**, only one element in the accumulation process was significantly altered by perceived threat: the valence-dependent drift rate bias. The drift rate bias is the difference in drift rates between desirable and undesirable factories, the greater the bias the greater the drift rate for desirable factories relative to undesirable ones. As can be observed in **Figure 2.5e** the valence-dependent bias in drift rate in the control group was significantly greater than in the threat group (control: β1 = 0.17 [0.07, 0.27]; threat: β1 = -0.08 [-0.20, 0.04]). For controls the bias in drift was significantly *positive* (95% CI does not include zero: β1 = 0.17 [0.07, 0.27]), leading to a drift rate that was more than double when participants were in the desirable factory ($v_{desirable}$ = 0.63) than undesirable factory ($v_{undesirable}$ = 0.46). In contrast, under perceived threat the bias in drift rate was numerically negative and not significantly different from zero (95% CI includes zero: β1 = -0.08 [-0.20, 0.04]), leading to a drift rate that was numerically and non-significantly larger when participants were in the undesirable factory ($v_{undesirable}$ = 0.63) than when they were in the desirable factory ($v_{desirable}$ = 0.55). We also corroborate these results using 95% HDI comparisons (**see Appendix 7.1 Supplementary Table 2.1 for HDI Comparisons**).

Estimate (Experimental Data)	**Control** [95% CI]	**Threat** [95% CI]
Distance between Decision Thresholds (α)	2.666 [2.491, 2.85]	2.474 [2.335, 2.62]
Non-Decision Time (t0)	7.546 [7.383, 7.714]	7.488 [7.334, 7.647]
Starting Point (z)	0.48 [0.47, 0.505]	0.516 [0.498, 0.535]
inter-trial starting point parameter (sz)	0.182 [0.066, 0.266]	0.185 [0.056, 0.276]
Drift Rate (β0)	0.456 [0.369, 0.548]	0.631 [0.544, 0.724]
Drift Rate Bias (β1)	0.174 [0.079, 0.268]	-0.085 [-0.201, 0.04]

Table 2.3. Parameter estimates of the evidence accumulation process. Displayed are the model estimates from the winning model for the control and threat groups. These include distance between decision thresholds (α), non-decision time (t0), starting point (0<z<1), inter-trial starting point parameter (sz), constant drift rate (β0) and drift rate bias (β1). The latter is the term reflecting the additional weight added to the drift rate as a function of factory desirability. Positive values indicate a bias towards desirable conclusions, and negative values indicate a bias towards undesirable conclusions. [CI].

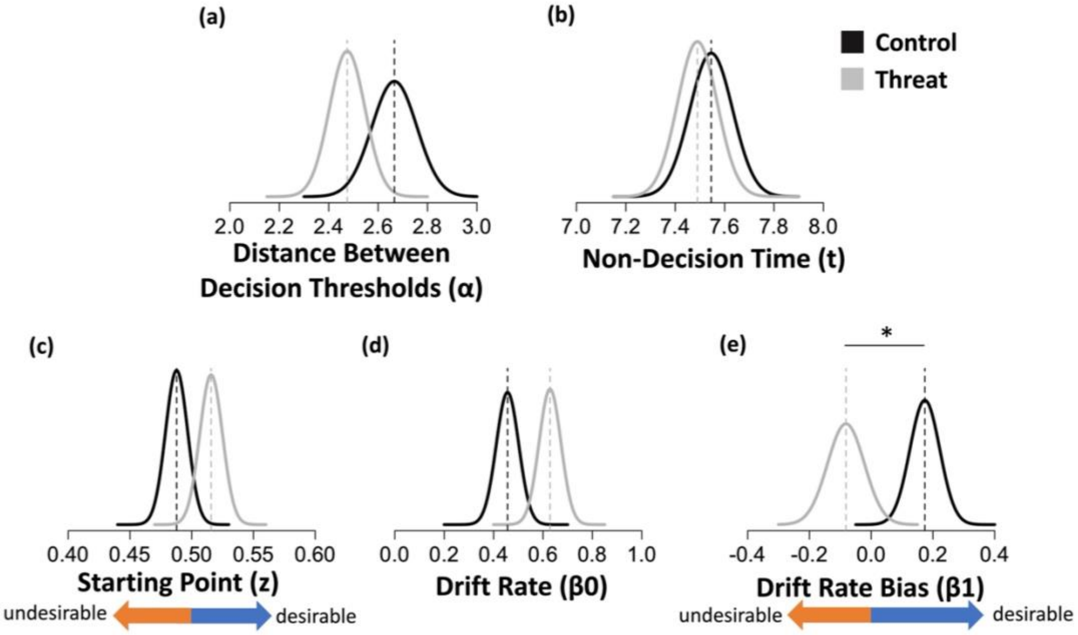

Figure 2.5. Under threat the valence-dependent drift rate bias is abolished. Displayed are the posterior distributions of parameter estimates for the threat group (light grey) and the control group (black). No significant difference is observed between groups for estimates of **(a)** distance between decision thresholds, **(b)** non-decision time, **(c)** starting point, and **(d)** drift-rate constant. **(e)** In contrast, a significant difference is observed for the valence-dependent drift rate bias. In the control group the bias indicates a significantly larger drift rate towards desirable than undesirable conclusions. This bias is corrected for under perceived threat and is numerically inverse (that is the bias is non-significantly negative under perceived threat but significantly positive in the control group). *indicates significant difference between parameters in the threat and the control group (i.e., CIs do not overlap). Dashed vertical lines indicate group mean.

We simulated data using group parameters from the threat and control group separately (**see Methods for details**). We first examined if the model parameters could be successfully recovered based on the simulated data. To do so the valence-dependent model was fit to simulated data, in the same way as for the experimental data. We sampled 2,000 times from the posteriors, discarding the first 500 as burn in. As shown in **Table 2.4** model parameters could be successfully recovered based on the simulated data. Additionally, we examined if the simulated data reproduced the same behavioural pattern of results as the participants' data. This was indeed the case (see **Figure 2.6d,** detailed explanation below).

Estimate (Simulated Data)	**Control** [95% CI]	**Threat** [95% CI]
Distance between Decision Thresholds (α)	2.667 [2.627, 2.722]	2.483 [2.434, 2.533]
Non-Decision Time (t0)	7.554 [7.51, 7.596]	7.421 [7.388, 7.458]
Starting Point (z)	0.505 [0.477, 0.532]	0.509 [0.483, 0.535]
inter-trial starting point parameter (sz)	0.332 [0.199, 0.43]	0.146 [0.008, 0.306]
Drift Rate (β0)	0.44 [0.383, 0.512]	0.68 [0.596, 0.769]
Drift Rate Bias (β1)	0.146 [0.075, 0.216]	-0.14 [-0.229, -0.049]

Table 2.4. Recovered Parameter estimates of the evidence accumulation process based on simulated data. Displayed are the winning model estimates recovered from the simulated data for the control and threat groups. These include distance between decision thresholds (α), non-decision time (t0), starting point (0<z<1), inter-trial starting point parameter (sz), constant drift rate (β0) and drift rate bias (β1). The latter is the term reflecting the additional weight added to the drift rate as a function of factory desirability. Positive values indicate a bias towards desirable conclusions, and negative values indicate a bias towards undesirable conclusions. [CI].

As DDM parameters are computed partially based on participants' responses we expected the model-based valence-dependent drift rate bias to correlate across individuals with a valence-dependent bias in judgments. Indeed, across participants there was a strong positive correlation between valence-dependent drift rate bias and the proportion of correctly identified desirable factories minus the proportion of correctly identified undesirable factories (threat group: r = 0.802, p < 0.001, **Figure 2.6a,** control: r = 0.918, p < 0.001, **Figure 2.6a)**, which we term 'valence-dependent judgement bias'. Individuals with a greater drift rate towards desirable than undesirable conclusions were more likely to correctly identify desirable factories as desirable when they observed them, than undesirable factories when they observed them. In contrast, the starting point bias did not correlate with the valence-dependent bias in judgements in the threat group (r = 0.223, p = 0.191), but did in the control group (r = 0.517, p < 0.001). In the latter, a larger starting point bias was related to the proportion of correctly identified desirable factories minus the proportion of correctly identified undesirable factories (**Figure 2.6b**).

As we have already shown that participants in the control group had a greater drift rate bias than those under perceived threat, it follows that they would also show a greater valence-dependent judgement bias. This is indeed what we observed. Entering the proportion of correctly identified factories as desirable or undesirable into a mixed 2 by 2 ANOVA with valence of factory (within-subject), group (between-subject) and their interaction, revealed a group by valence interaction (F(1,81) = 3.868, p = 0.049, partial η2 = 0.05, **Figure 2.6c**) as well as a main effect of factory valence (F(1,81) = 5.045, p = 0.027, partial η2 = 0.06) and no main effect of group (F(1,81) = 0.002, p = 0.985, partial η2 = 0.00). To tease apart the interaction we followed up with pairwise comparisons. This revealed that participants in the control group were less likely to correctly categorize undesirable factories (M = 0.727, SE = 0.022) than desirable factories (M = 0.80, SE = 0.015; t(42) = 3.099, p = 0.003, Cohen's d = 0.473). In contrast, under perceived threat the effect of valence dissappeared (t(39) = 0.19, p = 0.85, Cohen's d = 0.03); participants were just as likely to correctly categorize undesirable factories (M = 0.76, SE = 0.02) as they were desirable factories (M = 0.765, SE = 0.019). This suggests that under perceived threat the valence-dependent judgement bias is abolished.

We conducted the same analysis on our simulated data and find that it nicely reproduced the behavioural pattern of results (**Figure 2.6d, see Appendix 7.1 Supplementary Table 2.2**).

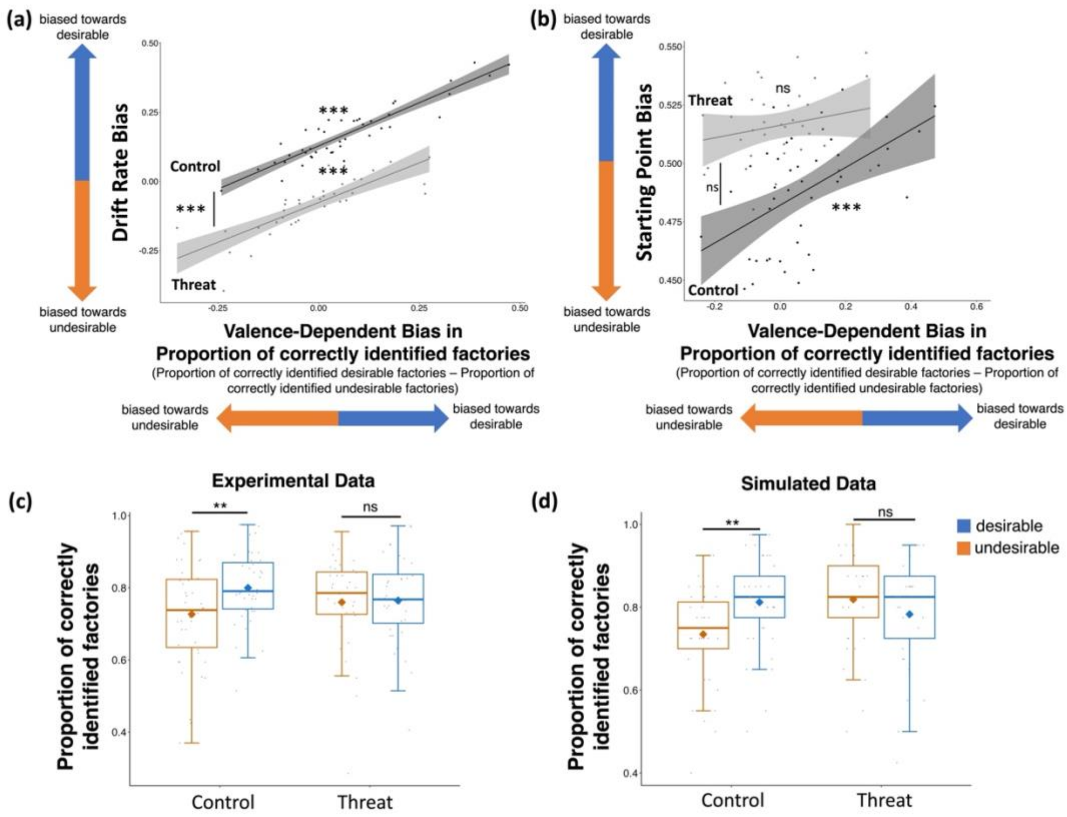

Figure 2.6. Threat-induced change in valence-dependent drift rate bias is expressed as a valence-dependent change in the proportion of correctly identified factories. (a) A positive relationship is observed between the valence-dependent drift rate bias (Y axis) and the valence-dependent bias in the proportion of correctly identified factories (X axis). Individuals with a greater drift rate towards desirable than undesirable conclusions are more likely to correctly categorize desirable than undesirable factories. This is true both for both participants in the control group (dark grey) and participants in the threat group (light grey). For those in the control group, the regression line is above that of participants in the threat group, which is due to the fact that their drift rate bias is significantly greater. The regression line for controls is also shifted to the right which indicates a significantly greater valence-dependent judgement bias. **(b)** By contrast we did not observe a relationship between the starting point bias (Y axis) and the valence-dependent bias in the proportion of correctly identified factories (X axis) in the threat group (light grey). A positive correlation was observed in the control group (dark grey). In the control group individuals with a large starting point bias were more likely to correctly identify desirable than undesirable factories. While the line for the threat group is above that of the control group this is not a significant difference. **(c)** Controls are less likely to correctly categorize undesirable factories (orange) than desirable factories (blue), while this is not the case for participants in the threat group. (d) Simulated data based on model parameters reproduced these findings. Data are plotted as boxplots for each condition, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers

indicate 1.5 x interquartile range. Diamond shape indicates the mean. $**p < 0.01$, ns = not significant. Clouds represent CIs.

2.5 Discussion

The findings show that perceived threat has a profound effect on the process by which evidence is accumulated. In particular, it leads to a reduction in the strength of the evidence needed to reach undesirable conclusions. Relative to controls, participants under perceived threat required a smaller proportion of negative stimuli to be observed before reaching an undesirable conclusion. In contrast, there was no between-group difference in the strength of evidence accumulated before reaching a desirable conclusion. We found this to be true despite the fact that the cause of the threat (anticipating a socially stressful event) was unrelated to the task performed (judging whether more phones or more TVs were observed).

Computationally, there are different mechanisms by which perceived threat can lower the strength of evidence needed to reach undesirable conclusions. First, under threat participants may be a priori more likely to believe they are in an undesirable state relative to controls leading to weaker evidence needed to reach that conclusion. Another possibility is that perceived threat can selectively increase the rate of negative evidence accumulation (drift rate) relative to control. This again will lead to weaker evidence required to reach an undesirable conclusion. To tease apart these possibilities we modelled responses as a drift-diffusion process (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013). We found support for the latter. Specifically, perceived threat altered only one feature of the accumulation process: the relative drift rate towards desirable and undesirable conclusions (the ‘valence-dependent drift rate bias’). For controls the bias in drift rate was positive – the rate of evidence accumulation was greater towards desirable than undesirable conclusions (as observed before in Gesiarz et al., 2019). Under threat, however, the bias disappeared due to the drift rate towards undesirable conclusions increasing.

These results fit with previous suggestions that perceived threat shifts neural valuation from desirable to potential aversive outcomes in risky choice (Engelmann et al., 2015), directs attention towards negative stimuli (Macatee et al., 2017) and leads to greater impact of such stimuli on belief updating (Garrett et al., 2018). Indeed, it is possible that the effect of perceived threat on the rate of negative evidence accumulation is partially due to increased attention towards negative stimuli. The current findings go beyond these previous demonstrations to illuminate the effects of perceived threat on the process of sequential evidence accumulation and show that weaker evidence is needed to reach undesirable conclusions under threat.

Here, we show a causal link between perceived threat and evidence accumulation in healthy individuals. It is interesting, however, to consider how our findings may be related to evidence accumulation in individuals with affective disorders, as these are often triggered by stressful events and/or characterized by high anxiety. With regards to individuals with high trait anxiety, a processing advantage for threatening words has been previously reported (White et al., 2010). While that study was correlational and thus could not determine whether anxiety caused the changes to the drift rate and/or vice versa, our results support the notion that anxiety can in fact alter the drift rate towards undesirable conclusions, even if the anxiety is short-lived rather than chronic. With regards to individuals with anxiety and mood disorders, one study (Aylward et al., 2019) found a lower drift rate towards desirable conclusions compared to healthy individuals. Interestingly, the latter study did not detect any effects of induced threat, which may be due to the fact that the task used in that study (as well as in all the above-mentioned studies) unlike ours, was a non-sequential perceptual decision-making task. The process by which pieces of evidence are accumulated over time may be especially impacted by perceived threat.

Our study suggests that evidence accumulation is a flexible process which quickly adjusts to the environment. In particular, the findings show that perceived threat leads to a valence-dependent change to the accumulation process, even when the evidence is not directly related to the cause of the threat. An increased rate of negative evidence accumulation can then enhance the probability of taking precautionary action to avoid aversive consequences. As aversive outcomes can be more severe and frequent in threatening environments, such generalization can be, on average, adaptive. However, in individuals who are hypersensitive to threat and/or falsely perceive situations as threatening, such as those suffering from anxiety and depression, an increased rate of negative evidence accumulation could be maladaptive. This is because such an increased rate can produce overly pessimistic predictions, which induce stress and anxiety, further worsening symptoms.

Chapter 3: Futile Rewards: Why Accuracy Incentives Fail to Reduce Biased Evidence Accumulation

3.1 Abstract

Intuitively, we assume that financially incentivizing individuals to provide a correct response will enhance the accuracy of their decisions. Here, we show that accuracy incentives fail to reduce a well-known bias in which people reach desirable (over undesirable) conclusions (= desirability bias), because the two operate on orthogonal aspects of the evidence accumulation process. Over three experiments, participants (n = 235) completed a perceptual evidence accumulation task in which they had to determine whether they were in a desirable state, which was associated with greater rewards, or an undesirable state. In some trials they were also financially incentivized for correct responses. Results show that while accuracy incentives led participants to take more time in reaching a conclusion, they did not impact participants' desirability bias. Fitting the data to an evidence accumulation model revealed that while accuracy incentives led to an increase in the distance between the decision thresholds, the desirability bias was associated with greater weight on desirable relative to undesirable evidence (drift rate bias). These results suggest that the desirability bias is likely unconscious.

3.2 Introduction

Humans demonstrate an impressive ability to understand the world around them. At the same time, however, people also exhibit a host of systematic errors in judgement (known as biases and heuristics, Gigerenzer & Gaissmaier, 2011). These errors can lead to poor and costly decisions in domains ranging from finance to health (Dunning et al., 2004; Shefrin, 2015). There is thus a clear need to identify ways to mitigate such systematic errors. One obvious solution is to financially incentivize people to form more accurate beliefs in situations where biases and heuristics are common. But would such incentives work? The answer may depend on the nature of the error.

One category of systematic errors, known as heuristics, are cognitive shortcuts used to make efficient and effortless decisions (Gigerenzer & Gaissmaier, 2011). These are not errors that people are motivated to make, but rather the result of ‘thoughtless’ processing. Past studies suggest financial incentives are generally successful at reducing such errors. For instance, financial incentives have been shown to reduce the tendency to falsely believe large-number ratios (e.g., 30/100) convey a higher probability than equivalent small-number ratios (3/10; Dale et al., 2007; Lefebvre et al., 2011). Incentives also reduce the anchoring effect, that is the tendency for decisions to be disproportionally influenced by an initial piece of observed information (Meub et al., 2013; Wright & Anderson, 1989; but see Epley & Gilovich, 2005; Enke et al., 2023), and the conjunction fallacy, which describes the tendency to incorrectly consider multi-attribute hypotheticals that are more specific (e.g., “The dog surfs and wears a hat”) to be more probable than singular hypotheticals (e.g., “The dog surfs", Zizzo et al., 2000; for a meta-analysis see Yechiam & Zeif, 2023; but see Charness et al., 2010). Thus, there is some evidence that incentives can increase (‘slow’) rational thinking and reduce the reliance on cognitive shortcuts.

Here, we focus on a distinct, second, category of systematic errors - motivational biases. These are errors that occur not due to ‘cognitive laziness’, but because the individual prefers one belief over the other (Montibeller & Von Winterfeldt, 2015). Only a handful of studies have examined whether incentives can overcome such motivational biases. Almost all have focused on the partisan bias, which is the tendency to process, interpret, and favour information in a way that aligns with one’s political ideology (Van Bavel & Pereira, 2018). Indeed, monetary incentives have been shown to reduce the impact of the partisan bias on judging the accuracy of political statements (Prior et al., 2015; Rathje et al., 2023), however not on non-political advice-taking (Zhang and Rand, 2023). Beyond the partisan bias, financial incentives also failed to mitigate the bias to interpret a perceptual stimulus as predicting the absence of an upcoming shock (Engelmann et al., 2019).

Thus, while muddy, the literature seems to suggest that monetary incentives are relatively unreliable in reducing motivational biases. Here, we propose and test a possible explanation. We hypothesize that financial incentives do indeed lead individuals to invest cognitive effort in reaching accurate conclusions, by for example accumulating more evidence. These efforts may indeed be fruitful when the cause of the error is fast, careless, processing. However, when the cause of the error is an unconscious bias in how information itself is processed (for example as when people put greater weight on desirable than undesirable evidence resulting in a desirability bias, as shown in Gesiarz et al., 2019; Globig et al., 2021), they will have little effect. In other words, if financial incentives alter a feature of the decision process that is orthogonal to the one the bias works on, they will fail.

To test this hypothesis we conducted three experiments, where participants (total N = 236) played a modified random-dot task. In the task, participants observed a cloud of moving dots. Their goal was to determine whether most of the dots were moving left or right. On some trials, they were rewarded for correct responses, while on other trials there was no reward for correct responses. Importantly, participants were told that one of the two directions was desirable. That is whenever the dots moved in that direction, they could receive an additional bonus. Crucially, participants had no control over the direction of the dots; their task was simply to detect it. In Experiment 1 and its replication, the reward for correct responses on incentivized trials was equal to this bonus, and in Experiment 2 it was five times greater. Thus, in all experiments, to maximize reward across the task, it was in participants’ interest to be as accurate as possible.

We fit a DDM to our data. This models the process of noisy evidence accumulation toward either of two decision thresholds from a given starting point (Ratcliff & Rouder, 1998; Voss et al., 2013) and enables us to determine which elements are altered by accuracy incentives. Previous work from our lab shows, that the motivation to form desirable conclusions alters the rate at which desirable evidence is accumulated (Gesiarz et al., 2019; Globig et al., 2021; see also Leong et al., 2019). We speculate that if accuracy incentives fail to mitigate biased evidence accumulation, this could be because they affect a different element of the accumulation process. For example, incentives may increase the distance between decision thresholds, thus making participants more cautious, but may not impact the relative rate at which desirable evidence is accumulated. Knowledge of the precise decision-making features that financial accuracy incentives impact, can shed light on when and why they may fail to reduce motivational biases.

3.3 Methods

Experimental Design

Participants (Experiment 1):

Seventy participants residing in the United States (US) completed the task on *Prolific Academic* (www.prolific.com). Participants received £7.50 per hour for their participation in addition to a performance-related bonus. Ethical approval was provided by the Research Ethics Committee at University College London and all participants gave informed consent. All experiments were performed in accordance with the principles expressed in the Declaration of Helsinki. Participants who failed the comprehension checks at the beginning of the experiment more than twice were not allowed to participate. We also adopted the following exclusion criteria: response times (RTs) faster than 200ms were discarded from further analysis, as recommended in previous literature (Ratcliff & Tuerlinckx, 2002; Rollwage et al., 2020; Wiecki et al., 2013). In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials (as in Gesiarz et al., 2019; Globig et al., 2021). Based on this we removed 1 participant and a further 64 trials from Experiment 1. Thus, data of 69 participants were analysed (Experiment 1: N = 69 M_{age} = 35.319, SD_{age} ± 10.222; female = 31, male = 38). Sample size was calculated based on a pilot study. Calculations were performed using g*Power (Faul et al., 2009) to achieve power of 0.8.

Participants (Replication):

Seventy-three participants residing in the US completed the task on *Prolific Academic*. Participants received £7.50 per hour for their participation in addition to a performance-related bonus. RTs faster than 200ms were discarded from further analysis (**see Participants Experiment 1 for details**). Forty-nine trials were discarded for this reason. No participants were removed. Thus, data of 73 participants were analysed (Replication: N = 73; M_{age} = 35.699, SD_{age} ± 9.253; female = 37, male = 36).

Participants (Experiment 2):

Ninety-three participants residing in the US completed the task on *Prolific Academic*. Participants received £7.50 per hour for their participation in addition to a performance-related bonus. RTs faster than 200ms were discarded from further analysis (**see Participants Experiment 1 for details**). In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials. Based on this we removed 1 participant from Experiment 2 and a further 134 trials. Thus, data of ninety-two participants were analysed (Experiment 2: N = 92; M_{age} = 29.396, SD_{age} ± 5.702; female = 44 male = 48).

Behavioural Task (Experiment 1 & Replication):

In each experiment, participants played a modified forced choice random dot kinetogram (RDK) task (Newsome & Pare, 1988; **see Figure 3.1**). The task consisted of four blocks with 60 trials each. On each trial participants observed a cloud of white dots (number of dots = 100, dot radius = 2°, dot life = 5 frames) moving within a circular aperture on a black background. The direction of motion of the clouds was either right or left along the horizontal meridian. Each set was replotted one aperture later in which a subset of dots, determined by the noise level, was offset from their previous location towards the target movement direction, and another subset was offset in the opposite direction, whereas the rest was replotted randomly. Participants had to judge whether the majority of the dots were moving in the right or left direction. They were free to respond as soon as they wished. The dot motion would either continue until they made their choice or up to 5s. Across trials we varied whether participants were incentivized for accuracy (i.e., paid for correct responses; reward = $5 vs no reward = $0) and the level of noise of the dot motion (i.e., low noise = 0.256 motion-coherence vs high noise = 0.064 motion-coherence). Participants were informed that at the end of the experiment, one trial would randomly be selected for pay off. If in this trial participants were incentivized for accuracy, and they responded correctly they would receive $5. In addition to this, in each block, one direction was assigned to be the desirable direction, and the other to be the undesirable direction. Participants were told that if the dots in the trial selected for pay off moved in the desirable direction, they would receive an additional $5 bonus. Crucially, this bonus was entirely outside of their control, i.e., it was not affected by their responses. Desirability was counterbalanced across participants and varied across blocks, such that in two blocks participants would receive a bonus if the dots moved to the left, and in the remaining two blocks participants would receive the bonus if the dots moved to the right. Trial type was randomly determined with replacement on every trial with an equal probability for each trial type.

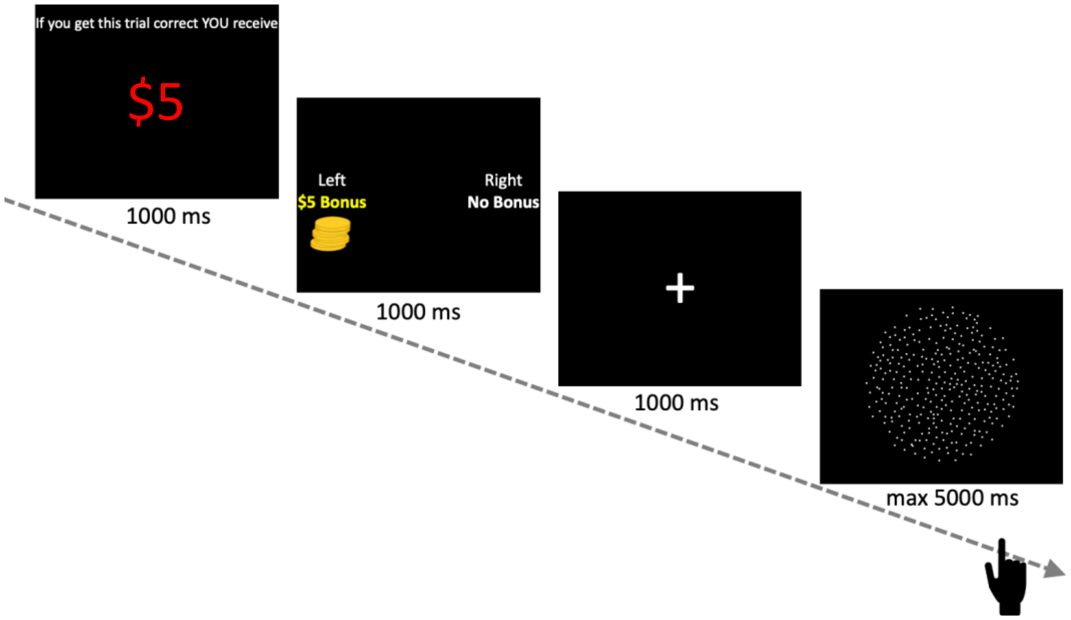

Figure 3.1. The Random Dot Kinetogram Task (Experiment 1 & Replication). In each trial, participants saw an animated cloud of dots moving

either rightward or leftward within a circular aperture. Their task was to determine the direction of the majority of the dots. Across trials, we varied the level of noise of the dot motion (i.e., % of dots moving in the same direction) and whether participants were incentivized for accuracy. The reward for correct responses on incentivized trials was $5. Participants were informed that at the end of the experiment one trial would randomly be selected for pay off. In each block, one direction was assigned to be desirable. On trials where the dots happened to move in that direction, participants could receive an additional $5 bonus, if this trial was selected for pay off. Notably, this bonus was beyond participants' control and was not affected by the actual decision made. Participants were told which direction was desirable at the start of each block. In each trial, they were first informed about whether the trial was incentivized for accuracy. Next, they were presented with a reminder about which direction was desirable. This was followed by a fixation cross. Each of these was presented for 1000ms. Afterwards the random-dot stimulus was displayed for a maximum of 5000ms or until button press.

For example, imagine that at the end of the experiment, a trial is selected in which there is no reward for correct responses. The trial is in a block where the right side is considered desirable. If a participant correctly deduces that the dots move to the right, they receive no performance-related reward despite answering correctly, because there is no reward for correct responses in this trial. However, they do receive a bonus ($5) because the dots are in fact moving in the desirable direction. The task was coded using *JsPsych* and *Javascript*. The task for the replication was identical to Experiment 1.

Training. Prior to playing the task, participants received extensive instructions and were required to answer multiple-choice comprehension check questions on the key points of the task, with the question repeated until they either chose correctly or failed twice, upon which the experiment was terminated. The questions addressed the following key points of how the task worked: that participants' task was to correctly identify the direction of the dot motion; that on some trials they could receive a reward for correct responses; and that the bonus for dots moving in the desirable direction was independent of the responses they gave. Participants then played two practice sessions of 10 trials each where they received feedback. During the first practice session, they were given the opportunity to familiarize themselves with RDKs before the bonus structure was introduced. The second practice session was included to give participants prior experience with the trial types. If their choice accuracy was below 50% in practice session 1 and 40% in practice session 2 respectively, they had to repeat the training procedures. If they failed the training procedure twice, the experiment was terminated.

Statistical Analysis (Experiment 1 & Replication):

Response Bias. To determine whether participants were biased towards reaching desirable conclusions we calculated each participant's response bias. We define this as their tendency to overestimate the proportion of desirable trials encountered. We calculated response bias separately for each participant as follows:

$$Response\ Bias\ =\ Prop._{desirable\ judgments\ made} -\ Prop._{desirable\ trials\ encountered}$$

Such that positive values indicate a bias towards desirable conclusions while negative values indicate a bias towards undesirable conclusions. Response bias values were then entered into a one-sample t-test against zero to assess whether participants showed a significant response bias in their behaviour for each incentive level separately. We then compared the average response bias for each incentive level ($0 vs $5) using paired t-tests. Bayes tests were calculated to corroborate non-significant findings (Ly et al., 2016). Finally, one-sample t-tests against zero were performed to assess whether participants showed a significant response bias in their behaviour for each incentive level separately. We compared the pattern of results obtained from participants' real data to those obtained from simulated data (**see Drift-Diffusion Modelling**). All statistical tests conducted in the present article are two-sided. Analysis was conducted using IBM SPSS 27 and R Studio (Version 1.3.1056). All results of interest hold when controlling for noise (**see Appendix 7.2 Supplementary Tables 3.9-3.16**).

Response Times. To determine whether participants cared about accuracy incentives, we calculated their mean log-transformed RTs for trials in which they were incentivized for accuracy ($5) and trials in which they were not incentivized ($0). We then computed a paired t-test to assess how accuracy incentives ($0 vs $5) affected log-transformed RTs.

Drift-Diffusion Modelling. Our aim in modelling our task using DDM was to gain a better understanding of the underlying mechanisms of how accuracy incentives and desirability may alter the evidence accumulation process.

To that end, we implemented and compared 65 different specifications of a DDM. The models included the following parameters: (1) t0—amount of non-decision time; (2) α—distance between decision thresholds; (3) z—starting point of the accumulation process; and (4) v—drift rate - the rate of evidence accumulation. To reduce computational load, we adopted a theory-driven approach: First, a baseline model was estimated where we allowed all parameters to vary - but did not include dependencies on any of the motives. Then, informed by prior research (Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019), we reasoned that desirability might (1) reflect as a shift in the starting point to be closer to the bound associated with the desirable conclusion; and/or (2) induce selective accumulation of evidence in line with the desirable conclusion, as evidenced by a valence-dependent drift-rate bias. Furthermore, we speculated that when individuals are motivated to make accurate decisions, (1) they may be more cautious about making a decision, thereby increasing the distance between decision thresholds; and/or (2) they might weigh evidence more carefully thus slowing the rate of evidence accumulation (Shevlin et al., 2022). We also controlled for noise. We hypothesized that noise could (1) increase the distance between decision thresholds; such that participants are more cautious when the evidence is noisy; and/or (2) slow the rate of evidence accumulation, by making it harder to separate the evidence for each response option.

We first compared 64 models in which we tested these possibilities as main effects; and then assessed whether adding interactions would improve model fit of the winning model (**see Appendix 7.2 Supplementary Tables 3.1 & 3.2**

for list of all models). RTs faster than 200 ms were discarded from the model fits and further analysis, as recommended in previous literature (Ratcliff & Tuerlinckx, 2002; Rollwage et al., 2020; Wiecki et al., 2013). In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials. Based on this, from Experiment 1 we removed 1 participant and a further 64 trials, and from its replication, we removed 49 trials.

We used the HDDM software toolbox (Wiecki et al., 2013) to estimate the parameters of our models. The HDDM package employs hierarchical Bayesian parameter estimation, using MCMC methods to sample the posterior probability density distributions for the estimated parameter values. We estimated both group-level parameters as well as parameters for each individual participant. Parameters for individual participants were assumed to be randomly drawn from a group-level distribution. Participants' parameters both contributed to and were constrained by the estimates of group-level parameters. In fitting the models, we used priors that assigned equal probability to all possible values of the parameters. Models were fit to log-transformed RTs because RTs were non-normally distributed and had a heavy positive skew. Also, since our 'error' RT distribution included relatively fast errors we included an inter-trial starting point parameter (sz) to improve model fit (Ratcliff & Rouder, 1998). We sampled 10,000 times from the posteriors, discarding the first 5000 as burn in and thinning set at 5. MCMC are guaranteed to reliably approximate the target posterior density as the number of samples approaches infinity. To test if the MCMC converged within the allotted time, we used Gelman-Rubin statistic (Rubin & Gelman, 1992) on 5 chains of our sampling procedure. The Gelman–Rubin diagnostic evaluates MCMC convergence by analysing the difference between multiple Markov chains. The convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter. In each case, the Gelman-Rubin statistic was close to one (<1.1), suggesting that MCMC were able to converge.

Model fits were then compared using the DIC (Spiegelhalter et al., 2002), which is a generalization of the AIC for hierarchical models. The DIC is commonly used when the posterior distributions of the models have been obtained by MCMC simulation (Gamerman & Lopes, 2006). It allows one to assess the goodness of fit, while penalizing for model complexity. To examine the robustness of these results, we performed a complementary analysis and calculated Bayesian Predictive Information Criterion (BPIC) for each model (Ando, 2007). This corrects for over-fitting by adjusting for the asymptotic bias of the posterior mean of the log-likelihood as an estimator for its expected log-likelihood.

To validate the winning model, we used group parameters obtained from participants' data to simulate log-transformed RTs and responses. We used the exact number of participants, total number of trials and trial structure as in the experiment. Simulated data was then used to (1) perform model recovery analysis and (2) to compare the pattern of participants' responses to the pattern of simulated responses. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. Simulation and model recovery analysis were performed using the HDDM software toolbox (Wiecki et al., 2013).

To assess if desirability, accuracy incentives and noise significantly altered the model parameters, we calculated 95% CIs of the parameters' values. If the 95% CI for a parameter does not overlap zero (or 0.5 for the starting point), we consider there to be a significant effect. We also computed 95% HDIs of the parameter's posterior distributions. If this HDI did not overlap zero, we considered there to be a meaningful effect. HDI testing was conducted in R using *HDInterval* (Meredith & Kruschke, 2016).

Behavioural Task and Analysis (Experiment 2):

The task and analysis in Experiment 2 was identical to that used in Experiment 1 and its replication except for the following differences:

1) As before, we varied whether participants were incentivized for accuracy across trials. But this time, we increased the potential reward for correct

responses five-fold. In some trials participants could receive $25 for correct responses and in others $0 **(see Figure 3.2)**. The bonus for dots moving in the desirable direction was $5. Thus, in Experiment 2, we compared $0 vs $25 accuracy incentives, instead of $0 vs $5 accuracy incentives in Experiment 1 and its replication.

2) We also increased the range of noise levels. Noise varied across trials between 0.03 (high noise) to 0.3 (low noise) in increments of 0.03.
3) RTs faster than 200ms were discarded from the DDM fits (**see Appendix 7.2 Supplementary Table 3.3 for list of all models**) and further analysis, as recommended in previous literature (Ratcliff & Tuerlinckx, 2002; Rollwage et al., 2020; Wiecki et al., 2013). In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials. Based on this, from Experiment 2 we removed 1 participant and a further 134 trials.

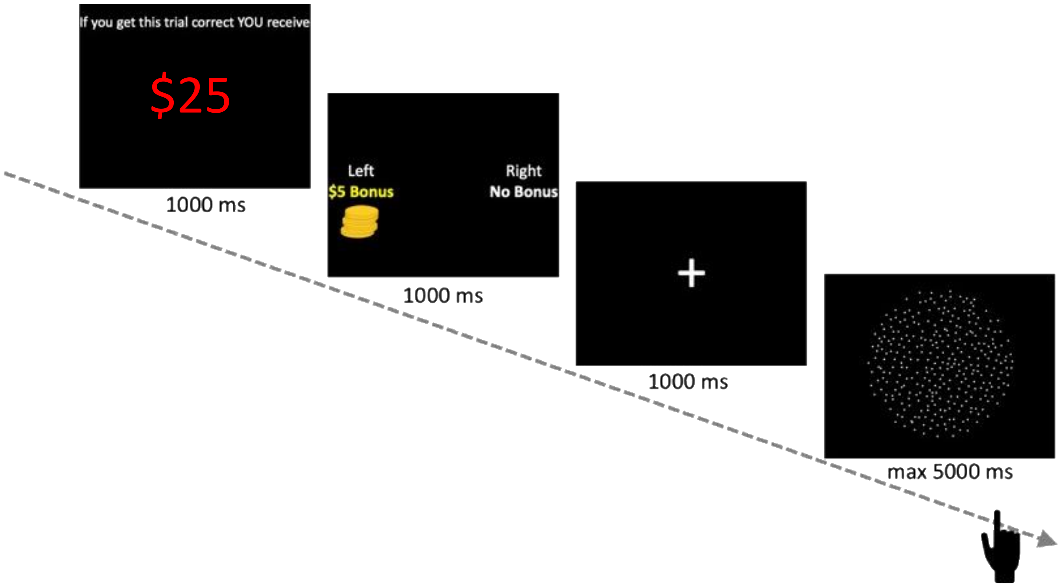

Figure 3.2. The Random Dot Kinetogram Task (Experiment 2). The task in Experiment 2 was identical to the task in Experiment 1 and its replication except for the following: Across trials, we varied the level of noise of the dot motion (0.03-0.3 in increments of 0.03) and whether participants were incentivized for accuracy. In Experiment 2 we increased the reward on incentivized trials to $25. On trials which were not incentivized the reward for correct responses was $0.

We maintained the bonus associated with the dots moving in a desirable direction at $5.

3.4 Results

In this study, we assessed whether rewarding participants for correct responses reduces the bias towards desirable conclusions in evidence accumulation. To test this, we ran two identical experiments ($N_{\text{Experiment 1}}$ = 70; $N_{\text{Replication}}$ = 73), in which participants completed a modified forced choice RDK task (Newsome & Pare, 1988, **see Figure 3.1**). In this task they judged whether most dots were moving right or left. Across trials we varied whether participants were incentivized for accuracy (i.e., paid for correct responses; reward = $5 vs no reward = $0) and the level of noise (low noise vs high noise). We also told participants that in each block, one direction was assigned to be the desirable direction, and the other the undesirable direction. Participants were told that each time the dots moved in the desirable direction they would win an additional $5 bonus. Crucially, participants had no control over which type of trial they were in (i.e., which direction the dots were moving), their task was simply to determine the direction, gaining rewards for accurate decisions on incentivized trials. We anticipated that this manipulation would cause participants to systematically err towards believing the dots were moving in the desirable direction even when they were not.

The manipulation successfully biased participants towards desirable conclusions. We first assessed whether our manipulation successfully induced a bias towards desirable conclusions. To that end, we calculated each participant's response bias as follows:

$$Response\ Bias\ =\ Prop._{desirable\ judgments\ made} -\ Prop._{desirable\ trials\ encountered}$$

such that positive values indicate a bias towards desirable conclusions while negative values indicate a bias towards undesirable conclusions.

Indeed, participants judged a larger proportion of trials as desirable than the proportion of desirable trials they actually encountered (Experiment 1: $M_{Response\ Bias}$ = 0.032, SE = 0.014, t(68) = 2.361, p = 0.021, Cohen's d = 0.284, Replication: $M_{Response\ Bias}$ = 0.047, SE = 0.014, t(72) = 3.323, p = 0.001, Cohen's d = 0.389). Thus, our manipulation was successful at inducing a bias towards desirable conclusions (replicating previous findings from Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019).

Participants become more cautious when incentivized for accuracy. We next examined whether accuracy incentives altered participants' behaviour. We speculated, that if participants were sensitive to accuracy incentives, they should be more cautious on incentivized trials ($5 reward) than on trials in which there was no reward for correct responses ($0 reward). Indeed, we found that participants were slower to respond when they were incentivized for accuracy (Experiment 1: $5 accuracy incentives – M_{logRT} = 7.25, SE = 0.048; $0 accuracy incentives: M_{logRT} = 7.219, SE = 0.05; t(68) = 3.478, p < 0.001; Cohen's d = 0.07; Replication: $5 accuracy incentive – M_{logRT} = 7.282, SE = 0.043; $0 accuracy incentives: M_{logRT} = 7.26, SE = 0.044; t(72) = 2.755, p = 0.007; Cohen's d = 0.059**; see Figure 3.3**). Adding the level of noise to the analysis does not alter the results (**see Appendix 7.2 Supplementary Tables 3.4 & 3.5**). While the effect of incentives is small, these results suggests that participants are sensitive to accuracy incentives and are slower about their decisions when incentivized.

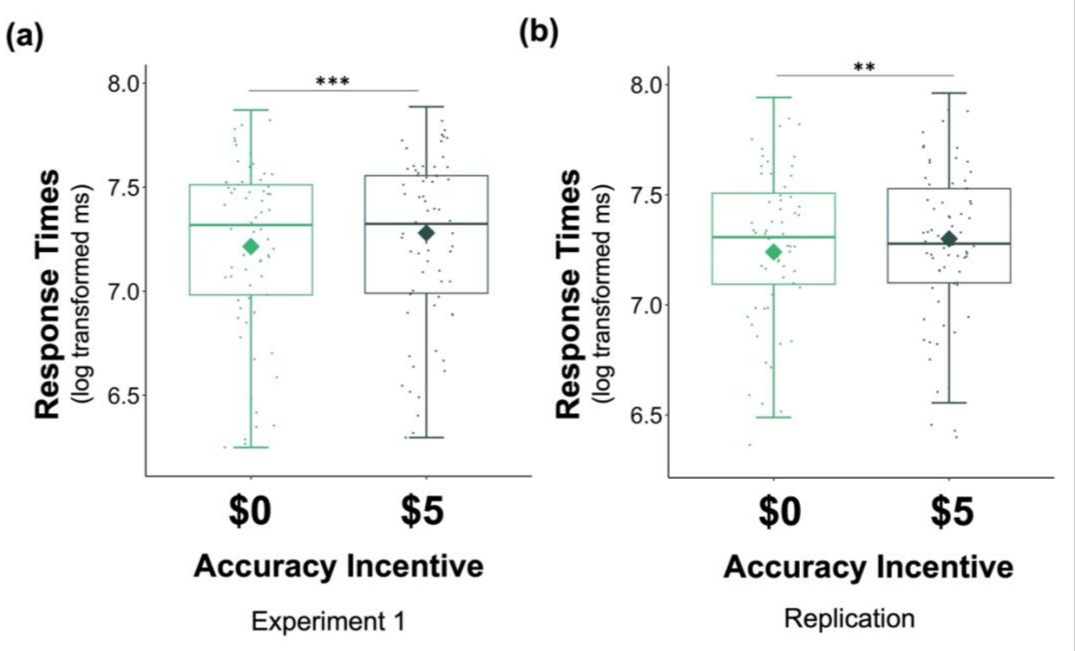

Figure 3.3. Participants are more cautious when incentivized for accuracy (Experiment 1 & Replication). In both **(a)** Experiment 1, and **(b)** its replication, participants took longer to reach a conclusion when incentivized for accuracy ($5), compared to when there was no reward for correct responses ($0). Y axis shows log-transformed RT. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean log-transformed RT per incentive level. Individuals' mean response time is shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001, **p < 0.01.

Accuracy incentives do not reduce the bias towards desirable conclusions.

We next turned to our primary question - do accuracy incentives reduce the bias towards desirable conclusions? We found that despite participants taking longer to reach a conclusion when incentivized for accuracy, their bias towards desirable conclusions remained unchanged. In particular, the magnitude of the response bias, measured as the proportion of trials judged as desirable minus the proportion of desirable trials actually encountered, remained the same regardless of whether accuracy was incentivized ($5 accuracy incentives – Experiment 1: $M_{\text{Response Bias}}$ = 0.035, SE = 0.014, Replication: $M_{\text{Response Bias}}$ = 0.042, SE = 0.015) or not ($0 accuracy incentives – Experiment 1: $M_{\text{Response Bias}}$ = 0.03, SE = 0.001, Replication: $M_{\text{Response Bias}}$ = 0.052, SE = 0.017; comparison between the two: Experiment 1: t(68) = -0.532, p = 0.96, Cohen's d = 0.064, Replication: t(72) = 0.765, p = 0.447, Cohen's d = 0.089, **see Figure 3.4**). Adding the level of noise to the analysis does not alter the results (**see Appendix 7.2 Supplementary Tables 3.6 & 3.7**). Bayes tests further provide moderate to strong support in favour of the null hypothesis (Experiment 1: BF_{01} = 9.181, Replication: BF_{01} = 8.139).

One-sample t-tests against zero revealed that for each incentive level participants concluded they were in a desirable trial significantly more often than an undesirable trial and thus overestimated the proportion of desirable trials they encountered ($0 accuracy incentives – Experiment 1: t(68) = 2.051, p = 0.044, Cohen's d = 0.247; Replication: M = 0.052, SE = 0.017, t(72) = 3.146, p = 0.002, Cohen's d = 0.368; $5 accuracy incentives – Experiment 1: M = 0.035, SE = 0.014, t(68) = 2.429, p = 0.018, Cohen's d = 0.119; Replication: M = 0.042, SE = 0.015, t(72) = 2.79, p = 0.007, Cohen's d = 0.327). Thus, participants are biased towards desirable responses irrespective of whether they are incentivized for accuracy. In line with this, we did not observe an improvement in overall discernment between signal and noise when incentivized vs when not incentivized (**see Appendix 7.2 Supplementary Results & Supplementary Tables 3.8 & 3.9 for analysis of dPrime**).

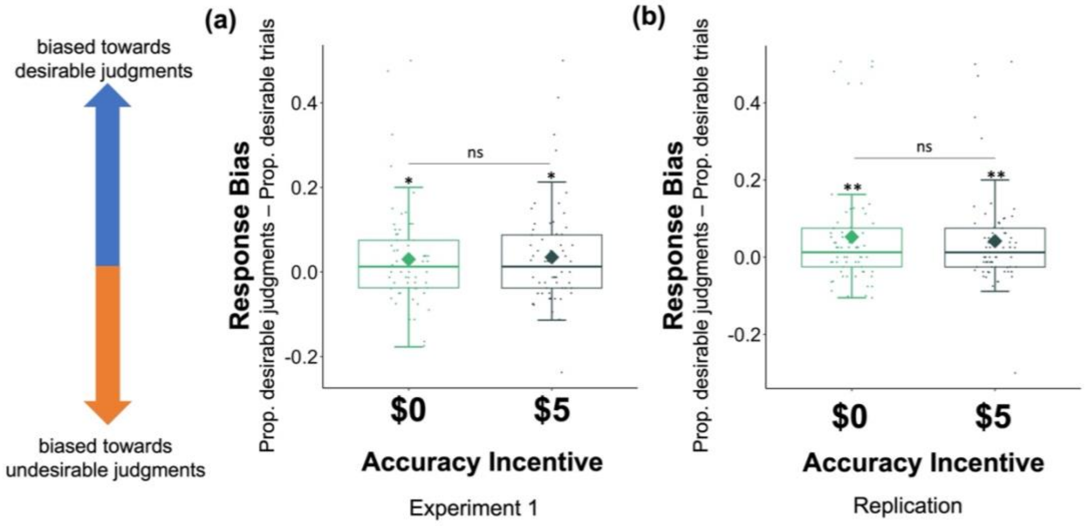

Figure 3.4. Accuracy incentives do not reduce the response bias (Experiment 1 & Replication). In both **(a)** Experiment 1, and **(b)** its replication participants were biased towards desirable responses. Accuracy incentives did not reduce this bias. For each accuracy incentive level, the proportion of trials they judged as desirable was significantly larger than the proportion of desirable trials they encountered. Y axis shows response bias, i.e., proportion of desirable judgements made minus proportion of desirable trials encountered. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean response bias per incentive level. Individuals' response bias is shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. *$p < 0.05$, ns = not significant.

Specifying the impact of accuracy incentives and desirability on the stages of evidence accumulation.

Thus far, our results suggest that although participants are sensitive to accuracy incentives, the way in which they accumulate evidence remains biased. We next considered different potential mechanisms through which this behaviour could be explained. To that end, we considered how accuracy incentives, desirability and noise may influence the evidence accumulation process.

Informed by prior research (Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019), we speculated that when participants are motivated to reach a desirable conclusion they may (1) selectively accumulate evidence in line with this desirable conclusion (drift rate); and/or (2) may be a priori more likely to believe they are in the desirable state even before seeing any evidence (starting point). Rewarding individuals for accurate responses makes them more cautious, potentially (1) increasing the amount of evidence they require before reaching a conclusion (distance between decision thresholds); and/or (2) causing them to weigh the evidence they gather in a less biased manner (drift rate, Shevlin et al., 2022). When the evidence is noisy, participants may (1) increase the amount of evidence they require before reaching a conclusion (distance between decision thresholds); and/or (2) slow the rate at which they accumulate evidence (drift rate).

To determine which, if any, of the above holds, we fit a DDM to the data. Specifically, we modelled participants' responses as a drift-diffusion process (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013) with the following parameters: (1) α—distance between decision thresholds; (2) t0—amount of non-decision time; (3) z—starting point of the accumulation process; and (4) v—drift rate, i.e., the rate of evidence accumulation.

In a first step, we set out to identify the model that best captures the data. To that end, we compared different model specifications, each allowing for different dependencies of each parameter on accuracy incentives, desirability, and noise. To reduce computational load, we first compared 64 different model specifications in which we tested all the above possibilities as main effects. We then added interactions to the winning model to assess whether this would improve model fit (**see Methods for further details**).

The DIC, a generalization of the AIC for hierarchical models, was calculated for each model (**see Appendix 7.2 Supplementary Tables 3.1 & 3.2**). DIC scores indicated that for both experiments the model that outperformed all other models allowed: (1) accuracy incentives to influence the distance between decision thresholds (2) desirability to influence starting point and drift rate and (3) noise to influence drift rate. Allowing for interaction effects did not improve model fit (**see Appendix 7.2 Supplementary Tables 3.1 & 3.2**). This suggests that the effect of desirability is not modulated by accuracy incentives, rather accuracy incentives and desirability influence different elements of the accumulation process. While desirability and noise alter the quality of evidence accumulation through the drift rate, accuracy incentives do not alter the quality of accumulation but instead influence the decision-making process by adjusting the caution with which decisions are made, i.e., the distance between decision thresholds. BPIC results corroborate these findings (**see Appendix 7.2 Supplementary Tables 3.1 & 3.2**).

Thus far, we have therefore established that a model which incorporates desirability, accuracy incentives and noise, without interactions among these variables fits the data best. To reveal the direction, magnitude, and significance of each effect, we, we calculated the 95% CIs for each parameter. If the 95% CI of the parameter distribution does not overlap with zero (or 0.5 for the starting point), we infer a significant effect.

This revealed that accuracy incentives increased the distance between decision thresholds (Experiment 1: α=0.059; 95% CI [0.021, 0.089], Replication: α=0.032; 95% CI [0.001, 0.072], **Table 3.1 & 3.2, Figure 3.5 a&b**). When rewarded for correct responses, participants were more cautious and needed more evidence before reaching a conclusion. In addition to this, desirability had a statistically meaningful effect on the drift rate (Experiment 1: v=0.125, 95% CI [0.002, 0.257]; Replication: v=0.278, 95% CI [0.12, 0.431], **Figure 3.5 c&d** replicating previous findings from Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019). Participants selectively accumulated evidence towards desirable conclusions, and thus had a larger drift rate when the dots moved in the desirable direction. By contrast, desirability did not significantly modulate the starting point (Experiment 1: z=0.504, 95% CI [0.494, 0.515], Replication: z=0.49, 95% CI [0.479, 0.50], replicating previous findings from Globig et al., 2021). Finally, we observed a significant effect of noise on the drift rate (Experiment 1: v=0.383; 95% CI [0.299, 0.461], Replication: v=0.396; 95% CI [0.298, 0.492], **Figure 3.5 e&f**). That is, when the evidence was noisy, it was harder for participants to separate the evidence for each response option, thereby slowing the rate of evidence accumulation. 95% HDI comparisons corroborate this result **(see Appendix 7.2 Supplementary Table 3.10 for HDI Comparisons**).

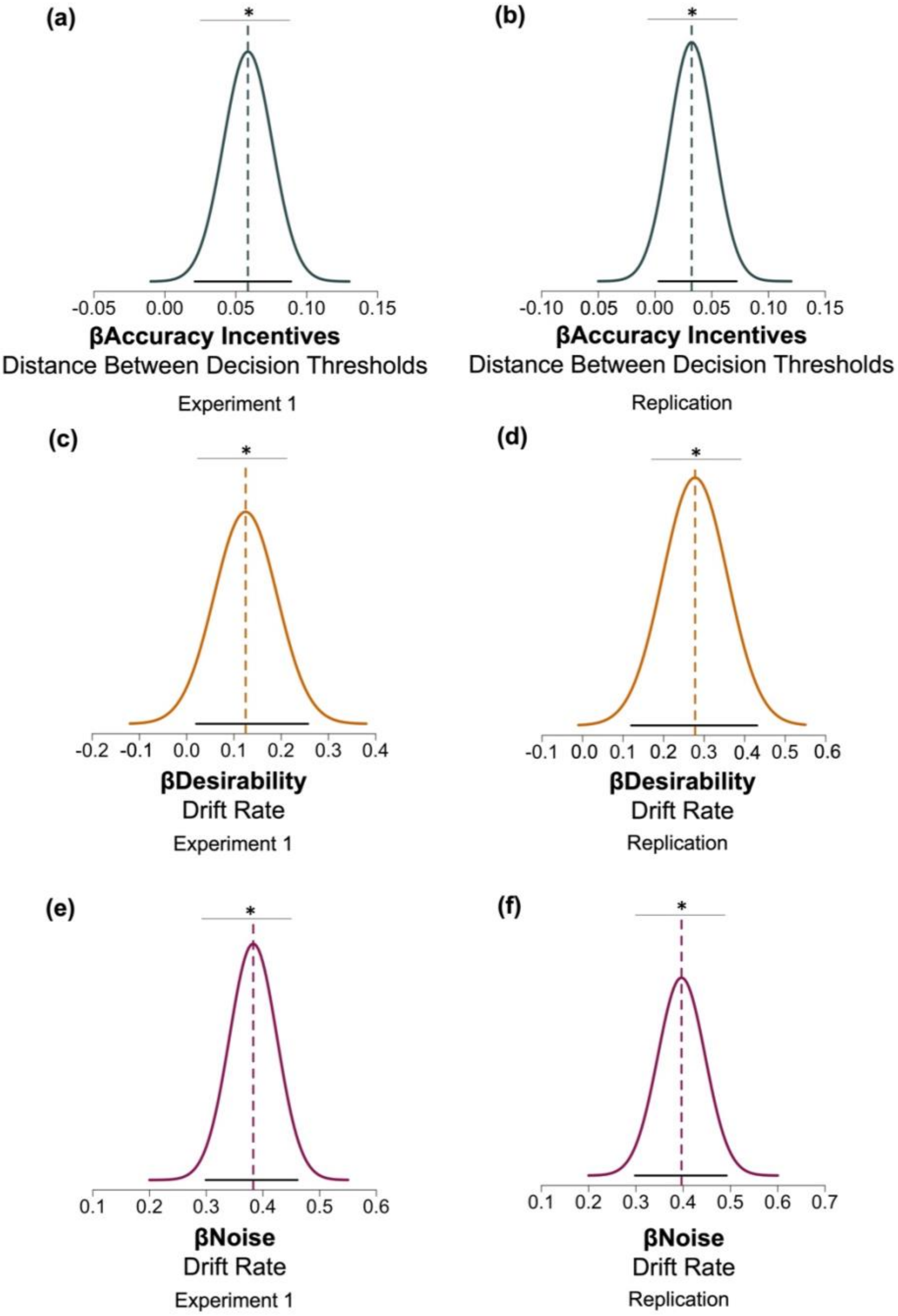

Figure 3.5. Different motives affect different aspects of the evidence accumulation process (Experiment 1 & Replication). When participants were incentivized for accuracy, **(a&b)** the distance between decision thresholds increased. **(c&d)** The rate of evidence accumulation was greater when the dots were moving in the desirable direction and **(e&f)** when noise was high. Displayed are the posterior distributions of the parameter estimates for each feature for the parameter it significantly modulates in Experiment 1 (left) and its

replication (right). Coloured dashed vertical lines indicate group means. Black line represents 95% CI. *indicates significant effect.

These findings lend support to some of our hypothesized mechanisms. Specifically, we find that (i) when individuals are motivated to make accurate decisions, they are more cautious **-** increasing the distance between decision thresholds; (ii) desirability induces selective accumulation of evidence towards desirable conclusions, and (iii) high levels of noise slow the rate of evidence accumulation. In line with the behavioural results, we do not find evidence for accuracy incentives modulating the effect of desirability on behaviour. This suggests that accuracy incentives fail to mitigate the influence of desirability on evidence accumulation, because the two act on different elements of the accumulation process. While participants required more information to make a decision, the way in which they accumulated evidence to reach that decision remained biased.

Model recovery is successful and simulated data reproduces experimental results.

To determine whether our winning model accurately captured the data we examined whether the model parameters could be successfully recovered based on simulated data. To that end, we first simulated data using the group parameters (**see Methods for details**). We then fit the winning model to the simulated data, in the same way as for the experimental data. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. As shown in **Table 3.1**-**3.2** model parameters could be successfully recovered based on the simulated data.

Estimate	**Experimental Data** [95% CI]	**Simulated Data** [95% CI]
Distance between Decision Thresholds (α)	2.294 [2.189, 2.39]	2.372 [2.349, 2.398]
βAccuracy Incentives Distance between Decision Thresholds	0.059 [0.021, 0.089]	0.051 [0.009, 0.093]
Non-decision Time (t0)	6.104 [5.983, 6.216]	5.989 [5.974, 0.004]
Starting Point (z)	0.504 [0.494, 0.515]	0.496 [0.484, 0.509]
inter-trial Starting Point (sz)	0.053 [0.002, 0.076]	0.069 [0.004, 0.159]
Drift Rate (β0)	0.093 [-0.004, 0.187]	0.091 [0.056, 0.126]
βDesirability Drift Rate	0.125 [0.002, 0.257]	0.131 [0.078, 0.182]
βNoise Drift Rate	0.382 [0.299, 0.461]	0.379 [0.332, 0.427]

Table 3.1. Real and Recovered Parameter estimates of the evidence accumulation process (Experiment 1). Displayed are the real (left) and recovered (right) model estimates from the winning model. These include a constant for Distance Between Decision Thresholds (α), βAccuracy Incentives Distance between Decision Thresholds, Non-decision Time (t0), Starting Point (0<z<1), Inter-trial Starting Point (sz), constant Drift Rate (β0), βDesirability Drift Rate (β1), and βNoise Drift Rate. [CI].

Estimate	**Experimental Data** [95% CI]	**Simulated Data** [95% CI]
Distance between Decision Thresholds (α)	2.318 [2.187, 2.449]	2.36 [2.358, 2.367]
βAccuracy Incentives Distance between Decision Thresholds	0.032 [0.001, 0.072]	0.032 [0.001, 0.055]
Non-decision Time (t0)	6.181 [6.062, 6.3]	6.068 [6.051, 6.086]
Starting Point (z)	0. 49 [0.479, 0.50]	0. 484 [0.474, 0.494]
inter-trial Starting Point (sz)	0.061 [0.005, 0.12]	0.118 [0.009, 0.206]
Drift Rate (β0)	0.039 [-0.062, 0.139]	0.019 [-0.009, 0.043]
βDesirability Drift Rate	0.278 [0.12, 0.431]	0.295 [0.263, 0.328]
βNoise Drift Rate	0.396 [0.298, 0.492]	0.363 [0.327, 0.40]

Table 3.2. Real and Recovered Parameter estimates of the evidence accumulation process (Replication). Displayed are the real (left) and recovered (right) model estimates from the winning model. These include a constant for Distance between Decision Thresholds (α), βAccuracy Incentives

Distance between Decision Thresholds, Non-decision Time (t0), Starting Point (0<z<1), Inter-Trial Starting Point (sz), constant Drift Rate (β0), βDesirability Drift Rate (β1), and βNoise Drift Rate. [CI].

Additionally, we examined whether the simulated data reproduced the same behavioural pattern of results as participants' data. This was indeed the case. In both the simulated data for Experiment 1 and its replication (**see Figure 5**) the response bias was not reduced by accuracy incentives ($5 accuracy incentives – Simulated Data Experiment 1: $M_{\text{Response Bias}}$ = 0.027, SE = 0.009, Simulated Data Replication: $M_{\text{Response Bias}}$ = 0.06, SE = 0.007; $0 accuracy incentives – Simulated Data Experiment 1: $M_{\text{Response Bias}}$ = 0.032, SE = 0.009, Simulated Data Replication: $M_{\text{Response Bias}}$ = 0.057, SE = 0.008; comparison between the two: Simulated Data Experiment 1: t(68) = 0.433, p = 0.666, Cohen's d = 0.052, Simulated Data Replication: t(72) = 0.291, p = 0.772, Cohen's d = 0.034, **see Figure 3.6**). Adding the level of noise to the analysis does not alter the results (**see Appendix 7.2 Supplementary Tables 3.11 & 3.12**). Bayes tests further provide moderate to strong support in favour of the null hypothesis (Simulated Data Experiment 1: BF_0 = 9.625, Simulated Data Replication: BF_{01} = 10.407).

One-sample t-tests against zero show that the response bias was significant in each incentive level ($0 accuracy incentives – Simulated Data Experiment 1: t(68) = 3.582, p < 0.001, Cohen's d = 0.431; Simulated Data Replication: t(72) = 6.959, p < 0.001, Cohen's d = 0.814; $5 accuracy incentives – Simulated Data Experiment 1: t(68) = 3.089, p = 0.003, Cohen's d = 0.372 Simulated Data Replication: t(72) = 7.993, p < 0.001, Cohen's d = 0.935). Thus, both the simulated data and the experimental data show a bias towards desirable conclusions irrespective of accuracy incentives.

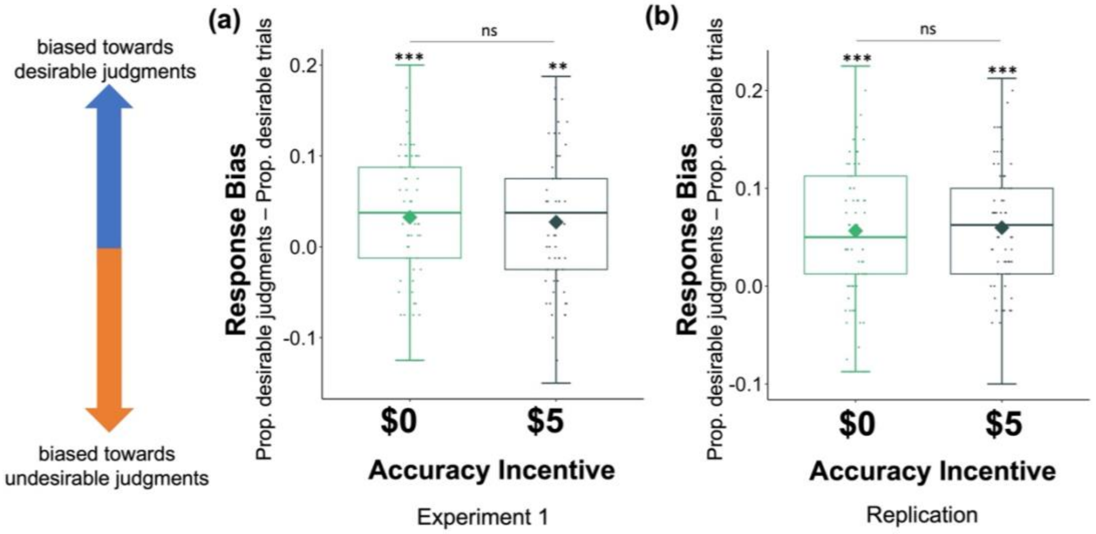

Figure 3.6. Simulated Data reproduces experimental data (Experiment 1 & Replication). As in the experimental data, simulated data from DDM parameters for both **(a)** Experiment 1, and **(b)** its replication shows a bias towards desirable responses. Accuracy incentives did not reduce this bias. Y axis shows response bias, i.e., the proportion of trials judged as desirable minus proportion of desirable trials encountered. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean response bias per incentive level. Simulated individuals' response bias is shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. ***p< 0.001, **p<0.01, ns = not significant.

Participants remain biased despite a five-fold increase in the reward for correct responses.

Up to this point, we show that modest accuracy incentives ($5), which match the magnitude of the bonus amount ($5) participants get when the dots move in the desirable direction, fail to mitigate biased evidence accumulation. Could a larger accuracy incentive, however, succeed in mitigating the bias? To test this, we increased the reward for correct responses five-fold, offering $25 for correct responses on some trials, and $0 in others. We maintained the bonus for desirable trials at $5 **(see Figure 3.7)**. As before, participants had no control over the trial type and were incentivized to give accurate response to maximize financial gain in the task.

We repeated the analysis as in Experiment 1 and its replication. Once again, participants judged a larger proportion of trials as desirable than the proportion of desirable they encountered (Experiment 2: M = 0.035, SE = 0.01, t(91) = 3.568, p < 0.001, Cohen's d = 0.372). Thus, our manipulation was successful. Consistent with our previous results, participants were also slower to respond when they were incentivized for accuracy (Experiment 2: $25 accuracy incentives – M = 7.211, SE = 0.043; $0 accuracy incentives – M = 7.157, SE = 0.044; t(91) = 2.649, p < 0.001; Cohen's d = 0.132, **see Figure 3.7).** Adding the level of noise to the analysis does not alter the results (**see Appendix 7.2 Supplementary Table 3.13**).

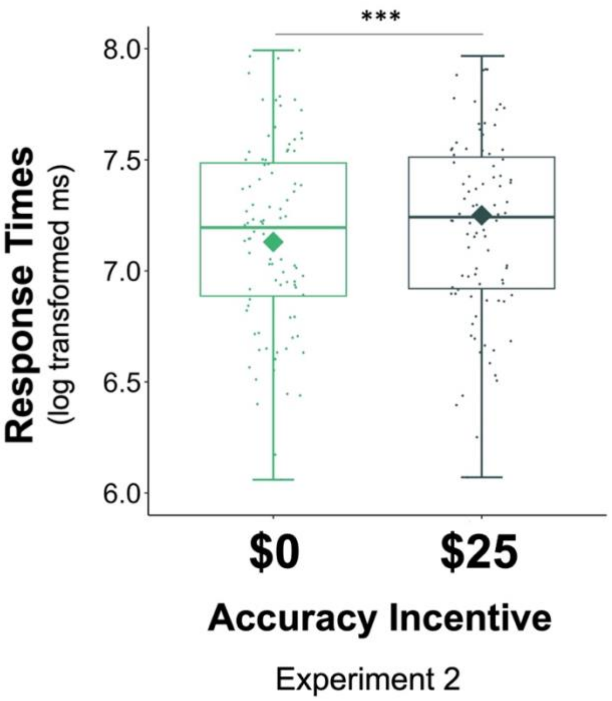

Figure 3.7. Participants are more cautious when incentivized for accuracy (Experiment 2). Participants in Experiment 2 took longer to reach a conclusion when incentivized ($25) for correct responses, compared to when they were not ($0). Y axis shows log-transformed RT. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean log-transformed RT per incentive level. Individuals' mean response time is

shown separately as dots. Symbols above boxplot indicate significance level compared to 0. ***p < 0.001.

Crucially, despite the large incentive for accuracy, the magnitude of participants' response bias was not significantly reduced ($25 accuracy incentives: $M_{\text{Response Bias}}$ = 0.023, SE = 0.012; $0 accuracy incentives: $M_{\text{Response Bias}}$ = 0.047, SE = 0.013; comparison between the two: t(91) = 1.508, p = 0.135, Cohen's d = 0.157, **see Figure 3.8**). Adding the level of noise to the analysis does not alter the results (**see Appendix 7.2 Supplementary Table 3.14**). A Bayes test further provides moderate to strong support in favour of the null hypothesis (Experiment 2: BF_{01} = 3.998).

One-sample t-tests against zero showed that participants were biased towards desirable conclusions in trials in which there was no reward for correct responses ($0 accuracy incentives: M = 0.047, SE = 0.013, t(91) = 3.567, p < 0.001, Cohen's d = 0.372) and in trials in which there was a $25 reward for correct responses ($25 accuracy incentives: M = 0.023, SE = 0.012, t(91) = 1.976, p = 0.05, Cohen's d = 0.206; **see Appendix 7.2 Supplementary Material & Supplementary Table 3.15 for analysis of dPrime**).

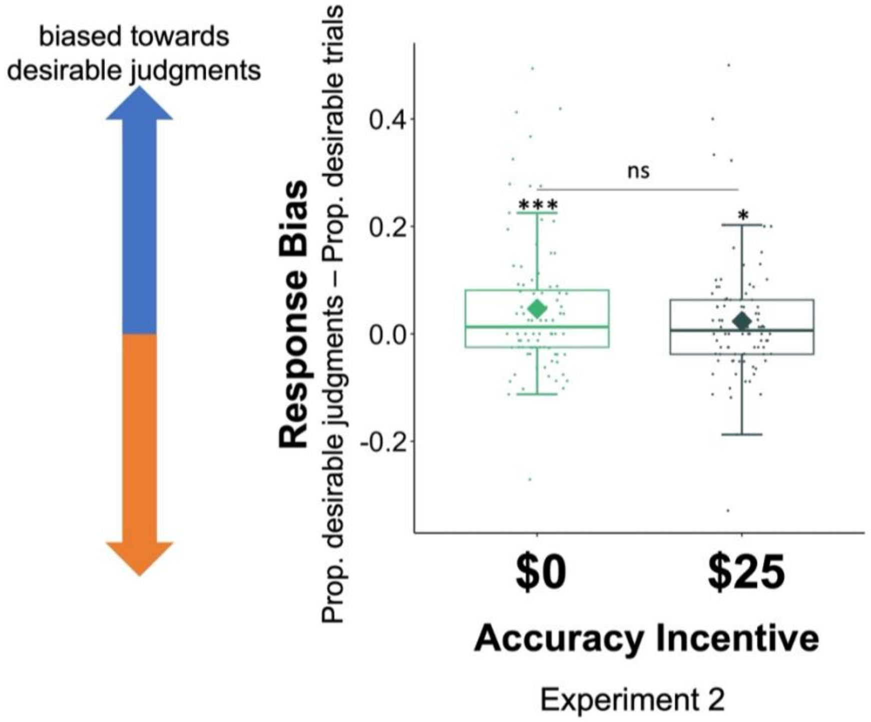

Figure 3.8. Accuracy incentives that are five times larger than the bonus do not reduce the response bias (Experiment 2). The proportion of trials participants judged as desirable was significantly larger than the proportion of desirable trials they encountered irrespective of whether the accuracy incentive was $0 or $25. Y axis shows response bias, i.e., proportion of desirable judgements made minus proportion of desirable trials encountered. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean response bias per incentive level. Individuals' response bias is shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. *p<0.05, ns = not significant.

As before, we then fit the DDM to our data to examine the mechanisms underlying these results. As in Experiment 1 and its replication, we found that the model that best explained participants' behaviour allowed (1) accuracy incentives to influence the distance between decision threshold; (2) desirability to influence starting point and drift rate; and (3) noise to influence drift rate. Allowing for interaction effects once again did not improve model fit (**see Appendix 7.2 Supplementary Table 3.3**).

As before, accuracy incentives increased the distance between decision thresholds such that participants were more cautious when incentivized for accuracy (Experiment 2: α=0.058; 95% CI [0.024, 0.093, **Table 3.3, Figure 3.9a**). The drift rate was larger when the majority of the dots moved in the desirable direction (Experiment 2: v=0.124; 95% CI [0.041, 0.209], **Figure 3.9b)** and decreased when the evidence was noisy, thus making it harder for participants to separate the evidence for each response option (Experiment 2: v=1.68; 95% CI [1.263, 2.105]; **Figure 3.9c**). Desirability did not significantly modulate the starting point (Experiment 2: z=0.503, 95% CI [0.493, 0.513]). We therefore replicate the results from Experiment 1 and its replication. 95% HDI comparisons corroborate this result **(see Appendix 7.2 Supplementary Table 3.16 for HDI Comparisons**).

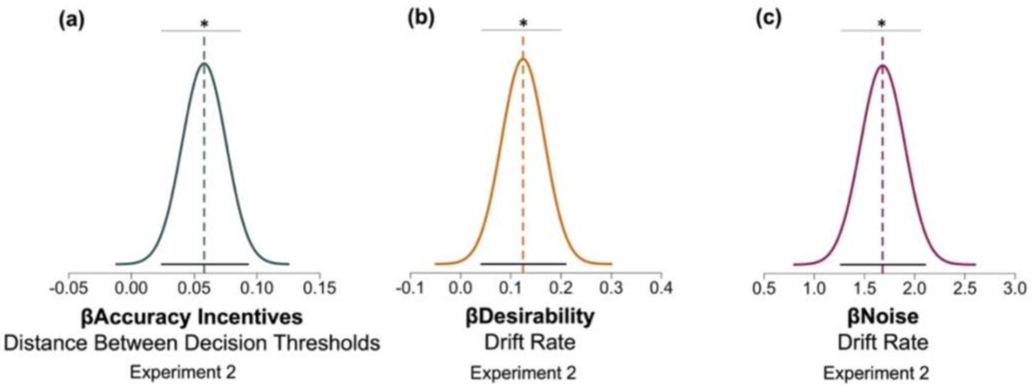

Figure 3.9. Different motives affect different aspects of the evidence accumulation process (Experiment 2). When participants were incentivized for accuracy, **(a)** the distance between decision thresholds increased**. (b)** The drift rate was greater when the dots were moving in the desirable direction and **(c)** when noise was low. Displayed are the posterior distributions of the parameter estimates for each motive for the parameter it significantly modulates in Experiment 2. Coloured dashed vertical lines indicate group means. Black line represents 95% CI%. *indicates significant effect.

Finally, we successfully recovered the model parameters based on simulated data (**see Table 3.3; see Methods for details**) and found that the simulated data reproduced the same behavioural pattern of results as participants' data. Accuracy incentives did not alter the bias towards desirable conclusions (Simulated Data Experiment 2: t(91) = 0.032, p = 0.975, Cohen's d = 0.003, **Figure 3.10).** The response bias across trials in which there was no reward for correct responses ($0 accuracy incentives – Simulated Data Experiment 2: 0.034, SE = 0.004) did not differ from trials in which there was a reward, even when the reward was five times larger than the bonus ($25 accuracy incentives – Simulated Data Experiment 2: M = 0.034, SE = 0.004). These results hold when controlling for noise (**see Appendix 7.2 Supplementary Table 3.17**). Bayes tests further provide moderate to strong support in favour of the null hypothesis (Simulated Data Experiment 2: BF_{01} = 12.143).

Estimate	**Experimental Data** [95% CI]	**Simulated Data** [95% CI]
Distance between Decision Thresholds (α)	2.3 [2.225, 2.374]	2.167 [2.141, 2.195]
βAccuracy Incentives Distance	0.058 [0.024, 0.093]	0.121 [0.085, 0.158]

between Decision Thresholds		
Non-Decision Time (t0)	6.037 [5.933, 6.144]	5.889 [5.879, 5.897]
Starting Point (z)	0.503 [0.493, 0.513]	0.501 [0.497, 0.506]
inter-trial starting point parameter (sz)	0.051 [0.003, 0.106]	0.044 [0.003, 0.087]
Drift Rate (β0)	-0.018 [-0.076, 0.041]	-0.022 [-0.039, -0.006]
βDesirability Drift Rate	0.124 [0.041, 0.209]	0.127 [0.113, 0.158]
βNoise Drift Rate	1.68 [1.263, 2.105]	1.56 [1.478, 1.628]

Table 3.3. Real and Recovered Parameter estimates of the evidence accumulation process (Experiment 2). Displayed are the real (left) and recovered (right) model estimates from the winning model. These include a constant for distance between decision thresholds (α), βAccuracy Incentives distance between decision thresholds, non-decision time (t0), starting point (0<z<1), inter-trial starting point parameter (sz), constant drift rate (β0), βDesirability drift rate (β1), and βNoise Drift Rate. [CI].

One-sample t-tests against zero show that the response bias was significant irrespective of incentive level ($0 accuracy incentives – Simulated Data: t(91) = 9.201, p < 0.001, Cohen's d = 0.959, $25 accuracy incentives – Simulated Data: t(91) = 9.551, p < 0.001, Cohen's d = 0.996). Thus, we observed a bias towards desirable responses, in both the simulated data and the experimental data, irrespective of the potential reward for correct responses. The results from Experiment 2 therefore illustrate that the desire to form positive beliefs and thus reach desirable conclusions, are not mitigated by rewarding participants for correct responses, even when the magnitude of the reward is relatively large.

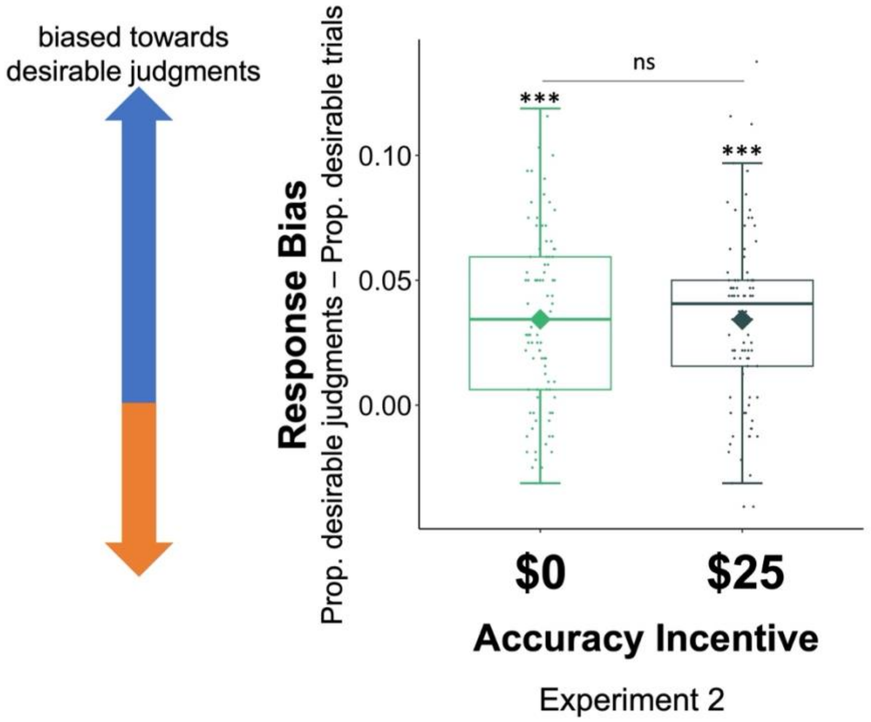

Figure 3.10. Simulated Data reproduces experimental data (Experiment 2). As in the experimental data, simulated data from DDM parameters shows a bias towards desirable responses. Accuracy incentives did not reduce this bias. For each accuracy incentive level, the proportion of trials judged as desirable was significantly larger than the proportion of desirable trials encountered. Y axis shows response bias, i.e., proportion of desirable judgements made minus proportion of desirable trials encountered. X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean response bias per incentive level. Individuals' response bias is shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. ***$p < 0.001$, ns = not significant.

3.5 Discussion

Over three experiments we find that although participants are more cautious in reaching conclusions when incentivized for accuracy, the magnitude of their bias is unaffected. Specifically, when incentivized participants take longer to reach a conclusion, but their tendency to falsely believe they are in a desirable state (a state where rewards are greater) remains unchanged. This was true even when the reward for correct responses was increased five-fold.

Importantly, we reveal that this failure is due to accuracy incentives impacting a feature of the accumulation process that is orthogonal to the one influenced by wishful thinking. Specifically, when incentivized for accuracy participants required more evidence before making a decision, which is signified as a larger distance between decision thresholds in the evidence accumulation model (the DDM). However, the desire to hold a preferred belief was associated with greater weight assigned to desirable evidence, signified in the model as a larger drift rate when in a desirable state (replicating previous findings from Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019). Thus, while participants accumulate more evidence before reaching a conclusion, the accumulation process remains biased.

These results fit with previous suggestions that even substantial accuracy incentives fail to alleviate certain biases, including the partisan bias in advice taking (Zhang & Rand), the repeated truth effect (Speckmann & Unkelbach, 2022), and anxiety-induced wishful thinking (Engelmann et al., 2019). The current findings go beyond these previous demonstrations and provide an explanation for why accuracy incentives may fail to reduce biased evidence accumulation.

Interestingly, our results suggest that participants consciously strive to maximize financial gains, slowing down when incentivized. Yet, the fact that this attempt is unfruitful may suggest they are not consciously aware of their bias towards forming desirable beliefs. This idea, that a desirability bias in evidence accumulation is automatic and unconscious, is in accord with findings showing that the magnitude of the bias is constant under time constraints and different degrees of cognitive load (Kappes & Sharot, 2019). These findings lend support to previous suggestions that motivational biases are not merely the product of overt preferences but instead beyond individuals' awareness (Chen & Krajbich, 2018; Desai & Krajbich, 2022; Krajbich, 2022; but see Sánchez-Fuenzalida et al., 2023).

A crucial feature of the task is that participants did not receive feedback. Thus, they did not have an opportunity to reflect on their errors and become aware of their bias. We speculate that to correct motivational biases, accuracy incentives should be coupled with feedback. Chapter 4 tests this assumption, by providing participants with direct feedback about the accuracy of the content they shared in simulated social media environments. We find that providing participants with accuracy-related feedback reduced the tendency to share misinformation in domains such as politics, culture, and health (Globig et al., 2023).

In summary, the study shows that accuracy incentives fail to mitigate the influence of the motivational bias on evidence accumulation. We suggest that this is because the desirability bias and accuracy incentives alter distinct elements of the evidence accumulation process. The former alters the relative weight assigned to desirable and undesirable information, while the latter increases the amount of evidence required to reach a conclusion. These findings are particularly relevant for policymakers and industry leaders as they may explain why financial bonuses may be ineffective in improving decision-making and performance.

Chapter 4: Changing the Incentive Structure of Social Media Platforms to Halt the Spread of Misinformation

4.1 Abstract

The powerful allure of social media platforms has been attributed to the human need for social rewards. Here we demonstrate that the spread of misinformation on such platforms is facilitated by existing social 'carrots' (e.g., 'likes') and 'sticks' (e.g., 'dislikes') that are dissociated from the veracity of the information shared. Testing 951 participants over six experiments, we show that a slight change to the incentive structure of social media platforms, such that social rewards and punishments are contingent on information veracity, produces a considerable increase in the discernment of shared information. Namely, an increase in the proportion of true information shared relative to the proportion of false information shared. Computational modelling revealed that the underlying mechanism of this effect is associated with an increase in the weight participants assign to evidence consistent with discerning behaviour. The results offer evidence for an intervention that could be adopted to reduce misinformation spread, which in turn could reduce violence, vaccine hesitancy and political polarization, without reducing engagement.

4.2 Introduction

In recent years, the spread of misinformation online has skyrocketed, increasing polarization, racism and resistance to climate action and vaccines (Barreto et al., 2021; Rapp & Salovich, 2018; Tsfati et al., 2020; Van Bavel et al., 2021). Existing measures to halt the spread, such as flagging posts, have had limited impact (Chan et al., 2017; Grady et al., 2021; Lees et al., 2022).

We hypothesize that the spread of misinformation on social media platforms is facilitated by the existing incentive structure of those platforms, where social rewards (in the form of 'likes' and 'shares') are dissociated from the veracity of the information (Sharot, 2021). The rationale for this hypothesis is as follows; users can discern true from false content to a reasonable degree (Allen et al., 2021; Pennycook & Rand, 2019). Yet, because misinformation generates no less retweets and ‘likes’ than reliable information (Lazer et al., 2018; Vosoughi et al., 2018), and online behaviour conforms to a reinforcement-learning model by which users are reacting to social rewards, users have little reason to use their discernment to guide sharing behaviour. Thus, people will share misinformation even when they do not trust it (Pennycook et al., 2021; Ren et al., 2021).

To halt the spread, an incentive structure is needed where ‘carrots’ and ‘sticks’ are directly associated with accuracy (**Figure 4.1a right panel**, (Sharot, 2021). Such a system will work with the natural human tendency to select actions that lead to the greatest reward and avoid those that lead to punishment (Skinner, 1966). Scientists have tested different strategies to reduce the spread of misinformation, including educating people about fake news (Guess et al., 2020; Traberg et al., 2022), using a prompt to direct attention to accuracy (Kozyreva et al., 2020; Pennycook et al., 2020, 2021) and limiting how widely a post can be shared (Jackson et al., 2022). Surprisingly, possible interventions in which the incentive structure of social media platforms is altered to reduce misinformation had previously been overlooked.

Here, we test the efficacy of such a structure by slightly altering the engagement options offered to users. Specifically, we add an option to react to posts using ‘trust’ and ‘distrust’ buttons (**Figure 4.1b**). We selected these buttons because trust by definition is related to veracity - it is defined as ‘a firm belief in the reliability, truth, ability, or strength of someone or something’ (Oxford Dictionary).

We hypothesize that (1) people will use the ‘trust’ and ‘distrust’ buttons to discern true from misinformation more so than the commonly existing engagement options (such as a ‘like’ button; **Figure 4.1b, top panel**). By ‘discernment’ we mean that true posts will receive more ‘trust’ reactions than ‘distrust’ reactions and vice versa for false posts. This will create an environment in which rewards (‘trusts’) and punishments (‘distrusts’) are more directly associated with the veracity of information. Thus, (2) when exposed to this environment, users will start sharing more true information and less false information in order to obtain more ‘trust’ carrots and fewer ‘distrust’ sticks (**Figure 4.1b, bottom panel**). The new feedback options could both reinforce user behaviour that generates trustworthy material and signal to others that the post is dependable.

We also test environments in which participants receive only ‘trusts’ (a different number of *trusts* for different posts) or only ‘distrusts’ (a different number of *distrusts* for different posts) to examine if and how the impact of small vs large positive feedback (‘trust’) on discernment differs from the impact of small vs large negative feedback (‘distrust’). It has been proposed that the possibility of reward is more likely to reinforce action than the possibility of punishment, while the possibility of punishment is more likely to reinforce inaction (Guitart-Masip et al., 2011, 2012, 2014). This may translate to a large number of ‘trusts’ selectively increasing sharing of true information without decreasing sharing of misinformation and vice versa for a large number of ‘distrusts’. Further, being mindful of potential differences in sharing behaviour across political parties (Grinberg et al., 2019; Guess et al., 2020) we test participants from both sides of the political divide (Republicans & Democrats).

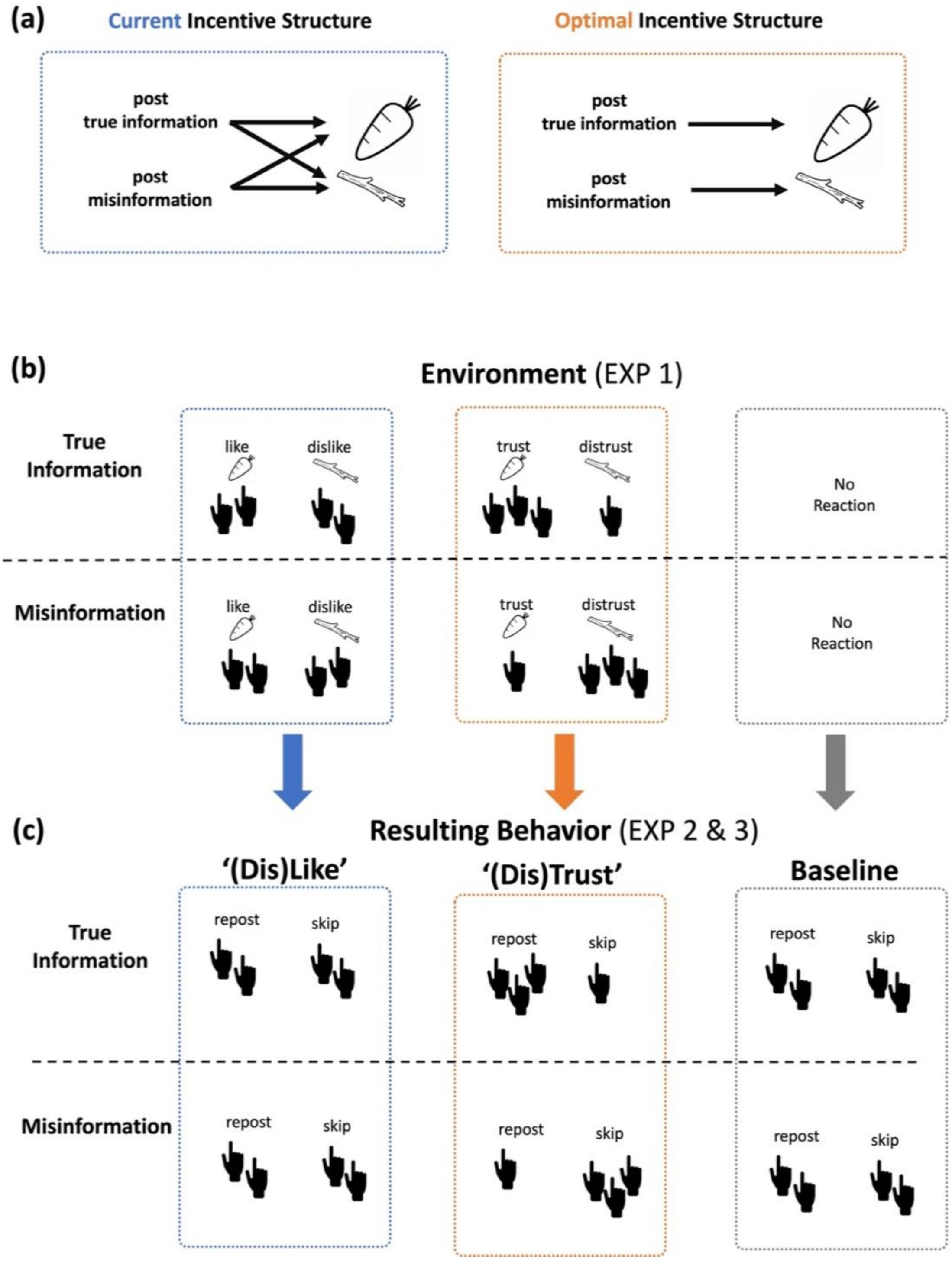

Figure 4.1. Theoretical Framework. (a) The current incentive structure (blue) is such that the veracity of shared information is dissociated from rewards ('carrots') and punishments ('sticks'). That is, true information and misinformation may lead to roughly equal number of rewards and punishments. An optimal incentive structure (orange) is such that sharing true information is rewarded with more 'carrots' than sharing misinformation, which in turn is penalized with more 'sticks' than true information. To create an optimal environment an intervention is needed by which the number of rewards and

punishments are directly associated with the veracity of information. **(b)** We test one such possible intervention (Experiment 1). In particular, we allow people to engage with posts using ‘trust’ reaction buttons and ‘distrust’ reaction buttons (orange). The rationale is that they will use these reactions to discern true from false information more so than ‘like’ and ‘dislike’ reaction buttons. **(c)** As a result, to obtain a greater number of ‘trust’ carrots and a smaller number of ‘distrust’ sticks in response to a post, people in the optimal environment (orange) will share more true than misinformation compared to those in the suboptimal environment which includes no feedback at all (grey), and those in an environment where the association between veracity of information and number of carrots and sticks is weak (blue). This second step is tested in Experiment 2 & 3.

To that end, over six experiments 951 participants engaged in simulated social media platforms where they encountered true and false information. In Experiment 1 we examined whether participants would use ‘trust’ and ‘distrust’ buttons to discern true from false information more so than existing ‘like’ and ‘dislike’ buttons (**Figure 4.1b,** replication: Experiment 4). In Experiment 2 and 3 we tested whether new groups of participants would share more true than false information in social media platforms that introduce real ‘trust’ and ‘distrust’ feedback from other participants (**Figure 4.1c,** replication: Experiment 5 & 6). The intuition is that ‘trust’ and ‘distrust’ reactions will naturally be used to indicate veracity and thus provide a reward structure contingent on accuracy, thereby reducing the sharing of misinformation and generating a healthier information ecosystem. Using computational modelling we provide insights into the specific mechanism by which our intervention improves sharing discernment.

4.3 Methods

Experimental Design

Power Calculations.

Sample sizes for all experiments were computed based on our pilot study (**see Appendix 7.3 Experiment 4-6**). Power calculations were performed using g*Power (Faul et al., 2009) to achieve power of 0.8 (beta = 0.2, alpha = 0.05; Experiment 1: partial η2 = 0.51; Experiment 2: Cohen’s d = 0.33; Experiment 3: Cohen’s d = 0.327).

Participants (Experiment 1). One-hundred and eleven participants residing in the US completed the task on *Prolific Academic* (www.prolific.com). Data of four participants who failed more than two memory checks were excluded from further analysis (**see Memory/Attention Check for details**). Thus, data of 107 participants were analysed (52 Democrats, 54 Republican, 1 Other, M_{age} = 40.579, SD_{age} ± 14.512; female = 55, male = 52; Non-White = 20, White = 87). Participants received £7.50 per hour for their participation in addition to a memory test performance-related bonus. For all experiments presented in this study, ethical approval was provided by the Research Ethics Committee at University College London and all participants gave informed consent. All experiments were performed in accordance with the principles expressed in the Declaration of Helsinki. All samples were politically balanced for Democrats and Republicans. All experiments were replicated (**see Appendix 7.3 Experiment 4-6**).

Participants (Experiment 2).

Three-hundred and twenty participants completed the task on *Prolific Academic*. Data of four participants who failed more than two memory checks were excluded from further analysis (**see Memory/Attention Check for details**). Thus, data of three-hundred and sixteen participants were analysed (146 Democrats, 142 Republican, 28 Other, M_{age} = 37.598, SD_{age} ± 13.60; female = 157, male = 157, other = 2, Non-White = 77, White = 239). Participants received £7.50 per hour for the participation in addition to a memory test performance-related bonus.

Participants (Experiment 3).

Four-hundred and nine participants completed the task on *Prolific Academic*. Data of three participants who failed more than two memory checks were excluded from further analysis (**see Participants Experiment 1 for details**). Further data of three participants who suspected that the feedback provided did not stem from real participants were excluded. Thus, data of four-hundred and three participants were analysed (194 Democrats, 197 Republican, 12 Other,

M_{age} = 35.179, SD_{age} ± 11.051; female = 204, male = 194, other = 4, Non-White = 85, White = 218). Participants received £7.50 per hour for their participation in addition to a memory test performance-related bonus.

Task (Experiment 1).

Participants engaged in a simulated social media platform where they saw 100 news posts, each consisting of an image and a headline (**see Figure 4.2 & see Appendix 7.3 Supplementary Table 4.1 for stimuli and ratings**). Half of the posts were true, and half were false. They covered a range of different topics including COVID-19, environmental issues, politics, health, and culture. They were all extracted from the fact-checking website Politifact (https://www.politifact.com). For each post, participants had the option to either 'like', 'dislike', 'trust' or 'distrust' the post, or they could choose to 'skip' the post. They could select as many options as they wished (e.g., 'like' and 'distrust') or none at all. Participants were informed that if they chose to react to a post other users would be able to see their reactions. They were asked to treat the platform as they would any other social media platform. The order in which reaction buttons appeared on screen was counterbalanced across participants. Participants also indicated their age, gender, ethnicity, and political orientation. The task was coded using the *Qualtrics* online platform (https://www.qualtrics.com).

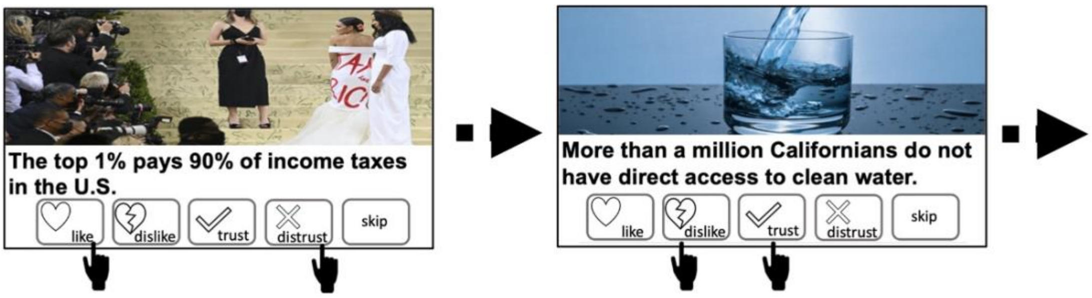
The top 1% pays 90% of income taxes in the U.S.
like
dislike
trust
distrust
skip
More than a million Californians do not have direct access to clean water.
like
dislike
trust
distrust
skip

Figure 4.2. Task (Experiment 1). Participants observed a series of 100 posts in random order (50 true, 50 false). Their task was to react using one or more of the 'like', 'dislike', 'trust' or 'distrust' buttons or to skip the post. The task was self-paced.

Memory/Attention check. At the end of the experiment, participants were presented with five posts and had to indicate whether these were old or new. This is to ensure that participants were attentive during the experiment. Participants who failed more than two of the memory checks were excluded from the analysis.

Task (Experiment 2).

In Experiment 2 participants engaged in a simulated social media platform where they saw the same 100 posts (50 true, 50 false) shown to participants in Experiment 1. Participants had to either ‘repost’ or ‘skip’ each post (**see Figure 4.3**). They were told that if they decided to repost, then the post would be shared to their feed, and they would observe other participants’ reactions to it. We used a between-subject design with five environments. Depending on the environment participants were randomly assigned to, they could either see (i) how many people *disliked* the post, (ii) how many people *liked* the post, (iii) how many people *distrusted* the post, or (iv) how many people *trusted* the post. We also included a *Baseline* environment, in which participants received no feedback. Due to logistic constraints, the feedback was not collected in real time but was instead taken from participants’ reactions in Experiment 1. The participants, however, believed the reactions were provided in real time as indicated by a rotating cogwheel (1s). If participants selected to skip, they would also observe a rotating cogwheel (1s) and then a screen asking them to click continue. The average duration of the white screen (M = 2.351s; SE = 0.281) was not different from the average duration of feedback (M = 2.625s; SE = 0.245; t(233) = 0.853, p = 0.395, Cohen’s d = 0.056). Though the duration of trials in which participants chose to skip (M = 9.046s, SE = 0.38) was slightly shorter than those in which they chose to share (M = 9.834s, SE = 0.358; t(233) = 2.044, p = 0.042, Cohen’s d = 0.134). Thereafter, participants were presented with all the posts again and asked to indicate if they believed the post was accurate or inaccurate on a continuous scale from *0 = inaccurate* to *100 = accurate*. Finally, participants completed a short demographic questionnaire assessing age, gender, ethnicity, and political orientation. The task was self-paced. The task was coded using *JsPsych* and *Javascript*.

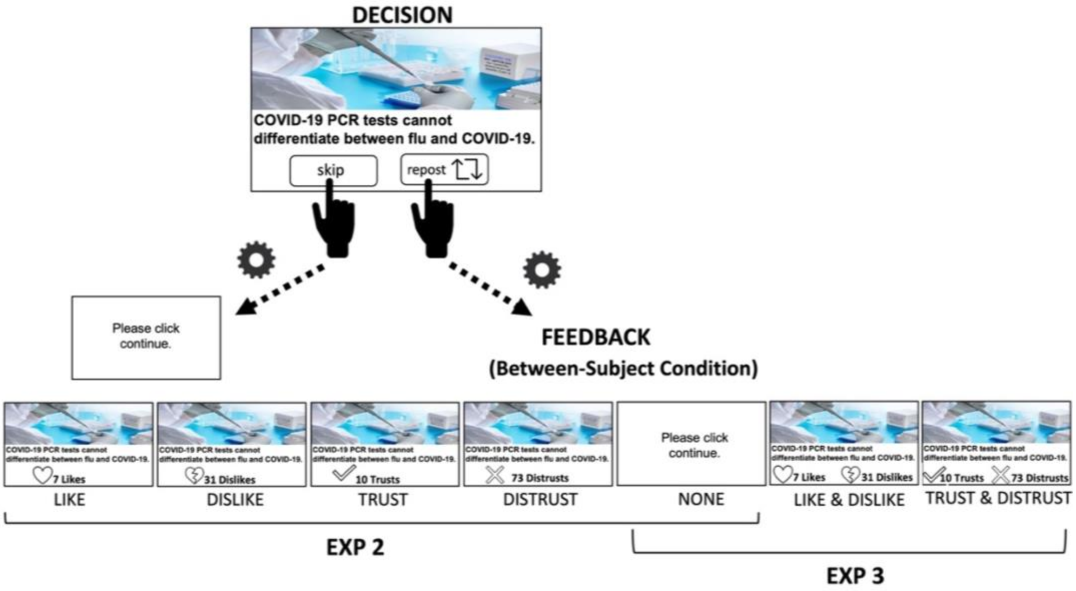

Figure 4.3. Task (Experiment 2 and 3). In Experiment 2 on each of 100 trials participants observed a post (50 true, 50 false). They then choose whether to share it or skip (self-paced). They were told that if they chose to share a post, it would be shared to their feed such that the other participants would be able to see the post and react to it in real time (feedback). Depending on the environment participants were in, they could either observe the number of (i) 'dislikes' (N = 45), (ii) 'likes' (N = 89), (iii) 'distrusts' (N = 49), or (iv) 'trusts' (N = 46) feedback. The feedback was in fact the number of reactions gathered from Experiment 1, though the participants believed the reactions were in real time as indicated by a rotating cogwheel (1s). Once the feedback appeared, participants could then click continue (self-paced). If participants selected to skip, they would observe a white screen asking them to click continue (self-paced). In the Baseline environment (N = 59) participants received no feedback. Experiment 3 was identical to Experiment 2 with two distinctions: (1) Depending on the environment participants were in, they could either observe the number of (i) both 'dislikes' and 'likes' (N = 128), (iii) both 'distrusts' and 'trusts' (N = 137) or (iii) no feedback (Baseline, N = 126). (2) In Experiment 3 we selected 40 posts (20 true, 20 false) to which Republicans and Democrats had on average reacted to similarly using the 'trust' button in Experiment 1. Discernment was calculated for each participant by subtracting the proportion of sharing false information from the proportion of sharing true information. High discernment indicates greater sharing of true than false information.

Task (Experiment 3).

Experiment 3 (**see Figure 4.3**) was identical to the task used in Experiment 2 with three exceptions:

(1) We selected 40 posts (20 true, 20 false), in which there was no significant difference in the way Republicans and Democrats reacted to them using the trust button during Experiment 1. This was done by entering participants' trust responses (0/1) into a vector for Democrats and Republicans for each post. We then performed Pearson Chi Square Tests for each of the 100 posts to identify whether Democrats and Republicans used the trust button differently. Posts where no significant difference was observed were included in Experiment 3.

(2) Three environments were included: a *Baseline* environment, in which participants received no feedback, a 'Trust & Distrust' environment, in which participants received both *Trust* and *Distrust* feedback whenever they chose to share a post, and a 'Like & Dislike' environment, in which participants received *Like* and *Dislike* feedback whenever they chose to share a post.

(3) At the end of the experiment, we asked participants: (1) "What do you think the purpose of this experiment is?"; and (2) "Did you, at any point throughout the experiment, think that the experimenter had deceived you in any way? If yes, please specify."

Statistical Analysis

Statistical Analysis (Experiment 1).

We examined whether participants used the different reaction buttons to discern true from false information. For positive reactions (e.g., *'likes'* and *'trusts'*) discernment is equal to the proportion of those reactions for true information minus false information, and vice versa for negative reactions (*'dislikes'* and *'distrusts'*). Proportions were calculated for each participant and then entered into a 2 (**type of reaction**: 'trust' and 'distrust' / 'like' and 'dislike') by 2 (**valence:** positive, i.e., 'like', 'trust'/ negative, i.e., 'dislike', 'distrust') within-subject ANOVA. Political orientation was also added as a between-subject factor (Republican/Democrat), allowing for an interaction of political orientation and type of reaction to assess whether participants with differing political beliefs used the reaction buttons in different ways. We performed one-sample t-tests to compare discernment (equal to the proportion of those reactions for true information minus false information, and vice versa for negative reactions) against zero to assess whether each reaction discerned between true and false information. To examine whether participants' frequency of use of each reaction option differed we again ran a within-subject ANOVA, but this time with percentage frequency of reaction option used as the dependent variable. We computed a Pearson's correlation across participants between frequency of skips and discernment.

One participant selected 'other' for political orientations. This participant was not included in the analysis because political orientation was included in the analysis, and such small group sizes could heavily skew results. Analysis was conducted using IBM SPSS 27 and R Studio (Version 1.3.1056). All statistical tests conducted in the present article are two-sided. All results of interest hold when controlling for demographics (age, gender and ethnicity, when not including political orientation in the analysis, and, if applicable, when allowing for an interaction between type of feedback and valence.

Discernment Analysis (Experiment 2 and 3).

Discernment is calculated for each participant by subtracting the proportion of sharing false information from the proportion of sharing true information. High discernment indicates greater sharing of true than false information. In Experiment 2 scores were submitted to an ANOVA with type of feedback ('(Dis)Trust' vs '(Dis)Like' vs Baseline), valence of feedback (positive, i.e., 'like', 'trust' vs negative, i.e., 'dislike', 'distrust'), political orientation and an interaction of political orientation and type of feedback. To assess whether frequency of posts shared differed we used the same ANOVA, this time with percentage of posts shared out of all trials as the dependent variable.

To test whether '(Dis)Trust' feedback improves belief accuracy, we transformed participants' belief ratings (which were given on a scale from 'post is accurate' =100 to 'post is inaccurate' = 0) to indicate error. If the post was false (inaccurate) error was equal to the rating itself, if the post was true (accurate) error was equal to 100 minus the rating. Participants' average error scores were then entered into a between-subject ANOVA with type of feedback (Baseline, '(Dis)Trust', '(Dis)Like'), valence of feedback, political orientation and an interaction of political orientation and type of feedback.

Analysis of Experiment 3 followed that of Experiment 2 with the difference being that we had three types of feedback environments (Baseline, 'Like & Dislike', 'Trust & Distrust') and of course no valence of feedback (as all environments were mixed valence or no valence). Data of participants who selected 'other' for political orientations (Experiment 2 = 28, Experiment 3 = 12) were not analysed, because political orientation was included in the analyses variable, and small group sizes of 'other' could heavily skew results.

Drift-Diffusion Modelling (Experiment 2 and 3).

To assess whether being exposed to an environment with '(Dis)Trust' feedback impacted the parameters of the evidence accumulation process in our data compared to Baseline and '(Dis)Like' feedback we modelled our data using DDM. To that end, we ran three separate models – one for each type of feedback and included the following parameters: (1) t0—amount of non-decision time; (2) α—distance between decision thresholds; (3) *z*—starting point of the accumulation process; and (4) v—drift rate, i.e., the rate of evidence accumulation.

We used the HDDM software toolbox (Wiecki et al., 2013) to estimate the parameters of our models. The HDDM package employs hierarchical Bayesian parameter estimation, using MCMC methods to sample the posterior probability density distributions for the estimated parameter values. We estimated both group-level and individual-level parameters. Parameters for individual participants were assumed to be randomly drawn from a group-level distribution. Participants' parameters both contributed to and were constrained by the estimates of group-level parameters. In fitting the models, we used priors that assigned equal probability to all possible values of the parameters. Models were fit to log-transformed RTs. We sampled 20,000 times from the posteriors, discarding the first 5000 as burn in and thinning set at 5. MCMCs are guaranteed to reliably approximate the target posterior density as the number of samples approaches infinity. To test whether the MCMC converged within the allotted time, we used Gelman–Rubin statistic (Rubin & Gelman, 1992) on five chains of our sampling procedure. The Gelman–Rubin diagnostic evaluates MCMC convergence by analysing the difference between multiple Markov chains. The convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter. In each case, the Gelman–Rubin statistic was close to one (<1.1), suggesting that MCMC were able to converge.

We then compared parameter estimates using 95% CIs. Specifically, for each parameter in each group (‘(Dis)Trust’ vs ‘(Dis)Like’, ‘(Dis)Trust’ vs Baseline, ‘(Dis)Like’ vs Baseline) we calculated the 95% CI. If the 95% CI of two groups do not overlap, we consider there to be a significant difference between the two feedback types compared. We also calculated 95% HDIs. For each comparison (‘(Dis)Trust’ vs ‘(Dis)Like’, ‘(Dis)Trust’ vs Baseline, ‘(Dis)Like’ vs Baseline) we calculated the difference in the posterior distributions and reported the 95% HDI of the difference. If this HDI did not overlap zero, we consider there to be a meaningful difference between the two feedback types compared**.** HDI testing was conducted in R using *HDInterval* (Meredith & Kruschke, 2016).

To validate the DDM, we used each group's parameters obtained from participants' data to simulate log-transformed RTs and responses separately for each feedback type. We used the exact number of participants and total number of trials as in the experiments. Simulated data were then used to (1) perform model recovery analysis and (2) to compare the pattern of participants' response to the pattern of simulated responses, separately for each group. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. Simulation and model recovery analysis were performed using the HDDM software toolbox (Wiecki et al., 2013). One-way ANOVAs were computed to examine if simulated data reproduced the behavioural pattern from the experimental data. To that end, discernment was entered into a one-way

ANOVA with type of feedback as the independent variable for Experiment 2 and 3 separately. Note, that we did not enter veracity of the post into our DDM and instead entered responses as either ‘veracity-promoting’ (true post shared or false post skipped) or ‘veracity-obstructing’ (false post shared or true post skipped). Thus, discernment was calculated as the proportion of true posts shared and false posts skipped minus the proportion of true posts skipped and false posts shared.

4.4 Results

Participants’ use ‘trust’ and ‘distrust’ buttons to discern true from false information (Experiment 1).

In a first step, we examined whether participants used ‘trust’ and ‘distrust’ reactions to discern true from false information more so than ‘like’ and ‘dislike’ reactions. In Experiment 1, participants saw 100 news posts taken from the fact-checking website Politifact (https://www.politifact.com; **see Figure 4.2**). Half of the posts were true, and half were false. Participants were given the opportunity to react to each post using ‘like’, ‘dislike’, ‘trust’ and ‘distrust’ reaction buttons. They could select as many buttons as they wished or none at all (‘skip’). Five participants were excluded according to pre-determined criteria (**see Methods for details**). Thus, one-hundred and six participants (52 Democrats, 54 Republican, M_{age} = 40.745, SD_{age} ± 14.479; female = 54, male = 52) were included in the analysis.

We then examined whether participants used the different reaction buttons to discern true from false information. Discernment was calculated as follows, such that high numbers always indicate better discernment:

For ‘like’:

$$Discernment = Prop_{likes\,true} - Prop_{likes\,false}$$

For ‘dislike’:

$$Discernment = Prop_{dislikes\ false} - Prop_{dislikes\ true}$$

For ‘trust’:

$$Discernment = Prop_{trusts\ true} - Prop_{trusts\ false}$$

For ‘distrust’:

$$Discernment = Prop_{distrusts\ false} - Prop_{distrusts\ true}$$

With *Prop* indicating the proportion of that response out of all true posts, or out of all false posts, as indicated.

These discernment scores were calculated for each participant separately and then entered into a 2 (**type of reaction**: ‘trust’ and ‘distrust’ / ‘like’ and ‘dislike’) by 2 (**valence:** positive, i.e., ‘like’, ‘trust’/ negative, i.e., ‘dislike’, ‘distrust’) within-subject ANOVA. Political orientation was also added as a between-subject factor (Republican/Democrat), allowing for an interaction of political orientation and type of reaction to assess whether participants with differing political beliefs used the reaction buttons in different ways.

The results reveal that participants’ use of ‘(Dis)Trust’ reaction buttons (M = 0.127; SE = 0.007) was more discerning than their use of ‘(Dis)Like’ reaction buttons (M = 0.047; SE = 0.005; F(1,104) = 95.832, p < 0.001**,** partial η2 = 0.48 **Figure 4.4**). We additionally observed an effect of valence (F(1,105) = 17.33, p < 0.001, partial η2 = 0.14), with negatively valenced reaction buttons (e.g., ‘dislike’ and ‘distrust’ M = 0.095, SE = 0.007) being used in a more discerning manner than positively valenced reaction buttons (e.g., ‘like’ and ‘trust’, M = 0.087, SE = 0.005) and an effect of political orientation (F(1,104) = 25.262, p < 0.001, partial η2 = 0.2), with Democrats (M = 0.115, SE = 0.007) being more discerning than Republicans (M = 0.06, SE = 0.005). There was also an interaction of type of reaction and political orientation (F(1,104) = 24.084, p < 0.001, partial η2 = 0.19), which was characterized by Democrats showing greater discernment than Republicans in their use of ‘(Dis)Trust’ reaction buttons (F(1,104) = 33.592, p < 0.001, partial η2 = 0.24), but not in their use of ‘(Dis)Like’ reaction buttons (F(1,104) = 2.255, p = 0.136, partial η2 = 0.02).

Importantly, however, both Democrats (F(1,51) = 93.376, p < 0.001, partial η2 = 0.65) and Republicans (F(1,53) = 14.715, p < 0.001, partial η2 = 0.22) used the ‘(Dis)Trust’ reaction buttons in a more discerning manner than the ‘(Dis)Like’ reaction buttons.

One-sample t-tests against zero further revealed that participants’ use of each reaction button discerned true from false information (‘like’: M = 0.06, SE = 0.006, t(105) = 10.483, p < 0.001, Cohen’s d = 1.018; ‘trust’: M = 0.099; SE = 0.01, t(105) = 9.744, p < 0.001, Cohen’s d = 0.946; ‘dislike’: M = 0.034; SE = 0.007; t(105) = 4.76, p < 0.001, Cohen’s d = 0.462; ‘distrust’: M = 0.156; SE = 0.01, t(105) = 15.872, p < 0.001, Cohen’s d = 1.542).

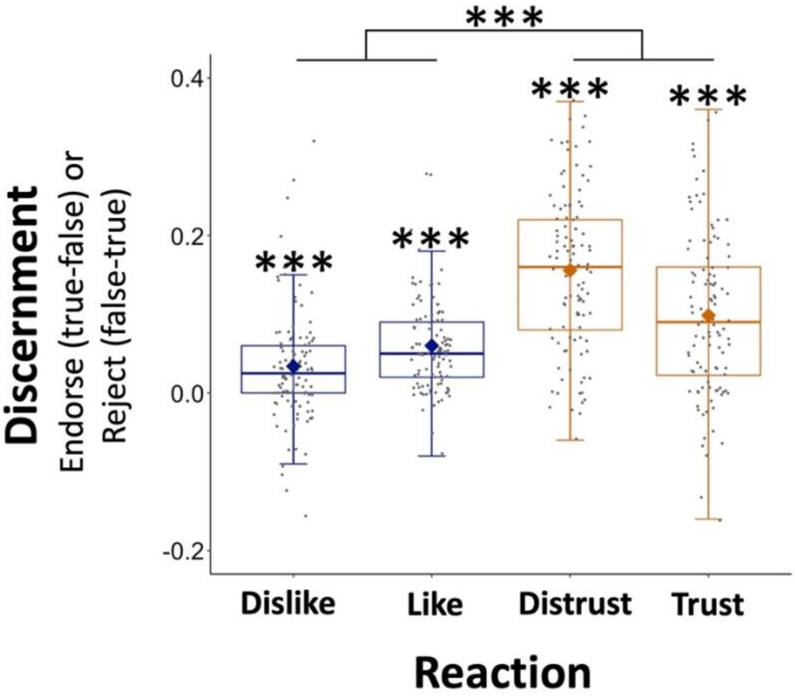

Figure 4.4. Participants use ‘trust’ and ‘distrust’ reactions to discern true from false information. ‘Trust and ‘distrust’ reactions were used in a more discerning manner than ‘like’ and ‘dislike’ reactions. Y axis shows discernment between true and false posts. For positive reactions (e.g., ‘likes’ and ‘trusts’) discernment is equal to the proportion of positive reactions for true information minus false information, and vice versa for negative reactions (‘dislikes’ and ‘distrusts’). X axis shows reaction options. Data are plotted as boxplots for each reaction button, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean discernment per reaction. Individuals' mean

discernment data are shown separately as grey dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001.

Thus far, we have shown that participants use ‘(Dis)Trust’ reaction buttons in a more discerning manner than ‘(Dis)Like’ reaction buttons. As social media platforms care about overall engagement not only its quality, we examined how frequently participants used the different reaction buttons. An ANOVA with the same specifications as above was conducted, but this time submitting frequency of reaction as the dependent variable. We found that participants used ‘(Dis)Trust’ reaction buttons more often than ‘(Dis)Like’ reaction buttons (Percentage use of reaction out of all trials: ‘trust’: M = 28.057%; ‘distrust’: M = 34.085%; ‘like’: M = 18.604%; ‘dislike’: M = 23.745%; F(1,104) = 36.672, p < 0.001, partial η2 = 0.26). In addition, negative reaction buttons (‘distrust’ and ‘dislike’: M = 28.915%, SE = 1.177) were used more frequently than positive reaction buttons (‘trust’ and ‘like’: M = 23.33%, SE = 1.133; F(1,105) = 16.96, p < 0.001, partial η2 = 0.07). No other effect was significant. Interestingly, we also found that participants who skipped more posts were less discerning (R = -0.414, p < 0.001). Together, the results show that the new reaction options increase engagement.

The results hold when controlling for demographics, when not including political orientation in the analysis, and allowing for an interaction between type of reaction and valence (**see Appendix 7.3 Supplementary Tables 4.2 & 4.3**). The results also replicate in an independent sample (**see Appendix 7.3 Experiment 4**).

‘Trust’ and ‘distrust’ incentives improve discernment in sharing behaviour (Experiment 2).

Thus far, we have shown that participants use ‘(Dis)Trust’ reaction buttons in a more discerning manner than ‘(Dis)Like’ reaction buttons. Thus, an environment which offers ‘(Dis)Trust’ feedback is one where the number of ’carrots’ (in the form of ‘trusts’) and the number of ‘sticks’ (in the form of ‘distrusts’) are directly associated with the veracity of the posts. It then follows that submitting participants to such an environment will increase their sharing of true information (to receive 'trusts') and reduce their sharing of misinformation (to avoid 'distrusts').

To test this, we ran a second experiment. A new group of participants (N = 320) were recruited to engage in a simulated social media platform. They observed the same 100 posts (50 true, 50 false) shown to the participants in Experiment 1, but this time instead of reacting to the posts they could either share the post or skip it (**see Figure 3**). They were told that if they chose to share a post, it would be shared to their feed such that the other participants would be able to see the post and would then be able to react to it in real time (*feedback*). Depending on the environment participants were in, which varied between-subjects, they could receive feedback in the form of the number of users who (i) '*disliked'*, or (ii) '*liked'*, or (iii) '*distrusted'*, or (iv) '*trusted'* their posts. We also included a (v) baseline condition, in which participants received no feedback. If participants selected to skip, they would observe a white screen asking them to click continue. Data of 32 participants were not analysed according to pre-determined criteria (**see Methods for details**). Two-hundred and eighty-eight participants (146 Democrats, 142 Republicans, M_{age} = 38.073, SD_{age} ± 13.683; female = 147, male = 141) were included in the analysis (**see Methods for details**).

$$Discernment\ =\ Prop_{reposts\ true} - Prop_{reposts\ false}$$

These scores were submitted to a between-subject ANOVA with type of feedback ('trust' & 'distrust'/ 'like' & 'dislike'/ Baseline), valence (positive, i.e., 'like' & 'trust' vs negative, i.e., 'dislike', 'distrust' vs neutral i.e., no feedback) and political orientation (Republican/Democrat) as factors. We also allowed for an interaction of political orientation and type of feedback.

We observed an effect of type of feedback (F(1,281) = 15.2, p < 0.001, partial η2 = 0.051), such that participants shared more true than false information in the '(Dis)Trust' environments (M = 0.18, SE = 0.018) than the '(Dis)Like' environments (M = 0.085, SE = 0.019, F(1,225) = 14.249, p < 0.001, partial η2 = 0.06) and Baseline environment (M = 0.084, SE = 0.025; F(1,150) = 10.906, p = 0.001, partial η2 = 0.068, **Figure 5a**). Moreover, participants who received 'trust' feedback (M = 0.176, SE = 0.026) were more discerning in their sharing behaviour than those who received 'like' feedback (M = 0.081, SE = 0.021, F(1,131) = 10.084, p = 0.002, partial η2 = 0.071). Those who received 'distrust' feedback (M = 0.175, SE = 0.026) were more discerning than those who received 'dislike' feedback (M = 0.092, SE = 0.039, F(1,90) = 5.003, p = 0.028, partial η2 = 0.053). We further observed a trend interaction between type of feedback and political orientation (F(1,281) = 2.939, p = 0.055, partial η2 = 0.02). While Democrats (M = 0.213; SE = 0.014) were generally more discerning than Republicans (M = 0.017; SE = 0.016; F(1,281) = 77.392, p < 0.001, partial η2 = 0.216), this difference was smaller in those who received '(Dis)Trust' feedback (M = 0.082, SE = 0.034) compared to those who received '(Dis)Like' feedback (M = 0.23, SE = 0.03; F(1,224) = 4.879, p = 0.028, partial η2 = 0.021) and by trend smaller than those who received no feedback (M = 0.229, SE = 0.045; F(1,149) = 3.774, p = 0.054, partial η2 = 0.025). There was no difference between the latter two (F(1,188) = 0.00, p = 0.988, partial η2 = 0.00). No other effects were significant. Overall engagement, measured as percentage of posts shared out of all trials, did not differ across environments (F(1,281) = 1.218, p = 0.271, partial η2 = 0.004; Mean % posts shared out of all trials: Baseline = 27.712%; Dislike = 35.889%; Like = 33.258%; Distrust = 32.51%; Trust = 30.435%; **see Appendix 7.3 Supplementary Table 4.4 for means for true and false posts**).

Results hold when controlling for demographics, when not including political orientation in the analysis, and allowing for an interaction between type of reaction and valence (**see Appendix 7.3 Supplementary Table 4.5 & 4.6**). Results replicate in an independent sample (**see Appendix 7.3 Experiment 5).**

To recap - participants in Experiment 2 decided whether to share content or skip. They then observed the reactions of other participants to their post (they believed this was happening in real-time, but for simplicity we fed them reactions of participants from Experiment 1). Each participant in Experiment 2 observed only one type of feedback. For example, only ‘distrusts’. How is it that observing ‘distrusts’ alone increases discernment? The rationale behind this design is that for any given post, true or false, some users will distrust the post. However, true posts will receive fewer ‘distrusts’ than false posts. It is the number of ‘distrusts’ per post that matters. Participants are motivated to minimize the average number of ‘distrusts’ they receive. To achieve this, they should post more true posts and fewer false posts. Of course, if participants were simply trying to minimize the *total* number of distrusts, they would just skip on every trial. Participants do not do that, however. Potentially because sharing in and of itself is rewarding (Tamir & Mitchell, 2012). The results indicate that participants are sensitive to the number of ‘distrusts’ per posts not just to the total number of ‘distrusts’ over all posts.

The same rationale holds for the participants that only observe ‘trusts’. They receive more ‘trusts’ for true than false posts. It is the magnitude of ‘trusts’ that is associated with veracity. This motivates participants to post more true posts and fewer false posts in order to maximize the average number of ‘trusts’ per post. Of course, if participants were simply trying to maximize the total number of ‘trusts’, they would just share on every trial. Participants do not do that, however. This indicates that they are sensitive to the number of ‘trusts’ per post not just to total number over all posts. Any user of social media platforms could relate to this; when posting a tweet, for example, many people will be disappointed with only a handful of ‘hearts’. The user’s goal is to maximize positive feedback per post. The same rationale as above holds for ‘likes’ and ‘dislikes’ except that those are less associated with veracity, thus impact discernment less.

The posts included in the experiment covered a range of topics including politics, science, health, environment, and society. As observed in **Figure 4.5b**, the effect of ‘(Dis)Trust’ environment on discernment is observed regardless of content type.

Thus far, our results show that changing the incentive structure of social media platforms by coupling the number of ‘carrots’ and ‘sticks’ with information veracity could be a valuable tool to reduce the spread of misinformation. If feedback promotes discernment in sharing behaviour, it is plausible that it may in turn improve belief accuracy. To test this, we asked participants at the end of the experiment to indicate how accurate they thought a post was on a scale from *inaccurate* (0) to *accurate* (100). Participants’ error in estimating whether a post was true or false was calculated as follows: for false posts error was equal to the participants’ accuracy rating and for true posts it was equal to 100 minus their rating. Participants’ average error scores were entered into a between-subject ANOVA with type of feedback and valence of feedback, as well as political orientation and its interaction with feedback type. We observed an effect of type of feedback (F(1,281) = 7.084, p = 0.008, partial η2 = 0.025), such that participants were more accurate (less errors) when they received ‘(Dis)Trust’ feedback (M = 47.24, SE = 0.938) compared to ‘(Dis)Like’ feedback (M = 50.553, SE = 0.851, F(1,224) = 7.024, p = 0.009, partial η2 = 0.03). We further observed an effect of political orientation (F(1,281) = 11.402, p < 0.001, η2 = 0.039), with Democrats (M = 47.264, SE = 0.773) being more accurate than Republicans (M = 51.117, SE = 0.802). No other effects were significant. All results hold when controlling for demographics, when not including political orientation in the analysis, and allowing for an interaction between type of feedback and valence (**see Appendix 7.3 Supplementary Table 4.7)**. Results are replicated in an independent sample (**see Appendix 7.3 Experiment 5**)**.**

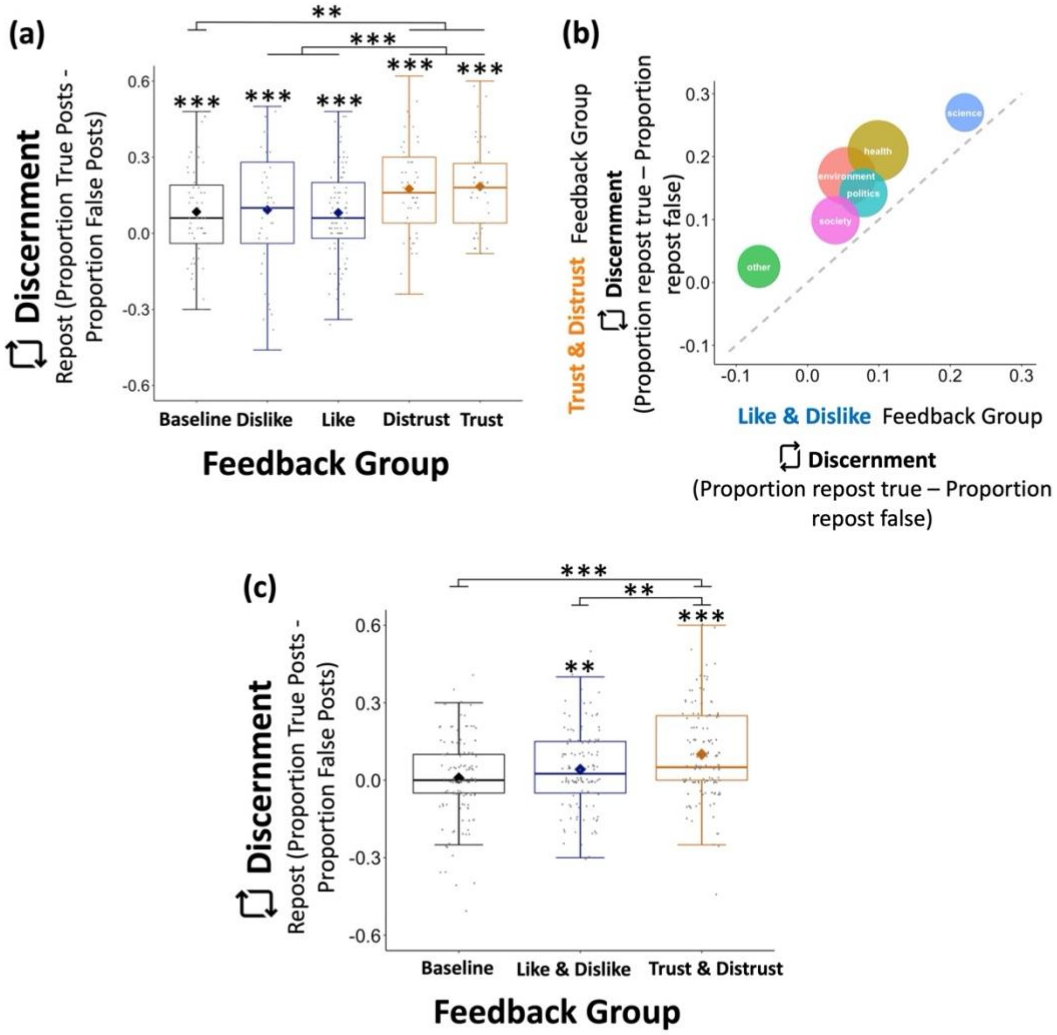

Figure 4.5. Altering the incentive structure of social media environments increases discernment of information shared. (a) Participants operating in an environment where ‘(Dis)Trust’ feedback was introduced shared more true information relative to false information than participants operating in an environment where only ‘(Dis)Like’ feedback was available, or no feedback at all (Baseline). Y axis shows discernment, i.e., proportion of true posts shared minus proportion of false posts shared. X axis shows the group environment (type of feedback). **(b)** This was the case regardless of the topic of the post (politics = turquoise, science = blue, health = olive, environment = salmon, society = pink, other = green). Bubble size corresponds to number of the posts included in the study. Diagonal dashed line indicates point of equivalence, where discernment in equal across the ‘(Dis)Like’ and ‘(Dis)Trust’ environments. As can be seen, all bubbles are above the dashed line indicating that in all cases discernment is greater in an environment that offers ‘(Dis)Trust’ feedback. Y axis shows discernment in the ‘(Dis)Trust’ environment, X axis shows discernment in the ‘(Dis)Like’ environment. **(c)** Experiment 3 showed the same results as Experiment 2. Data are plotted as boxplots for each reaction, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond

shape indicates the mean discernment per reaction. Individuals' mean discernment data are shown separately as grey dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001, **p < 0.01.

‘Trust’ and ‘distrust’ incentives together improve discernment in sharing behaviour (Experiment 3).

Given that Experiment 2 revealed that receiving ‘trust’ or ‘distrust’ feedback separately improves discernment, it is likely that the coupled presentation of both will jointly also improve discernment. To test this, we ran a third experiment with a new group of participants. The task was identical to Experiment 2 (**see Figure 4.3**), but this time we included three between-subject environments: a *Baseline* environment, in which participants received no feedback, a 'Trust & Distrust’ environment, in which participants observed both the number of *trust* and the number of *distrust* feedback, and a ‘Like & Dislike’ environment, in which participants observed both the number of *like* and the number of *dislike* feedback.

Additionally, to ensure posts align equally with Democratic and Republican beliefs, in Experiment 3 we selected 40 posts (20 true, 20 false) in response to which Republicans and Democrats utilized the ‘trust’ button in a similar manner in Experiment 1 (**see Methods**). Data of 18 participants were not analysed according to pre-determined criteria (**see Methods for details**). Analysis of Experiment 3 (N = 391, 194 Democrats, 197 Republican, M_{age} = 35.304, SD_{age} ± 11.089; female = 196, male = 192, other = 3) was the same as in Experiment 2 except that there were three environments (Baseline, ‘Like & Dislike’, ‘Trust & Distrust’) and no valence of feedback, because all environments either include both positive and negative feedback or no feedback.

Discernment was submitted to a between-subject ANOVA with **type of feedback** (Baseline/ ‘Like & Dislike’ / ‘Trust & Distrust’), political orientation and their interaction as factors. Again, we observed an effect of type of feedback (F(1,385) = 11.009, p < 0.001, partial η2 = 0.054, **Figure 4.5c**), with participants in the ‘Trust & Distrust’ feedback group posting more true relative to false information (M = 0.101, SE = 0.015) than those in the ‘Like & Dislike’ group (M = 0.042, SE = 0.013; F(1,261) = 8.478, p = 0.00, partial η2 = 0.031) or those who received no feedback at all (M = 0.008, SE = 0.014, F(1,259) = 20.142, p < 0.001, partial η2 = 0.0724). By contrast there was no difference between the latter two groups (F(1,250) = 2.981, p = 0.085, partial η2 = 0.012). As observed in Experiment 2, Democrats (M = 0.073, SE = 0.011) were more discerning than Republicans (M = 0.031, SE = 0.012; F(1,385) = 6.409, p = 0.012, partial η2 = 0.016). No other effects were significant.

Interestingly participants shared more frequently in the ‘Trust & Distrust’ environment compared to the other two environments (% of all trials: ‘Trust & Distrust’ = 36.2%, ‘Like & Dislike’ = 30.41%; Baseline = 25.853%; F(1,385) = 8.7692, p < 0.001, partial η2 = 0.044). This illustrates that ‘(Dis)Trust’ feedback improves discernment without reducing engagement. No other effects were significant.

All results hold when controlling for demographics, when not including political orientation in the analysis, and allowing for an interaction between type of reaction and valence (**see Appendix 7.3 Supplementary Tables 4.8 & 4.9**). Results replicate in an independent sample (**see Appendix 7.3 Experiment 6).**

At the end of Experiment 3, we again asked participants to indicate how accurate they thought a post was. Participants’ average error scores were calculated as in Experiment 2 and entered into a between-subject ANOVA with type of feedback, political orientation and their interaction as factors. Democrats (M = 40.591; SE = 6.371) were more accurate than Republicans (M = 42.056; SE = 5.633; F(1,385) = 5.723, p = 0.017, partial η2 = 0.015). No other effects were significant (for results when controlling for demographics, when not including political orientation in the analysis **see Appendix 7.3 Supplementary Table 4.10)**. Note, that in the replication study (Experiment 6) we did observe an effect of type of feedback (F(1,147) = 4.596, p = 0.012, partial η2 = 0.059), with ‘(Dis)Trust’ being most accurate. Thus, we see accuracy effects in three (Experiment 2, 5, 6) out of our four studies.

Taken together these findings suggest that changing the incentive structure of social media platforms, such that 'carrots' and 'sticks' are strongly associated with veracity promotes discernment in sharing behaviour, thereby reducing the spread of misinformation.

'(Dis)Trust' incentives improve discernment in sharing behaviour by increasing the relative importance of evidence consistent with discerning behaviour. Next, we set out to characterize the mechanism by which the new incentive structure increased discernment. Imagine you observe a post on social media, and you need to decide whether to share it – how do you make this decision? First, you examine the post. Second, you retrieve your existing knowledge. For example, you may think about what you already know about the topic, what you heard others say, you may try to estimate how others will react to the post if you share it, and so on. This process is called 'evidence accumulation' - you gradually accumulate and integrate external evidence and internal evidence (memories, preferences etc.) to decide. Some of the evidence you retrieve will push you towards a 'good' response that promotes veracity (that is posting a true post and skipping a false post) and some will push you towards a 'bad' response that obstructs veracity (that is posting a false post and skipping a true post). We can think of the evidence that pushes you toward a response that promotes veracity as 'signal'. Using computational modelling it is possible to estimate how much a participant is influenced ('pushed') by signal relative to noise, by calculating a parameter known as a 'drift rate' in a class of models known as DDM. One possibility then is that in the '(Dis)Trust' environment evidence towards responses that promote veracity is given more weight than towards responses that obstruct veracity (that is the drift rate is larger), thus people make more discerning decisions.

Another, non-exclusive possibility, is that in the '(Dis)Trust' environment participants are more careful about their decisions. They require more evidence before making a decision. For example, they may spend more time deliberating the post. In DDM this is estimated by calculating what is known as the distance between the decision thresholds (that is how much total distance do I need to be 'pushed' in one direction or the other to finally make a choice).

To test the above possible mechanisms, we modelled our data using the DDM (Lin et al., 2023; Ratcliff, 1978; Ratcliff & McKoon, 2008). We modelled participants' responses ('veracity-promoting' vs 'veracity-obstructing' choice) separately for each type of feedback ('(Dis)Trust', '(Dis)Like', Baseline) and each Experiment (Experiment 2 and 3). The following parameters were included: (1) t0—the amount of non-decision time, capturing encoding and motor response time; (2) α—the distance between decision thresholds ('veracity-promoting' response vs 'veracity-obstructing' response); (3) z—starting point of the accumulation process; and (4) v—the drift rate (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013).

We next examined which of the parameters were different in the different environments (**see Table 4.1 & 4.2**). To that end, we calculated 95% CIs of each parameter for each pair of incentive environments ('(Dis)Trust' vs '(Dis)Like', '(Dis)Trust' vs Baseline, '(Dis)Like' vs Baseline). If the 95% CIs do not overlap, we infer a significant difference between the two incentive environments.

For both Experiment 2 (**see Figure 4.6a**) and Experiment 3 (**see Figure 4.6c**) we observed a significant difference in the drift rate. In particular, in the '(Dis)Trust' environments the drift rate was larger (Experiment 2: v = 0.216; 95% CI [0.17, 0.262]; Experiment 3: v = 0.12; 95% CI [0.086, 0.155]) than in the'(Dis)Like' environments (Experiment 2: v = 0.01; 95% CI [0.056, 0.145]; Experiment 3: 0.037; 95% CI [0.002, 0.069]) or no feedback environment (Experiment 2: v = 0.098; 95% CI [0.039, 0.158]; Experiment 3: v = 0.006; 95% CI [-0.027, 0.037]). The Baseline and '(Dis)Like' environments did not differ for drift rate. This suggests that relative to the other environments, in the '(Dis)Trust' environments evidence consistent with a 'veracity-promoting' response is weighted more than 'evidence' consistent with a 'veracity-obstructing' response. 95% HDI comparisons corroborate this result **(see**

Appendix 7.3 Supplementary Tables 4.11 & 4.12 for HDI Comparisons). We replicate these results in Experiment 5 and Experiment 6 **see Appendix 7.3 Experiment 5 & 6**).

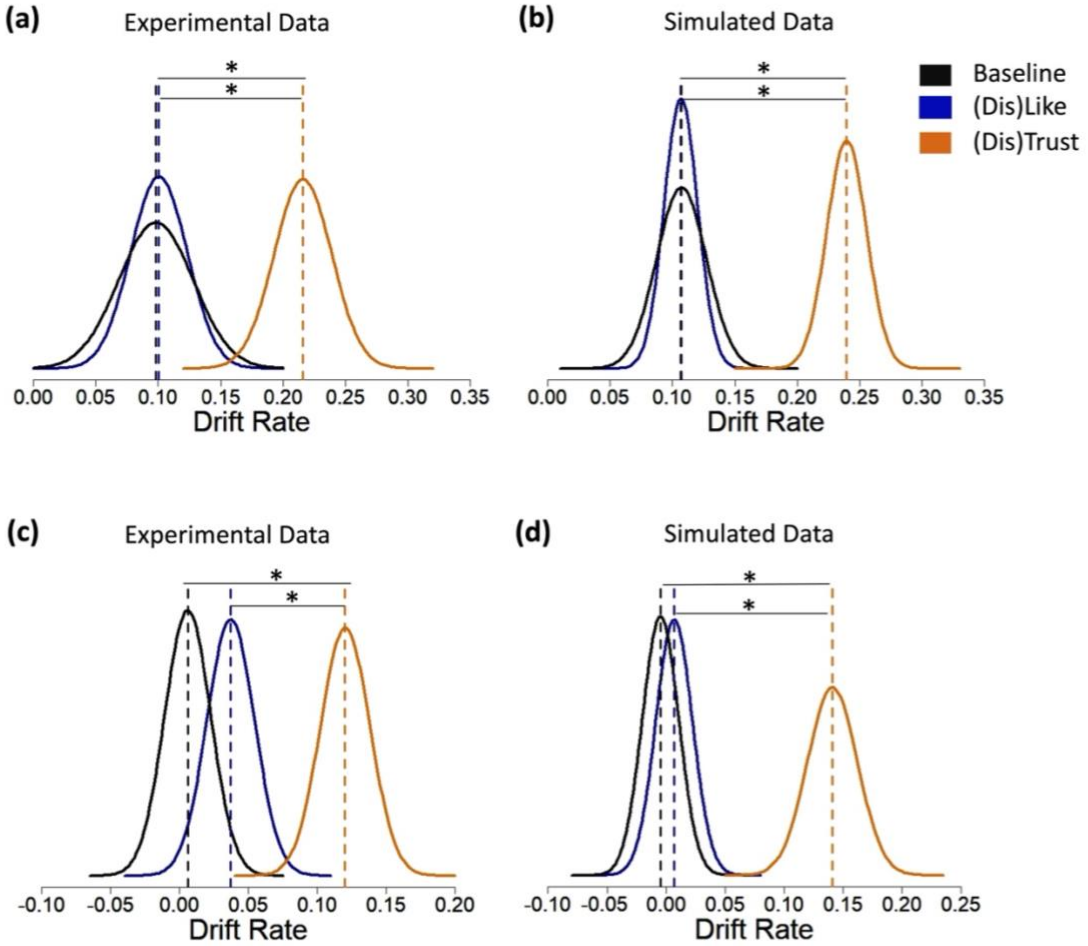

Figure 4.6. '(Dis)Trust' feedback increases the drift rate. Displayed are the posterior distributions of parameter estimates for the Baseline environment, the '(Dis)Like' environment and the '(Dis)Trust' environment. Dashed vertical lines indicate respective group means. **(a)** In both Experiment 2 and **(c)** Experiment 3 95% CI comparison revealed that participants had a larger drift rate in the '(Dis)Trust' environments than in the other environments. No significant difference was observed between the latter two environments. Recovered model parameter estimates reproduced experimental results for both **(b)** Experiment 2 and **(d)** Experiment 3. *indicates significant difference between parameters (i.e., CI do not overlap).

Estimate	**Baseline** [95% CI]	**'(Dis)Like'** [95% CI]	**'(Dis)Trust'** [95% CI]
Distance between Decision Thresholds (α)	2.153 [2.09, 2.214]	2.373 [2.281, 2.466]	2.403 [2.280, 2.529]

Non-Decision Time (t0)	7.025 [6.898, 7.154]	6.936 [6.802, 7.071]	6.681 [6.425, 6.94]
Starting Point (z)	0.497 [0.486, 0.508]	0.491 [0.483, 0.50]	0.48 [0.471, 0.49]
Drift Rate (v)	0.098 [0.039, 0.158]	0.10 [0.056, 0.145]	0.216 [0.17, 0.262]

Table 4.1. Group estimates for DDM in Experiment 2.

Estimate	**Baseline** [95% CI]	**'(Dis)Like'** [95% CI]	**'(Dis)Trust'** [95% CI]
Distance between Decision Thresholds (α)	2.238 [2.153, 2.328]	2.207 [2.132, 2.286]	2.209 [2.134, 2.286]
Non-Decision Time (t0)	6.9 [6.762, 7.04]	7.051 [6.918, 7.186]	7.076 [6.944, 7.208]
Starting Point (z)	0.5 [0.49, 0.51]	0.5 [0.49, 0.511]	0.489 [0.476, 0.5]
Drift Rate (v)	0.006 [-0.027, 0.037]	0.037 [0.002, 0.069]	0.12 [0.086, 0.155]

Table 4.2. Group estimates for DDM in Experiment 3.

While in Experiment 2 the distance between decision thresholds in the Baseline environment was lower than in the other two environments, and non-decision time (t0) higher than in the '(Dis)Trust' environment, these differences are not replicated in Experiment 3**.** More importantly, neither distance between decision thresholds nor non-decision time differed between the '(Dis)trust' and '(Dis)like' environments (see **Table 4.1 & 4.2 and Appendix 7.3 Supplementary Tables 4.11 & 4.12 for HDI Comparisons**).

Model parameters could be successfully recovered with data simulated using group-level parameters from Experiment 2 and Experiment 3 separately (**see Methods for details**, **see Figure 4.6 b, d, Appendix 7.3 Supplementary Tables 4.13 & 4.14)**. This was done by fitting the model to simulated data, in the same way as for the experimental data. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. The same pattern of results was reproduced with the simulated data as with real participants' data (**Figure 4.7**). For each Experiment we ran two separate one-way ANOVAs to assess the effect of type of feedback on discernment: one for the real data and one for the simulated data. We remind the reader that we entered responses into our DDM as either ‘veracity-promoting’ (true post shared or false post skipped) or ‘veracity-obstructing’ (false post shared or true post skipped). Thus, discernment here is calculated as:

$$Discernment = Prop_{veracity-promoting\ responses} - Prop_{veracity-obstructing\ responses}$$

Which is equal to:

$$Discernment = Prop_{reposts\ true\ posts+reposts\ false\ posts} - Prop_{reposts\ false\ posts+skips\ true\ posts}$$

As expected, we observed an effect of type of feedback in both the simulated (Experiment 2: F(1,285) = 3.795, p = 0.024, η2 = 0.026; Experiment 3: F(1,388 = 7.843, p = 0.001, η2 = 0.039, **Figure 4.7 b, d**), and the experimental data (Experiment 2: F(1,287) = 7.049, p = 0.001, η2 = 0.047; Experiment 3: F(1,388) = 11.166, p < 0.001, η2 = 0.054). That is, discernment was higher in '(Dis)Trust’ environments relative to ‘(Dis)Like’ environments or no feedback environments **(Figure 4.7 a, c, see Appendix 7.3 Supplementary Tables 4.15 & 4.16 for pairwise comparisons** and **Supplementary Table 4.17 for correlations between real and recovered individual-level parameters**).

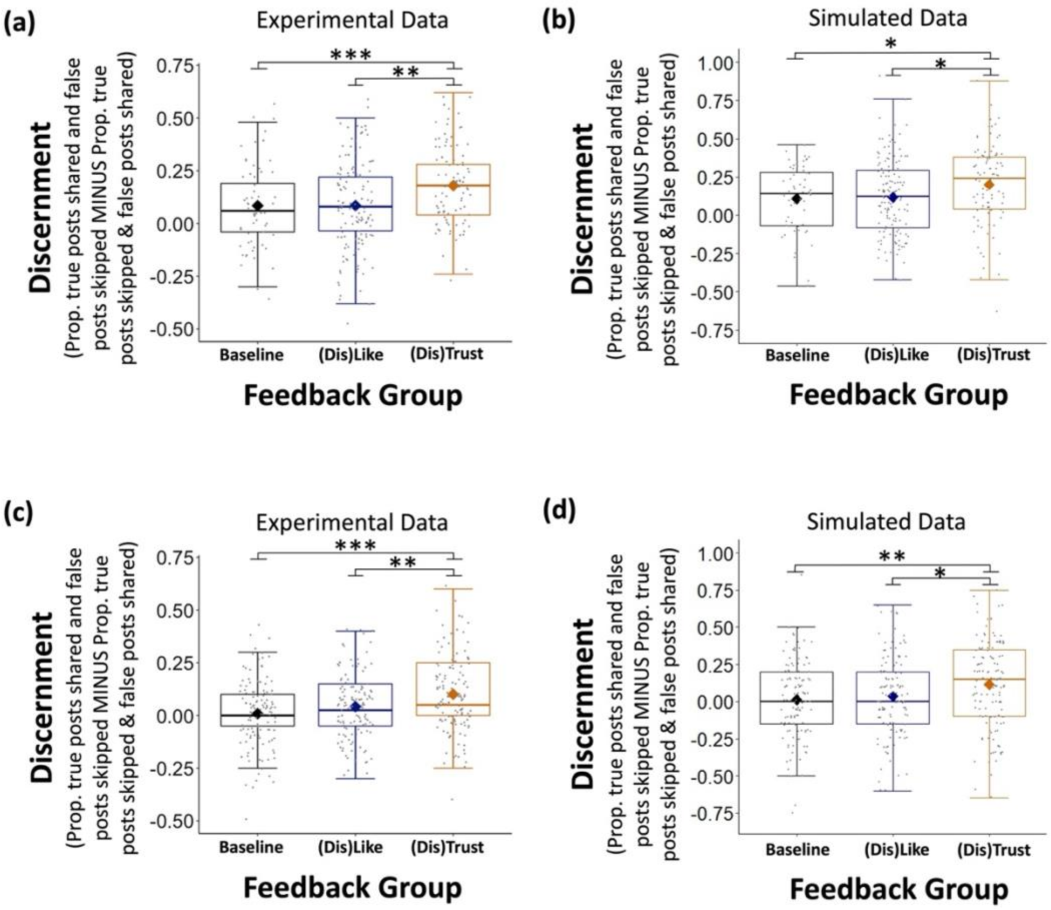

Figure 4.7. Simulated Data reproduced experimental findings. In both **(a)** Experiment 2 and **(c)** Experiment 3 participants who received '(Dis)Trust' feedback were more discerning than participants in the '(Dis)Like' and Baseline environments. Simulated data reproduced these findings **(b&d)**. Y axis shows discernment, i.e., proportion of true posts shared and false posts skipped minus the proportion of true posts skipped and false posts shared. X axis shows feedback group. Data are plotted as boxplots for each reaction, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean discernment per reaction. Individuals' mean discernment data are shown separately as grey dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001, **p < 0.01, *p < 0.05.

4.5 Discussion

Here, we created a novel incentive structure that significantly reduced the spread of misinformation and provide insights into the cognitive mechanisms that make it work. This structure can be adopted by social media platforms at no cost. The key was to offer reaction buttons (social 'carrots' and 'sticks') that participants were likely to use in a way that discerned between true and false information. Users who found themselves in such an environment, shared more true than false posts in order to receive more ‘carrots’ and less ‘sticks’.

In particular, we offered ‘trust’ and ‘distrust’ reaction buttons, which in contrast to ‘likes’ and ‘dislikes’, are by definition associated with veracity. For example, a person may dislike a post about Joe Biden winning the election, however this does not necessarily mean that they think it is untrue. Indeed, in our study participants used ‘distrust’ and ‘trust’ reaction buttons in a more discerning manner than ‘dislike’ and ‘like’ reaction buttons. This created an environment in which the number of social rewards (‘carrots’) and punishments (‘sticks’) were strongly associated with the veracity of the information shared. Participants who were submitted to this new environment were more discerning in their sharing behaviour compared to those in traditional environments who saw either no feedback or ‘dislike’ and/or ‘like’ feedback. The result was a reduction in sharing of misinformation without a reduction in overall engagement. All the effects were replicated, and effect sizes of misinformation reduction were large to medium.

Using computational modelling we were able to pin-point the changes to participants’ decision-making process. In particular, DDM revealed that participants in the new environment assigned more weight to evidence consistent with discerning than non-discerning behaviour relative to traditional environments. In other words, the possibility of receiving rewards that are consistent with accuracy led to an increase in the weight participants assigned to accuracy-consistent evidence when making a decision. ‘Evidence’ likely includes external information that can influence the decision to share a post (such as the text and photo associated with the post) as well as internal information (e.g., retrieval of associated knowledge and memories).

Our results held when the potential feedback was only negative (‘distrust’), only positive (‘trust’), or both (‘trust’ and ‘distrust’). While negative reaction buttons were used in a more discerning manner and more frequently than positive reaction buttons, we did not find evidence for a differential strength of positively or negatively valenced feedback on discernment of sharing behaviour itself.

The findings also held across a wide range of different topics (e.g., politics, health, science, etc.) and a diverse sample of participants, suggesting that the intervention is not limited to a set group of topics or users, but instead relies more broadly on the underlying mechanism of associating veracity and social rewards. Indeed, we speculate that these findings would hold for different ‘carrots’ and ‘sticks’ (beyond ‘trust’ and ‘distrust’), as long as people use these ‘carrots’ and ‘sticks’ to reward true information and punish false information. However, we speculate that the incentives were especially powerful due to being provided by fellow users and easily quantifiable (just as existing buttons including ‘like’ and ‘heart’). This may contrast with incentives which are either provided by the platform itself and/or not clearly quantified such as verification marks (Edgerly & Vraga, 2019) or flagging false news (Brashier et al., 2021; Chan et al., 2017; Grady et al., 2021; Lees et al., 2022). Interestingly, a trust button has also been shown to increase sharing of private information (Bălău & Utz, 2016).

Finally, we observed that feedback not only promotes discernment in sharing behaviour but may also increase the accuracy of beliefs. Though, while we see an increase in accuracy of beliefs in three of the four experiments, we did not observe this effect in Experiment 3. Thus, the new incentive structure reduces the spread of misinformation and may help in correcting false beliefs. It does so without drastically diverging from the existing incentive structure of social media networks by relying on user engagement. Thus, this intervention may be a powerful addition to existing intervention such as educating users on how to detect misinformation (Lewandowsky & van der Linden, 2021; Maertens et al., 2021; Pilditch et al., 2022; Roozenbeek & van der Linden, 2019; Traberg et al., 2022) or prompting users to think about accuracy before they engage in the platform (Capraro & Celadin, 2022; Fazio, 2020; Pennycook & Rand, 2022a). Over time, these incentives may help users build better habits online (Anderson & Wood, 2021; Ceylan et al., 2023).

As real-world platforms are in the hands of private entities, studying changes to existing platforms requires testing simulated platforms. The advantage of this approach is the ability to carefully isolate the effects of different factors. However, real-world networks are more complex and involve additional features which may interact with the tested factors. Our hope is that the science described here will eventually impact how privately owned platforms are designed, which will reveal whether the basic mechanisms reported here hold in more complex scenarios.

This study lays the groundwork for integration of the new incentive structure into existing (and future) social media platforms to further test the external validity of the findings. Rather than removing existing forms of engagement, we suggest an addition that complements the existing system and could be adopted by social media platforms at no cost. The new structure could subsequentially help reduce violence, vaccine hesitancy and political polarization, without reducing user engagement.

Chapter 5: General Discussion

5.1 Synthesis

The research detailed in this thesis sheds new light on whether and how, prominent incentives to gather and share information can be altered to reduce biased evidence accumulation and make sharing decisions more discerning.

In the first part of this thesis, I focus on evidence accumulation as the pathway of decision-making (Platt & Glimcher, 1999). Prior research illustrates that individuals selectively accumulate evidence to form positive beliefs (Gesiarz et al., 2019; Leong et al., 2019) from which they derive internal rewards, such as positive emotions, and increased sense of self-efficacy (Bromberg-Martin & Sharot, 2020; Loewenstein, 2006; Sharot et al., 2023). Often, they prioritize internal incentives over external incentives, such as financial gains. This results in a bias towards desirable, internally rewarding conclusions (Gesiarz et al., 2019; Leong et al., 2019). Such positively biased beliefs can have detrimental consequences, such as lack of preparation for natural disasters (Paton, 2003).

In Chapter 2 and 3, I investigate if and how incentives can be altered to make evidence accumulation less biased. I hypothesize that evidence accumulation becomes less biased towards desirable conclusions when false beliefs resulting in false conclusions are costly. I test this in two ways: In Chapter 2, I examine how perceived threat impacts evidence accumulation. I find that under threat the relative rate at which negative evidence is accumulated increases and the bias towards desirable conclusions disappears. This may be adaptive, as it can lead to increased precaution in environments in which the risk of adverse consequences is high. Building on this work, in Chapter 3, I then examine whether increasing the external incentives to reach accurate conclusions can also alter the influence of internal incentives on evidence accumulation. Results show that while accuracy incentives led participants to take more time to reach a conclusion, they did not impact participants' bias. I provide a mechanistic explanation for why this might be. In particular, DDM reveals that accuracy incentives and the desirability bias act on orthogonal aspects of the accumulation process. While accuracy incentives led to an increase in the distance between decision thresholds, the bias was associated with greater weight on desirable relative to undesirable evidence (drift rate bias). These results suggest that participants may not be aware of their own bias. I suggest that when accuracy incentives are coupled with direct feedback, decision-making becomes more discerning.

One decision that is informed by evidence accumulation is the decision to share information (Globig et al., 2023; Huang et al., 2015; Lin et al., 2023). In recent years, the rising ease of sharing, brought about by the advent of the internet, has also facilitated the dissemination of misinformation (Kreps, 2020). This has been attributed to the existing incentive structure of social media platforms (Brady et al., 2021; Lindström et al., 2021; Scissors et al., 2016), in which social rewards ('likes') and punishments ('dislikes') are dissociated from the veracity of the information shared. In the Chapter 4, I therefore test the hypothesis that the role of this incentive structure in promoting misinformation spread can be curtailed by making social rewards and punishment contingent on the veracity of the information shared. Results show that when social incentives are contingent on the veracity of information, discernment in sharing behaviour increases which in turn reduces the spread of misinformation.

Taken together, this thesis provides insights into how incentive structures can be altered to improve the quality of evidence accumulation and sharing behaviour. This discussion will summarize the key findings of the studies in each chapter, and delve into their implications, limitations, and potential future pathways for this research.

5.2 Under threat weaker evidence is required to reach undesirable conclusions

5.2.1 Summary

Many important decisions are made in threatening environments. Such settings, often characterized by heightened stress and anxiety, are known to have adverse effects on both learning and decision-making (FeldmanHall et al., 2015; Porcelli & Delgado, 2009, 2017; Raio et al., 2013; Starcke & Brand, 2012). More recent work, however, highlights that stress can also have an adaptive function (Akinola & Mendes, 2012; Garrett et al., 2018; Graybeal et al., 2011).

In Chapter 2, I demonstrate that stress induced by perceived threat also alters the process by which evidence is accumulated in a way that may be adaptive. Ninety-one participants completed a sequential sampling task in which they were incentivized to accurately judge whether they were in a desirable state, which was associated with greater rewards than losses, or an undesirable state, which was associated with greater losses than rewards (Gesiarz et al., 2019). Participants were assigned to one of two groups: a 'threat group' and a 'control group'. Prior to the task, participants in the threat group experienced a social-threat manipulation. As expected, we found that participants in the control group were biased towards desirable conclusions. They weighed desirable evidence more than undesirable evidence (replicating previous findings from Gesiarz et al., 2019). Under threat this bias disappears. Relative to the control group, participants in the threat group required weaker evidence to reach an undesirable conclusion. DDM revealed that this was due to an increase in the relative rate at which negative evidence is accumulated. This is line with previous findings that stress increases attention to negative stimuli, resulting in fewer decision errors (Akinola & Mendes, 2012), and bolsters the integration of negative information, thereby moderating the bias for positive information in belief updating (Garrett et al., 2018). Such increased attention may also, in part, explain the observed effect of stress on the rate of evidence accumulation. The results reported in Chapter 2 extend previous research by shedding light on how perceived threat alters the bias towards desirable conclusions in evidence accumulation and show that perceived threat reduces the strength of evidence required to reach an undesirable conclusion. As the source of the threat in this study was unrelated to the behavioural task this threat-induced change in evidence accumulation points to a global effect of stress. Taken together, these results lend support to the hypothesis that when participants are exposed to an environment in which the costs of false beliefs are high, such as a threatening environment, they are less biased towards internally rewarding conclusions. This study thus suggests that individuals adaptively prioritize internal and external incentives depending on the environment.

5.2.2 Limitations and Future Directions

Chapter 2 sheds light on how perceived threat alters evidence accumulation and provides support for the idea that when the costs of false beliefs are high, the motivational bias towards internally rewarding conclusions is mitigated. This is in line with prior research, showing that stress improves the ability to flexibly adjust previously learned behaviours to fit new task requirements (Graybeal et al., 2011) and triggers defence mechanisms (Baas et al., 2006; Cornwell et al., 2007), especially in the face of aversive stimuli (Blanchard et al., 2011; Davis et al., 2010; Grillon, 2008; Robinson et al., 2012). However, it is important to bear in mind some limitations of this study and future directions for this work.

Anecdotal evidence illustrates that psychophysiological stress-responses often provide a global rather than specific signal of threat. For example, stress elicited by personal conflict may impact professional performance (Piotrkowski, 1979). This fits with prior research reporting a general effect of stress on behaviour and neural responses, that goes beyond the source of the threat itself, in both humans (Cavanagh, Frank, et al., 2011; Lenow et al., 2017; Otto & Daw, 2019; Robinson, Overstreet, et al., 2013; Youssef et al., 2012) and non-human animals (Harding et al., 2004; Rygula et al., 2013). As such, in this study, the source of the threat (anticipating a stressful event) was intentionally disconnected from the behavioural task (the ‘Factory Task’). As expected, the results point to a global effect of perceived threat on evidence accumulation. This may be adaptive, as it enhances the likelihood of taking precautions against potential aversive outcomes, which tend to be more extreme and/or common in threatening environments.

However, there is some evidence to suggest that the effect of global stress may differ from the effect of task-specific stress. In some instances, the latter has even been found to induce biased cognition (for a review see Yu, 2016). For example, a study Engelmann et al., (2019) observed that participants who engaged in a visual pattern recognition task, in which some patterns were associated with receiving an electric shock, were worse at recognizing these patterns, relative to patterns that were not associated with a shock. This suggests that participants engaged in 'wishful thinking' to deceive themselves about the looming electric shock. This stands in contrast to the study described in Chapter 2, in which the threat-induction was decoupled from the behavioural task and participants were motivated to hold one belief over another using financial incentives. In Engelmann et al., (2019) the threat itself motivated participants to form a desirable conclusion. This in line with suggestions, that people at risk of terminal illnesses avoid negative information to remain optimistic and protect their emotional well-being (Ganguly & Tasoff, 2017; Lerman et al., 1998; Oster et al., 2013). As participants have no (perceived) control over the incoming threat, this response is likely adaptive. This may also explain reports of a reduction in attentional bias towards aversive stimuli under threat (Jiang et al., 2017). In the study outlined in Chapter 2, the threat response is unspecific, and thus participants are alerted to potential dangers in their environment. I thus speculate, that when accumulating evidence under threat, prioritizing external over internal incentives is an adaptive response to allow participants to adopt defensive mechanisms (see also Baas et al., 2006; Cornwell et al., 2007); especially in response to threatening stimuli (Blanchard et al., 2011; Davis et al., 2010; Grillon, 2008). Taken together, these studies suggest, that stress has an adaptive function designed to restore homeostasis (de Kloet et al., 2008) and therefore its' precise effect on decision-making depends on the specific context (Starcke & Brand, 2012). Future studies should further investigate this hypothesis and could explore to what extent the findings from Chapter 2 hold in different contexts, using different threat induction methods, such as cold-pressor tasks (Lenow et al., 2017; Otto et al., 2013), and threat of electric shock (Robinson, Overstreet, et al., 2013), or different task designs. Moreover it would be interesting to explore how global threat as used in Chapter 2 interacts with a task-induced motivational manipulation which uses threat as adopted by Engelmann et al., (2019).

In Chapter 2, I establish a causal relationship between perceived threat and evidence accumulation in healthy individuals. An intriguing avenue for future research is to examine how these findings might apply to chronic stress and individuals with affective disorders, who often report heightened anxiety. It has been suggested that chronic stress has adverse effects on cognitive and neural functions (for a review see Sousa & Almeida, 2012) and may have a more pronounced effect on decision-making compared to acute stress (Hales et al., 2016). In individuals, who are hypersensitive to stress, enhanced accumulation of undesirable stimuli could therefore be maladaptive and result in overly pessimistic predictions, further aggravating distress. Indeed, Aylward et al. (2019) observed a diminished drift rate toward desirable conclusions in individuals with anxiety and mood disorders, when studying the effect of stress in a two-alternative-forced-choice task, in which participants had to discriminate between two types of auditory tones (high reward tone vs low reward tones). Future research should explore this effect in a sequential sampling task, like the ‘Factory Task’ (Gesiarz et al., 2019) used here. This would allow us to examine whether the mechanism by which stress impacts evidence accumulation identified in Chapter 2 could serve as a potential target mechanism for therapeutic interventions to restore healthy cognitive processing.

A further potential limitation of the study is the absence of physiological measures, such as salivary cortisol levels, to confirm the efficacy of the elicited stress response. In this experiment, anxiety was induced using a variation of the Trier Social Stress Test (Birkett, 2011; Marteau & Bekker, 1992; Spielberger et al., 1970) and assessed the success of the manipulation using a well-established state anxiety questionnaire (Allen et al., 2014). While this procedure was validated using salivary cortisol levels and skin conductance in a previous study from our lab (Garrett et al., 2018), recording cortisol as a biomarker for stress, may have provided a more objective assessment of the stress response, complementing the study's findings based on behavioural data (for a detailed discussion on this topic see Allen et al., 2014). Furthermore, the effect of perceived threat on decision-making varies as a function of sensitivity to cortisol response (Kudielka et al., 2009; Van den Bos et al., 2009). Thus, the precise effect of perceived threat on evidence accumulation may also be contingent on individual differences in sensitivities to threat. Future studies should record salivary cortisol changes to provide further insights into the impact of these and other differences on the desirability bias in evidence accumulation. Neuroimaging could further complement these findings. Additionally, it is plausible that a moderate level of stress may enhance behaviour adaptively, while minimal or excessive stress, as often seen in clinical disorders, might impair performance, thereby suggesting the presence of an inverted U-shaped relationship. This hypothesis aligns with existing literature, which reports an inverted U-shaped correlation between stress and memory performance (for a review see Lupien et al., 2007). Future research should aim to empirically test this hypothesis using cortisol levels to further elucidate the nuanced effects of stress on biased cognition.

In conclusion, despite its limitations, the present study supports the idea that evidence accumulation is a dynamic process in which individuals adaptively prioritize internal and external incentives depending on their costs and benefits in a given environment. This discussion has also highlighted potential directions for future research, which would contribute to a more comprehensive understanding of this process.

5.3 Futile Rewards: Why Accuracy Incentives Fail to Reduce Biased Evidence Accumulation

5.3.1 Summary

When accumulating evidence to make a decision, individuals are biased by the desire to form positive beliefs, from which they derive internal rewards, such as positive emotions. They assign greater weight to evidence that aligns with their desired belief. As a result, they are more likely to reach a conclusion that confirms this belief, even if it is incorrect (Gesiarz et al., 2019; Globig et al., 2021; Leong et al., 2019). Such false beliefs can lead to suboptimal decision-making (for a review see Karayanni & Nelken, 2022) which can have detrimental consequences, such as falsely optimistic investment decisions preceding the financial crisis (Shefrin, 2015). One obvious solution is to financially incentivize people to form more accurate beliefs in situations where biases and heuristics are common. This idea is predicated on the understanding that biases and heuristics can be moderated through deliberate and effortful thinking (Bonner & Sprinkle, 2002; Botvinick & Braver, 2015; Smith & Walker, 1993). Yet, despite the intuitive appeal of this approach, the empirical evidence is mixed (Engelmann et al., 2019; Prior et al., 2015; Rathje et al., 2023; Zhang & Rand, 2023).

In Chapter 3, I demonstrate that accuracy incentives fail to counteract biased evidence accumulation and provide a mechanistic explanation for why this might be. Over three experiments, participants (N = 235) completed a perceptual evidence accumulation task in which they had to judge whether they were in a desirable state, which was associated with greater rewards, or an undesirable state, associated with no reward. In some trials they were also financially incentivized for correct responses. Crucially participants had no control over which type of trial they were in and had to be as accurate as possible to maximize their rewards. The results show that while participants were more cautious about their conclusions when incentivized for accuracy, they remained biased towards desirable conclusions. This was true even when the reward for correct responses was increased five-fold.

These results fit with previous suggestions that even substantial accuracy incentives fail to alleviate certain biases, including the partisan bias in advice taking (Zhang & Rand), and anxiety-induced wishful thinking (Engelmann et al., 2019). The current findings go beyond these previous demonstrations and provide a mechanistic explanation for why accuracy incentives may fail to reduce biased evidence accumulation. DDM revealed that accuracy incentives and the bias towards internally rewarding conclusions alter orthogonal elements of the accumulation process. While the bias selectively increases the rate of evidence accumulation towards desirable conclusions, accuracy incentives increase the distance between decision thresholds, thereby increasing the amount of evidence participants required to make a decision, but not changing the way in which this evidence is accumulated.

Taken together, these findings lend support to previous suggestions that internal incentives alter pre-conscious processing of information and the resultant motivational biases are not merely the product of overt preferences but instead beyond individuals' awareness (Chen & Krajbich, 2018; Desai & Krajbich, 2022; Krajbich, 2022; but see Sánchez-Fuenzalida et al., 2023). This study lends support to the hypothesis that the inherent subjective value of internal rewards exceeds the subjective value of external rewards in safe environments. The findings are particularly relevant for policymakers and industry leaders as they may explain why financial bonuses may be ineffective in improving decision-making.

5.3.2 Limitations and Future Directions

Chapter 3 provides a mechanistic explanation for why accuracy incentives may fail to mitigate the influence of the motivational bias on evidence accumulation. These results offer valuable insights for both private and public sectors and may help explain why financial bonuses do not always result in more rational decision-making. Nevertheless, it is essential to note the limitations inherent in this study and to outline future research directions.

The findings from Chapter 3 seemingly stand in contrast to those of Chapter 2. In theory, both perceived threat and accuracy incentives can increase the external costs of holding false beliefs. Participants assign greater weight to undesirable than desirable evidence under threat, but not when incentivized for accuracy. This could be because stress has a global effect on evidence accumulation. The stress response thus alters both conscious and unconscious processes. In contrast monetary incentives seemingly operate via more high-level conscious processes.

Our results indicate that when incentivized for accuracy, participants are more cautious about their responses. Yet, despite their increased effort, they remain biased towards desirable conclusions. This suggests that participants' biased perception is sincere, and thus cannot be easily modulated through financial incentives. This fits with prior work on motivated perception positing that the influence of internal incentives extends to pre-conscious processing of information (Balcetis & Dunning, 2006,). By contrast, perceived threat elicits a global stress response, which can be both implicit and explicit (Verkuil et al., 2009) and thus also alters the implicit influence of internal incentives on evidence accumulation. I speculate, that the reason accuracy incentives fail to reduce biased evidence accumulation, is that participants are not aware of their mistakes. Their biased processing of information is genuine and not a result of deliberation. This is in line with suggestions that preferential belief updating for good news compared to bad news occurs even when there is a time limit or participants have reduced cognitive load (Kappes & Sharot, 2019). It is possible that allowing participants to learn about the inaccuracy of their beliefs, for instance by providing feedback would help mitigate biased processing (but see Engelmann et al., 2019). Indeed, in Chapter 4, I found that participants who received feedback about the accuracy of the content they shared in simulated social media environments, were better at discerning between true and false posts, than those who did not receive accuracy-related feedback. Similar results were also reported by Zhang and Rand (2023) who observed that while accuracy incentives failed to mitigate partisan bias in advice-taking, allowing participants to learn about their bias through feedback, successfully reduced the partisan bias and improved accuracy. I thus hypothesize that if participants learn that their beliefs are false, they adjust how they weigh the evidence they observe, reducing the bias towards desirable conclusions. Future studies, for instance using neuroimaging and eyetracking, should test this hypothesis and investigate whether the preferential sampling of desirable evidence is indeed sincere and not deliberate (for a discussion see Krajbich, 2022).

In addition to this, Chapter 2 and Chapter 3 are not entirely comparable, as they utilize different paradigms. While Chapter 2 utilizes a sequential sampling task, Chapter 3 uses a perceptual accumulation task. Although the bias in the drift rate towards desirable conclusions is consistently observed across both tasks, a reliable comparison between the two ways of increasing external incentives requires identical tasks designs. Future research should therefore test if accuracy incentives remain ineffective when a sequential sampling task, like the ‘Factory Task’, is used.

Finally, it is important to acknowledge, that although the effect of accuracy incentives on response times was observed consistently across experiments in Chapter 3, the effect sizes were small. Interestingly, the effect of accuracy incentives on response times increased when the reward was increased five-fold. While it is possible that further increasing the magnitude of accuracy incentives could augment caution in such a way that participants become less biased, I speculate that while caution would increase, a reduction of the bias is unlikely given that the two alter distinct elements of the accumulation process. Indeed, prior research using substantially larger financial accuracy incentives also failed to show a reduction of motivational biases (Enke et al., 2023; Zhang & Rand, 2023). Nevertheless, future studies should examine whether changing the type of accuracy incentive, e.g., using a non-monetary accuracy incentive, or using punishment instead of reward would alter these results.

In conclusion, Chapter 3 sheds light on why accuracy incentives might fail to debias evidence accumulation. This discussion has highlighted important limitations and potential directions for future research, which would contribute to a more comprehensive understanding of the influence of internal incentives on evidence accumulation and how to mitigate it.

5.4 Changing the Incentive Structure of Social Media Platforms to halt the Spread of Misinformation

5.4.1 Summary

The wide-spread success of social media platforms has been attributed to the human need for social rewards (Nadkarni & Hofmann, 2012). People share information in order to receive rewards, in the form of engagement, and in turn this engagement reinforces sharing behaviour (Lindström et al., 2021; Rosenthal-von der Pütten et al., 2019; Scissors et al., 2016). In Chapter 4, I demonstrate that the existing incentive structure of social media platforms, in which social rewards (e.g., ‘likes’) and punishments (e.g., ‘dislikes’) are dissociated from the veracity of the information shared, contributes to the spread of misinformation online. I show that slightly altering this incentive structure, such that social rewards and punishments are contingent on the veracity of information, improves discernment in sharing behaviour. Recent studies corroborate this result (Alizadeh et al., 2023; Butler et al., 2023; Kapoor et al., 2023; Pretus et al., 2023; Rathje et al., 2023).

Over six experiments, 951 participants engaged in simulated social media platforms where they encountered both true and false information. In some platforms ‘trust’ and ‘distrust’ reaction buttons, which are, by definition, related to veracity, were added to the existing ‘like’ and ‘dislike’ reaction buttons. We found that participants used the buttons in a more discerning manner than the existing engagement options, thereby creating an environment in which the number of social rewards and punishments were strongly associated with the veracity of the information shared. Other participants who were then exposed to this environment subsequently shared more true relative to false posts than those in traditional environments. This improved the overall quality of the information shared thereby reducing the spread of misinformation without reducing overall engagement. DDM revealed that the underlying mechanism of this effect is associated with an increase in the weight participants assign to evidence consistent with discerning behaviour.

Taken together, these results offer evidence for an intervention that capitalizes on the reward-based engagement structures of social media platforms. By creating an environment in which engagement is contingent on the veracity of the information shared, platforms could create healthier information ecosystem without reducing overall engagement. This study therefore lays the groundwork for an intervention which could easily be incorporated into a variety of different existing social media platforms to reduce the spread of misinformation and thus mitigate its devastating consequences for society.

5.4.2 Limitations and Future Directions

The results of this study provide compelling evidence for an intervention that leverages the existing incentive structure of social media platforms and could easily be adopted in the real world. However, there are several limitations which bear consideration.

First, it is unclear to what extent the efficacy of the intervention generalizes to real-world contexts. The study utilizes a highly controlled simulated social media environment. Although such approximations of real-world situations are vital to establish proof-of-concept (Charness & Fehr, 2015; Falk & Heckman, 2009), they also bring about several limitations which constrain the ecological validity of the study. For instance, participants in this study were either given the opportunity to react to posts (Experiment 1 & 4), or to share them (Experiment 2-3 & 5-6), never both. Moreover, in the latter experiments, participants were assigned to one of three between-subject conditions. They either saw no feedback, ‘like’ and/or ‘dislike’ feedback, or ‘trust’ and/or ‘distrust’ feedback. By contrast, in real social media platforms users can react and share at the same time and are also able to see a range of social feedback simultaneously. Furthermore, the design of these experiments was such, that participants first decided whether to repost a piece of content or not, and only then would they receive feedback. This was done to ensure that learning what others thought about a specific piece of content could not directly impact whether participants would repost that *specific* piece. Instead, the idea was that knowing that others would be able to 'trust' or 'distrust' the post would motivate participants to consider the veracity of the post when making their decision. However, when users decide to repost something in real social media platforms, they can often already see how others have reacted to the original post. Thus, future studies should examine whether the observed effects hold when users can react and share at the same time and are able to see how others have reacted to the original post.

The prevalence of misinformation online has been attributed to large rumour cascades which originate from so-called 'super-spreaders' with high-follower counts (Mosleh & Rand, 2022; Vosoughi et al., 2018). In line with this, 0.01% of Twitter (now X) users generated 80% of the misinformation surrounding the US election (Grinberg et al., 2019). As such, the reposting of other-generated content plays a pivotal role in the spread of misinformation. In Chapter 4, I therefore focused on sharing other-generated, rather than sharing self-generated content. This not only mimics prevalent real-world behaviour, but also ensures that all participants in the study observe the same content thereby increasing the reliability of the study. Nevertheless, in real platforms users are able to share their own content, as well as repost that of others. While I predict that anticipating others' '(Dis)Trust' feedback influences reposting and sharing behaviour equally, future studies should directly test this hypothesis. Furthermore, participants in this study only observed news-style content which, was either true or false. Future studies should investigate how users utilize the '(Dis)Trust' buttons on other types of content, such as personal interest updates or opinion pieces. As '(Dis)trust' feedback is intended to complement existing '(Dis)Like' feedback, I predict that participants will primarily use the latter when interacting with social content, and the former when interacting with verifiable content.

In addition to utilizing feedback to make sharing decisions, social media users may also (rightly or wrongly) use a range of other signals to infer the reliability of the content they encounter online, such as the source of the information (Epstein et al., 2019). To reduce confounds, this study did not provide information about the identity of the original poster or the source of the content. Future studies should investigate how these different signals interact. Furthermore, it is important to bear in mind that the sample in this study was politically balanced. As such, the feedback participants received was also politically balanced. Future studies should investigate how this intervention behaves in organic polarized networks. For instance, in this study Democrats were more discerning in their use of '(Dis)Trust' buttons than Republicans, as such it is possible that feedback obtained from a primarily Republican sample would be less discerning. In line with this, due to ethical considerations users remained anonymous in this study. However, it is possible that the use of the '(Dis)Trust' buttons as well as the sensitivity to such feedback may vary depending on the social proximity of the parties involved. Indeed it has been suggested that social media interactions differ depending on the proximity of the users (Stopczynski et al., 2018). Another limitation of this study is that it only investigated behaviour at one timepoint. It did not assess the longevity of the intervention. Future research should investigate whether the results hold long-term or whether users habituate to '(Dis)Trust' feedback over time and become less responsive to it.

To further aid in the adoption of novel interventions into real-world platforms, it is essential not only to consider whether they achieve their desired effect in reducing the spread of misinformation, but also to investigate how they influence holistic user behaviour. For instance, some platforms use algorithms which do not primarily rely on user interactions, but instead utilize more passive methods of engagement (Geers et al., 2024), such as dwell-time, that is the time spent looking at a piece of content (Epstein et al., 2022). As such, another limitation of this study is that it solely focusses on discernment in sharing behaviour. Future research should therefore also take into account other variables, such measures of attention, that can be utilized by algorithms to determine the dissemination of content. Relevant insights could for instance be obtained using eye tracking to assess attention.

To systematically address these limitations and examine the holistic impact of this (and other) intervention(s) on user behaviour, future studies should test the efficacy of this intervention in real-world settings. However, as platforms are seemingly wary of directly implementing proposed interventions, possibly for fear of adverse effects, novel solutions are necessary. For instance, some researchers are trying to develop immersive social media simulations, that are modelled on existing platforms, such as Twitter (now X) or Facebook and have varying degrees of experimental control (for an overview see Jagayat & Choma, 2023). While the existing prototypes of these platforms are either not fully interactive or not scalable to larger audiences, they hold great promise for both academia and industry alike. By bridging the gap between highly controlled experimental settings and the real world, this type of hyper-realistic social media simulation could not only provide insights into how this intervention impacts user behaviour 'in the wild' but it would also aid in the development of more optimal solutions to combatting the spread of misinformation online.

Another limitation of the study is that it does not contain a baseline measure of participants' belief accuracy. This was due to a variety of reasons. First, it has been shown that directly asking users to indicate whether they think a post is true or false alters sharing behaviour (for a review see Pennycook & Rand, 2022b). To rule out that such accuracy prompts could confound the influence of social feedback, belief accuracy was only assessed once participants had already made their sharing decisions. Second, it has been shown that the repeated exposure of information alters its perceived accuracy and thus increases the likelihood of sharing. To avoid such interference, no initial (baseline) beliefs were recorded. Thus, it is possible that existing between-group differences prior to the intervention contributed to the observed effect on belief accuracy. Furthermore, it has been shown that participants asymmetrically update their beliefs when learning about others' opinions (Kappes et al., 2020). Finding out others agree with them significantly increases their confidence. By contrast, finding out that others disagree only marginally diminishes their confidence. It then follows that when participants observe social feedback they disagree with, they do not update their beliefs, as much as for positive feedback (Kappes et al., 2020). Future studies should strive to tease apart this effect of valence on belief updating in response to online feedback from others.

In recent years there are also growing concerns about non-genuine user activity online. It has been suggested that some social media users deliberately spread misinformation to pursue specific geopolitical objectives (Bradshaw & Howard, 2019). Moreover, many of the accounts disseminating misinformation have been identified as social bots (Polychronis & Kogan, 2023; Shao et al., 2018). Therefore, at a time where large language models are continuously gaining popularity and provide unprecedented access to this technology, there is growing concern that social media companies provide platforms which can easily be leveraged to maliciously shape public opinion (Hajli et al., 2022; Himelein-Wachowiak et al., 2021). It is thus crucial to examine whether the '(Dis)Trust' button is susceptible to this type of abuse. Thus far, I have studied the impact of genuine '(Dis)Trust' feedback. Future studies should assess whether the results generalize to environments in which the feedback is non-sensical, i.e., the buttons are used randomly and not in a manner that signals the veracity of the content, as well as whether the results generalize to environments in which those using the '(Dis)Trust' button pursue a specific misinformation narrative, and thus does not provide a genuine 'trust' signal.

Misinformation is thought to have contributed to increasing polarization, racism and resistance to climate action and vaccines (Barreto et al., 2021; Rapp & Salovich, 2018; Tsfati et al., 2020; Van Bavel et al., 2021). In light of these detrimental consequences, it is unsurprising that researchers are devoting significant amounts of time to developing new interventions (for a review see Ziemer & Rothmund, 2024). Until now however, few have compared the efficacy of these interventions (but see Hameleers, 2022; Heuer & Glassman, 2022). This is important to ensure that policy and industry resources are assigned to the most promising efforts. Future research should therefore systematically compare the ‘(Dis)Trust buttons to other interventions by determining their efficacy in the same environment. This would also help identify which (or if any) combination of interventions is most effective (Bak-Coleman et al., 2022)

It is also important to acknowledge that participation in our study was limited to the US. Emerging research indicates that cultural norms significantly influence sharing behaviour on social media (Hsu et al., 2021). While US users are more reactive to negative content that others’ have generated, users in Japan respond more to positive content. This disparity might also influence how they use the ‘(Dis)Trust’ buttons and underscores the importance of considering cultural variances in future research. Thus, cultural differences could significantly affect the generalizability and efficacy of the proposed intervention.

Finally, it is important to recognize that this study focussed solely on external incentives. However, it has been shown that sharing is also internally rewarding (Baek et al., 2017; Tamir et al., 2015; Tamir & Mitchell, 2012). Indeed, it is unclear to what extent external incentives provided on social media platforms affect internal incentives. It has been suggested that when external rewards are introduced, activities become less intrinsically rewarding (Deci, 1971). As yet, it is unclear how this relates to sharing decisions. Future studies should strive to explore how external and internal incentives interact to determine sharing behaviour and whether internal incentives could also be utilized to reduce the spread of misinformation online. Related to this, it is plausible that platforms are more likely to implement interventions which improve the user experience and increase their enjoyment of the platform. In the experiments outlined in Chapter 4, I found that users utilized the new trust and distrust buttons more frequently than existing options, and there was no reduction in frequency of sharing. As such, I speculate, that the ‘(Dis)Trust’ buttons did not negatively affect the user experience. Nevertheless, it would be interesting to ask users about their subjective experience of using the buttons.

Taken together, this study presents an innovative approach to reducing misinformation spread on social media. Here I have outlined several limitations which constrain its generalizability to the real world and have outlined future research directions which will help determine the efficacy of this intervention in real-world contexts and will shed light on how social dynamics and network effects impact the efficacy of this (and other) interventions at a population-level.

5.5 Conclusion

Every day humans are faced with copious opportunities to accumulate evidence and sharing information. This behaviour is driven by (1) external incentives, such as financial or social gains (Gold & Shadlen, 2002; Rosenthal-von der Pütten et al., 2019), and (2) internal incentives, such as positive emotions and self-efficacy (Baek et al., 2017; Gesiarz et al., 2019; Leong et al., 2019; Tamir et al., 2015; Tamir & Mitchell, 2012). This thesis explores how the incentive structure people face can be altered to improve the quality of evidence accumulation and sharing behaviour.

First, I test the assumption that evidence accumulation should be less biased when false beliefs are costly. I test this assumption in two ways: In Chapter 2, I examine the impact of exposing participants to a threatening environment in which external punishments for holding negative beliefs may be high; and in Chapter 3, I examine the impact of increasing monetary incentives to form accurate beliefs. I find that while perceived threat induces a global shift towards negative evidence such that evidence accumulation becomes less biased, in a manner that may be adaptive, monetary incentives to form accurate conclusions fail to mitigate biased evidence accumulation. This is because external and internal incentives alter orthogonal elements of the accumulation process. I speculate that stress elicits a global psychophysiological response which elicits an automatic shift in evidence accumulation towards negative evidence. By contrast accuracy incentives alter deliberate processing. While participants try to be more cautious when incentivized for accuracy, this does not change their biased perception, as they believe it to be genuine.

I speculate that feedback will allow participants to become aware of their biases and thus will enable them to update their beliefs in order to maximize external rewards. I test this assumption in Chapter 4 by providing participants with social feedback about the accuracy of the information they share online. I demonstrate that the existing incentive structure of social media platforms, in which social rewards (e.g., ‘likes’) and punishments (e.g., ‘dislikes’) are dissociated from the veracity of the information shared, contributes to the spread of misinformation online. I show that slightly altering this incentive structure, such that social rewards and punishments are contingent on the veracity of information, improves discernment in sharing behaviour and reduces the spread of misinformation. These findings highlight the importance of considering both automatic and deliberate processes when designing interventions to increase the accuracy of decision-making.

In conclusion, this thesis not only sheds light on the mechanisms by which incentives alter evidence accumulation and sharing behaviour, but also bridges the gap between theory and practice, offering actionable strategies that hold the potential to mitigate biased decision-making across digital and physical realms.

Chapter 6: Bibliography

AFR. (2015). *Even money can't end Goldman Sachs' gender "hand-to-hand" combat*. https://www.afr.com/work-and-careers/management/even-money-cant-end-goldman-sachs-gender-hand-to-hand-combat-20150312-141sa0

Akinola, M., & Mendes, W. B. (2012). Stress-induced cortisol facilitates threat-related decision making among police officers. *Behavioral Neuroscience*, *126*(1), 167. https://doi.org/10.1037/a0026657

Alizadeh, M., Hoes, E., & Gilardi, F. (2023). Tokenization of social media engagements increases the sharing of false (and other) news but penalization moderates it. *Scientific Reports*, *13*(1), 13703. https://doi.org/10.1038/s41598-023-40716-2

Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., & Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test. *Neuroscience & Biobehavioral Reviews*, *38*, 94–124. https://doi.org/10.1016/j.neubiorev.2013.11.005

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36). https://doi.org/10.1126/sciadv.abf4393

Allison, P. J., Guichard, C., Fung, K., & Gilain, L. (2003). Dispositional optimism predicts survival status 1 year after diagnosis in head and neck cancer patients. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *21 3*, 543–548. https://doi.org/10.1200/JCO.2003.10.092

Anderson, I. A., & Wood, W. (2021). Habits and the electronic herd: The psychology behind social media's successes and failures. *Consumer Psychology Review*, *4*(1), 83–99. https://doi.org/10.1002/arcp.1063

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, *94*(2), 443–458. https://doi.org/10.1093/biomet/asm017

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*(3), 486. https://doi.org/10.1037/0033-2909.110.3.486

Aylward, J., Hales, C., Robinson, E., & Robinson, O. J. (2019). Translating a rodent measure of negative bias into humans: The impact of induced anxiety and unmedicated mood and anxiety disorders. *Psychological Medicine*, *50*(2), 237–246. https://doi.org/10.1017/S0033291718004117

Baas, J. M., Milstein, J., Donlevy, M., & Grillon, C. (2006). Brainstem Correlates of Defensive States in Humans. *Biological Psychiatry*, *7*(59), 588–593. https://doi.org/10.1016/j.biopsych.2005.09.009

Baek, E. C., Scholz, C., O'Donnell, M. B., & Falk, E. B. (2017). The Value of Sharing Information: A Neural Account of Information Transmission. *Psychological Science*, *28*(7), 851–861. https://doi.org/10.1177/0956797617695073

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, *6*(10), 1372–1380. https://doi.org/10.1038/s41562-022-01388-6

Bălău, N., & Utz, S. (2016). Exposing information sharing as strategic behavior: Power as responsibility and "Trust" buttons. *Journal of Applied Social Psychology*, *46*(10), 593–606. https://doi.org/10.1111/jasp.12388

Balcetis, E., & Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of personality and social psychology*, *91*(4), 612. https://doi.org/10.1037/0022-3514.91.4.612

Barasch, A. (2020). The consequences of sharing. *Current Opinion in Psychology*, *31*, 61–66. https:// doi.org/10.1016/j.copsyc.2019.06.027

Barber, B. M., & Odean, T. (1999). The Courage of Misguided Convictions. *Financial Analysts Journal*. https://doi.org/10.2469/faj.v55.n6.2313

Barreto, M. D. S., Caram, C. D. S., Santos, J. L. G. D., de Souza, R. R., Goes, H. L. D. F., & Marcon, S. S. (2021). Fake news about the COVID-19 pandemic: Perception of health professionals and their families. *Revista Da Escola de Enfermagem*, *55*, 1–9. https://doi.org/10.1590/1980-220X-REEUSP-2021-0007

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21767–21772. https://doi.org/10.1073/pnas.0908104107

Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to*. Simon and Schuster.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586–607. https://doi.org/10.1016/j.jcps.2014.05.002

Berlinghieri, R., Krajbich, I., Maccheroni, F., Marinacci, M., & Pirazzini, M. (2023). Measuring utility with diffusion models. *Science Advances*, *9*(34), eadf1665. https://doi.org/10.1126/sciadv.adf1665

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *Wiley Interdisciplinary Reviews. Cognitive Science*, *5*(1), 61–73. https://doi.org/10.1002/wcs.1266

Birkett, M. A. (2011). The Trier Social Stress Test protocol for inducing psychological stress. *Journal of Visualized Experiments*, *56*, 1–6. https://doi.org/10.3791/3238

Blanchard, D. C., Griebel, G., Pobbe, R., & Blanchard, R. J. (2011). Risk assessment as an evolved threat detection and analysis process. *Neuroscience & Biobehavioral Reviews*, *35*(4), 991–998. https://doi.org/10.1016/j.neubiorev.2010.10.016

Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*(3), 602–614. https://doi.org/10.1016/j.neuron.2014.12.050

Bodaghi, A., & Oliveira, J. (2020). The characteristics of rumor spreaders on Twitter: A quantitative analysis on real data. *Computer Communications*, *160*, 674–687. https://doi.org/10.1016/j.comcom.2020.07.017

Bond, K., Rasero, J., Madan, R., Bahuguna, J., Rubin, J., & Verstynen, T. (2023). Competing neural representations of choice shape evidence accumulation in humans. *Elife*, *12*, e85223. https://doi.org/10.7554/eLife.85223

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4–5), 303–345. https://doi.org/10.1016/S0361-3682(01)00052-6

Bonnevie, E., Sittig, J., & Smyser, J. (2021). The case for tracking misinformation the way we track disease. *Big Data & Society*, *8*(1), 205395172110138. https://doi.org/10.1177/20539517211013867

Botvinick, M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annual Review of Psychology*, *66*(1), 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. https://digitalcommons.unl.edu/scholcom/207/

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), 1–15. https://doi.org/10.1126/sciadv.abe5641

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(5), 2–4. https://doi.org/10.1073/pnas.2020043118

Bromberg-Martin, E. S., Feng, Y.-Y., Ogasawara, T., White, J. K., Zhang, K., & Monosov, I. E. (2024). A neural mechanism for conserved value computations integrating information and rewards. *Nature Neuroscience*, 1–17. https://doi.org/10.1038/s41593-023-01511-4

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., & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neuroscience*, *14*(9), 1209–1216. https://doi.org/10.1038/nn.2902

Bromberg-Martin, E. S., & Sharot, T. (2020). The value of beliefs. *Neuron*, *106*(4), 561–565. https://doi.org/10.1016/j.neuron.2020.05.001

Brudner, E. G., Fareri, D. S., Shehata, S. G., & Delgado, M. R. (2023). Social feedback promotes positive social sharing, trust, and closeness. *Emotion*, *23*(6), 1536. https://psycnet.apa.org/doi/10.1037/emo0001182

Butler, L. H., Prike, T., & Ecker, U. K. H. (2023). *Nudge-Based Misinformation Interventions are Effective in Information Environments with Low Misinformation Prevalence* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-3736230/v1

Calabro, R., Lyu, Y., & Leong, Y. C. (2023). Trial-by-trial fluctuations in amygdala activity track motivational enhancement of desirable sensory evidence during perceptual decision-making. *Cerebral Cortex*, *33*(9), 5690–5703. https://doi.org/10.1093/cercor/bhac452

Calder, M., Craig, C., Culley, D., De Cani, R., Donnelly, C. A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., Hinds, D., Lane, D. C., Mitchell, D., Pavey, G., Robertson, D., Rosewell, B., Sherwin, S., Walport, M., & Wilson, A. (2018). Computational modelling for decision-making: Where, why, what, who and how. *Royal Society Open Science*, *5*(6), 172096. https://doi.org/10.1098/rsos.172096

Capraro, V., & Celadin, T. (2022). “I think this news is accurate”: Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin*. *49*(12), 2635-1645. https://doi.org/10.1177/01461672221117691

Carver, C. S., & Scheier, M. F. (2014). Dispositional optimism. *Trends in Cognitive Sciences*, *18*(6), 293–299. https://doi.org/10.1016/j.tics.2014.02.003

Cavanagh, J. F., Frank, M. J., & Allen, J. J. B. (2011). Social stress reactivity alters reward and punishment learning. *Social Cognitive and Affective Neuroscience*, *6*(3), 311–320. https://doi.org/10.1093/scan/nsq041

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467. https://doi.org/10.1038/nn.2925

Castells, M. (1996). *The information age: Economy, society and culture* (3 volumes). Blackwell, Oxford, 1997, 1998.

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. pui 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

Charness, G., & Fehr, E. (2015). From the lab to the real world. *Science*, *350*(6260), 512–513. https://doi.org/10.1126/science.aad4343

Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, *68*(2), 551–556. https://doi.org/10.1016/j.geb.2009.09.003

Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1800547115

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-05994-9

Chen, X., Sun, M., Wu, D., & Song, X. Y. (2019). Information-sharing behavior on WeChat moments: The role of anonymity, familiarity, and intrinsic motivation. *Frontiers in Psychology*, *10*, 2540. https://doi.org/10.3389/fpsyg.2019.02540

Clithero, J. A. (2018). Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, *69*, 61–86. https:// doi.org/10.1016/j.joep.2018.09.008

Cornwell, B. R., Baas, J. M., Johnson, L., Holroyd, T., Carver, F. W., Lissek, S., & Grillon, C. (2007). Neural responses to auditory stimulus deviance under threat of electric shock revealed by spatially-filtered magnetoencephalography. *NeuroImage*, *37*, 282–289. https://doi.org/10.1016/j.neuroimage.2007.04.055

Cropley, M., & MacLeod, A. K. (2003). Dysphoria, attributional reasoning and future event probability. *Clinical Psychology & Psychotherapy*, *10*(4), 220–227. https://doi.org/10.1002/cpp.360

Dale, D., Rudski, J., Schwarz, A., & Smith, E. (2007). Innumeracy and incentives: A ratio bias experiment. *Judgment and Decision Making*, *2*(4), 243–250. https://doi.org/10.1017/S1930297500000577

Davey, C. G., Allen, N. B., Harrison, B. J., Dwyer, D. B., & Yücel, M. (2010). Being liked activates primary reward and midline self-related brain regions. *Human Brain Mapping*, *31*(4), 660–668. https://doi.org/10.1002/hbm.20895

Davis, M., Walker, D. L., Miles, L., & Grillon, C. (2010). Phasic vs sustained fear in rats and humans: Role of the extended amygdala in fear vs anxiety. *Neuropsychopharmacology*, *35*(1), 105–135. https://doi.org/10.1038/npp.2009.109

De Angelis, M., Bonezzi, A., Peluso, A. M., Rucker, D. D., & Costabile, M. (2012). On Braggarts and Gossips: A Self-Enhancement Account of Word-of-Mouth Generation and Transmission. *Journal of Marketing Research*, *49*(4), 551–563. https://doi.org/10.1509/jmr.11.0136

de Kloet, E. R., Karst, H., & Joëls, M. (2008). Corticosteroid hormones in the central stress response: Quick-and-slow. *Frontiers in Neuroendocrinology*, *29*(2), 268–272. https://doi.org/10.1016/j.yfrne.2007.10.002

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105. https://doi.org/10.1037/h0030644

Delgado, M. R., Fareri, D. S., & Chang, L. J. (2023). Characterizing the mechanisms of social connection. *Neuron*, *111*(24), 3911–3925. https://doi.org/10.1016/j.neuron.2023.09.012

Desai, N., & Krajbich, I. (2022). Decomposing preferences into predispositions and evaluations. *Journal of Experimental Psychology: General*, *151*(8), 1883. https://doi.org/10.1037/xge0001162

Dörnemann, A., Boenisch, N., Schommer, L., Winkelhorst, L., & Wingen, T. (2022). How do Good and Bad News Impact Mood During the Covid-19 Pandemic? The Role of Similarity. *Journal of European Psychology Students*, *13*(1), 107–116. https://doi.org/10.5334/jeps.566

Dunning, D. (2009). Misbelief and the neglect of environmental context. *Behavioral and Brain Sciences*, *32*(6), 517–518. https://doi.org/10.1017/S0140525X09991208

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Edgerly, S., & Vraga, E. K. (2019). The blue check of credibility: Does account verification matter when evaluating news on Twitter? *Cyberpsychology, Behavior, and Social Networking*, *22*(4), 283–287. https://doi.org/10.1089/cyber.2018.0475

Eliaz, K., & Schotter, A. (2007). Experimental Testing of Intrinsic Preferences for NonInstrumental Information. *American Economic Review*, *97*(2), 166–169. https://doi.org/10.1257/aer.97.2.166

Engelmann, J., Lebreton, M., Schwardmann, P., Van Der Weele, J. J., & Chang, L.-A. (2019). Anticipatory Anxiety and Wishful Thinking. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3408017

Engelmann, J. B., & Hein, G. (2013). Contextual and social influences on valuation and choice. *Progress in Brain Research*, *202*, 215–237. https://doi.org/10.1016/B978-0-444-62604-2.00013-7

Engelmann, J. B., Meyer, F., Fehr, E., & Ruff, C. C. (2015). Anticipatory Anxiety Disrupts Neural Valuation during Risky Choice. *The Journal of Neuroscience*, *35*(7), 3085–3099. https://doi.org/10.1523/JNEUROSCI.2880-14.2015

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., & Van De Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of*

Economics and Statistics, *105*(4), 818–832. https://doi.org/10.1162/rest_a_01093

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212. https://doi.org/10.1002/bdm.495

Epstein, Z., Lin, H., Pennycook, G., & Rand, D. (2022). *Quantifying attention via dwell time and engagement in a social media browsing environment* (arXiv:2209.10464). arXiv. https://doi.org/10.48550/arXiv.2209.10464

Epstein, Z., Pennycook, G., & Rand, D. G. (2019). *Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/z3s5k

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*(9), eabo6169. https://doi.org/10.1126/sciadv.abo6169

Fainman, E. Z., & Kucukyazici, B. (2020). Design of financial incentives and payment schemes in healthcare systems: A review. *Socio-Economic Planning Sciences*, *72*, 100901. https://doi.org/10.1016/j.seps.2020.100901

Falk, A., & Heckman, J. J. (2009). Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

Falk, E. B., O'Donnell, M. B., & Lieberman, M. D. (2012). Getting the word out: Neural correlates of enthusiastic message propagation. *Frontiers in Human Neuroscience*, *6*, 313. https://doi.org/10.3389/fnhum.2012.00313

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-009

FeldmanHall, O., Raio, C. M., Kubota, J. T., Seiler, M. G., & Phelps, E. A. (2015). The Effects of Social Context and Acute Stress on Decision Making Under Uncertainty. *Psychological Science*, *26*(12), 1918–1926. https://doi.org/10.1177/0956797615605807

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1), 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Fu, P.-W., Wu, C.-C., & Cho, Y.-J. (2017). What makes users share content on Facebook? Compatibility among psychological incentive, social capital focus, and content type. *Computers in Human Behavior*, *67*, 23–32. https://doi.org/10.1016/j.chb.2016.10.010

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. CRC Press.

Ganguly, A., & Tasoff, J. (2017). Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future. *Management Science*, *63*(12), 4037–4060. https://doi.org/10.1287/mnsc.2016.2550

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *Journal of Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Geers, M., Swire-Thompson, B., Lorenz-Spreen, P., Herzog, S. M., Kozyreva, A., & Hertwig, R. (2024). The Online Misinformation Engagement Framework. *Current Opinion in Psychology*, *55*, 101739. https://doi.org/10.1016/j.copsyc.2023.101739

Gesiarz, F., Cahill, D., & Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLoS Computational Biology*, *15*(6), e1007089. https://doi.org/10.1371/journal.pcbi.1007089

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Giltay, E. J., Geleijnse, J. M., Zitman, F. G., Buijsse, B., & Kromhout, D. (2007). Lifestyle and dietary correlates of dispositional optimism in men: The Zutphen Elderly Study. *Journal of Psychosomatic Research*, *63*(5), 483–490. https://doi.org/10.1016/j.jpsychores.2007.07.014

Globig, L. K., Holtz, N., & Sharot, T. (2023). Changing the incentive structure of social media platforms to halt the spread of misinformation. *Elife*, *12*, e85767. https://doi.org/10.7554/eLife.85767

Globig, L. K., Witte, K., Feng, G., & Sharot, T. (2021). Under threat, weaker evidence is required to reach undesirable conclusions. *Journal of Neuroscience*, *41*(30), 6502–6510. https://doi.org/10.1523/JNEUROSCI.3194-20.2021

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*(31), 10686–10698. https://doi.org/10.1523/JNEUROSCI.0727-12.2012

Gluth, S., Rieskamp, J., & Büchel, C. (2013). Deciding Not to Decide: Computational and Neural Evidence for Hidden Behavior in Sequential Choice. *PLoS Computational Biology*, *9*(10). https://doi.org/10.1371/journal.pcbi.1003309

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*(2), 299–308. https://doi.org/10.1016/S0896-6273(02)00971-6

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Grady, R. H., Ditto, P. H., & Loftus, E. F. (2021). Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications*, *6*(1). https://doi.org/10.1186/s41235-021-00315-z

Grant, S., Kajii, A., & Polak, B. (1998). Intrinsic preference for information. *Journal of Economic Theory*, *83*(2), 233–259. https://doi.org/10.1006/jeth.1996.2458

Graybeal, C., Feyder, M., Schulman, E., Saksida, L. M., Bussey, T. J., Brigman, J. L., & Holmes, A. (2011). Paradoxical reversal learning enhancement by stress or prefrontal cortical damage: Rescue with BDNF. *Nature Neuroscience*, *14*(12), 1507–1509. https://doi.org/10.1038/nn.2954

Grillon, C. (2008). Models and mechanisms of anxiety: Evidence from startle studies. *Psychopharmacology*, *199*(3), 421–437. https://doi.org/10.1007/s00213-007-1019-1

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Gu, R., Ao, X., Mo, L., & Zhang, D. (2020). Neural correlates of negative expectancy and impaired social feedback processing in social anxiety. *Social Cognitive and Affective Neuroscience*, *15*(3), 285–291. https://doi.org/10.1093/scan/nsaa038

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends in Cognitive Sciences*, *18*(4), 194–202. https://doi.org/10.1016/j.tics.2014.01.003

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J. M., Dayan, P., Dolan, R. J., & Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *Journal of Neuroscience*, *31*(21), 7867–7875. https://doi.org/10.1523/JNEUROSCI.6376-10.2011

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

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management*, *33*(3), 1238–1253. https://doi.org/10.1111/1467-8551.12554

Hales, C. A., Robinson, E. S. J., & Houghton, C. J. (2016). Diffusion modelling reveals the decision making processes underlying negative judgement bias in rats. *PLoS ONE*, *11*(3), 1–25. https://doi.org/10.1371/journal.pone.0152592

Hameleers, M. (2022). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society*, *25*(1), 110–126. https://doi.org/10.1080/1369118X.2020.1764603

Harding, E. J., Paul, E. S., & Mendl, M. (2004). Cognitive bias and affective state. *Nature*, *427*(6972), 312–312. https://doi.org/10.1038/427312a

Haselton, M. G., & Nettle, D. (2005). *The paranoid optimist: An integrative evolutionary model of cognitive biases.*

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hasnain, Z., & Pierskalla Henryk, N. (2012). Performance-related pay in the public sector: A review of theory and evidence. *World Bank Policy Research Working Paper*, *6043*. https://ssrn.com/abstract=2043471

Hausmann, D., & Läge, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, *3*(3), 229–243. https://doi.org/10.5167/uzh-6226

Hernandez, R., Kershaw, K. N., Siddique, J., Boehm, J. K., Kubzansky, L. D., Diez-Roux, A., Ning, H., & Lloyd-Jones, D. M. (2015). Optimism and cardiovascular health: Multi-ethnic study of atherosclerosis (MESA). *Health Behavior and Policy Review*, *2*(1), 62–73. https://doi.org/10.14485/HBPR.2.1.6

Heuer, H., & Glassman, E. L. (2022). A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517717

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5), e26933. https://doi.org/10.2196/26933

Hsu, T. W., Niiya, Y., Thelwall, M., Ko, M., Knutson, B., & Tsai, J. L. (2021). Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality and Social Psychology*, *121*(5), 969. https://doi.org/10.1037/pspa0000282

Huang, Y., White, C., Xia, H., & Wang, Y. (2015). Modeling Sharing Decision of Campus Safety Reports and Its Design Implications to Mobile Crowdsourcing for Safety. *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 400–409. https://doi.org/10.1145/2785830.2785889

Jackson, M. O., Malladi, S., & McAdams, D. (2022). Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences*, *119*(34), e2205549119. https://doi.org/10.1073/pnas.2205549119

Jagayat, A., & Choma, B. L. (2023). A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology*, 101726. https://doi.org/10.1016/j.copsyc.2023.101726

Jiang, C., Buchanan, T. W., Yao, Z., Zhang, K., Wu, J., & Zhang, L. (2017). Acute psychological stress disrupts attentional bias to threat-related stimuli. *Scientific Reports*, *7*(1), 14607. https://doi.org/10.1038/s41598-017-14138-w

Jobin, J., Wrosch, C., & Scheier, M. F. (2014). Associations between dispositional optimism and diurnal cortisol in a community sample: When stress is perceived as higher than normal. *Health Psychology*, *33*(4), 382–391. https://doi.org/10.1037/a0032736

Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320. https://doi.org/10.1038/nature10384

Johnston, W. M., & Davey, G. C. L. (1997). The psychological impact of negative TV news bulletins: The catastrophizing of personal worries. *British Journal of Psychology*, *88*(1), 85–91. https://doi.org/10.1111/j.2044-8295.1997.tb02622.x

Kapoor, H., Rezaei, S., Gurjar, S., Tagat, A., George, D., Budhwar, Y., & Puthillam, A. (2023). Does incentivization promote sharing “true” content online? *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-120

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others’ opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Kappes, A., & Sharot, T. (2019). The automatic nature of motivated belief updating. *Behavioural Public Policy*, *3*(1), 87–103. https://doi.org/10.1017/bpp.2017.11

Karayanni, M., & Nelken, I. (2022). Extrinsic rewards, intrinsic rewards, and non-optimal behavior. *Journal of Computational Neuroscience*, *50*(2), 139–143. https://doi.org/10.1007/s10827-022-00813-z

Kelly, S. P., & O’Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441. https://doi.org/10.1523/JNEUROSCI.3355-13.2013

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Kobayashi, K., & Hsu, M. (2019). Common neural code for reward and information value. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1820145116

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, *29*, 6–11. https://doi.org/10.1016/j.copsyc.2018.10.008

Krajbich, I. (2022). Decomposing Implicit Bias. *Psychological Inquiry*, *33*(3), 181–184. https://doi.org/10.1080/1047840X.2022.2106758

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Kreps, S. (2020). The role of technology in online misinformation. *Foreign Policy*. https://www.brookings.edu/wp-content/uploads/2020/06/The-role-of-technology-in-online-misinformation.pdf

Kudielka, B. M., Hellhammer, D. H., & Wüst, S. (2009). Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoneuroendocrinology*, *34*(1), 2–18. https://doi.org/10.1016/j.psyneuen.2008.10.004

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, *28*(2), 331–339. https://doi.org/10.1016/j.chb.2011.10.002

Lees, J., McCarter, A., & Sarno, D. M. (2022). Twitter’s disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among Democrats and Independents. *Journal of Online Trust and Safety*, *1*(3). https://doi.org/10.54501/jots.v1i3.39

Lefebvre, M., Vieider, F. M., & Villeval, M. C. (2011). The ratio bias phenomenon: Fact or artifact? *Theory and Decision*, *71*(4), 615–641. https://doi.org/10.1007/s11238-010-9212-9

Lenow, J. K., Constantino, S. M., Daw, N. D., & Phelps, E. A. (2017). Chronic and Acute Stress Promote Overexploitation in Serial Decision Making. *The Journal of Neuroscience*, *37*(23), 5681–5689. https://doi.org/10.1523/JNEUROSCI.3618-16.2017

Leong, Y. C., Hughes, B. L., Wang, Y., & Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. *Nature Human Behaviour*, *3*(9), 962–973. https://doi.org/10.1038/s41562-019-0637-z

Lerman, C., Hughes, C., Lemon, S. J., Main, D., Snyder, C., Durham, C., Narod, S., & Lynch, H. T. (1998). What you don’t know can hurt you: Adverse psychologic effects in members of BRCA1-linked and BRCA2-linked families who decline genetic testing. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.1998.16.5.1650

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), 1–10. https://doi.org/10.1038/s41467-020-19607-x

Loewenstein, G. (2006). The Pleasures and Pains of Information. *Science*, *312*(5774), 704–706. https://doi.org/10.1126/science.1128388

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition*, *65*(3), 209–237. https://doi.org/10.1016/j.bandc.2007.02.007

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

Lyratzopoulos, G., Vedsted, P., & Singh, H. (2015). Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. *British Journal of Cancer*, *112*(1), S84–S91. https://doi.org/10.1038/bjc.2015.47

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. In *Journal of Experimental Psychology: Applied* (Vol. 27). American Psychological Association. https://doi.org/10.1037/xap0000315

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

McFarland, C., & Ross, M. (1982). Impact of causal attributions on affective reactions to success and failure. *Journal of Personality and Social Psychology*, *43*(5), 937–946. https://doi.org/10.1037/0022-3514.43.5.937

McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*(6), 493–510. https://doi.org/10.1017/S0140525X09990975

Meredith, M., & Kruschke, J. (2016). HDInterval: Highest (posterior) density intervals. *R Package Version 0.1*, *3*.

Meshi, D., Morawetz, C., & Heekeren, H. R. (2013). Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in Human Neuroscience*, *7*(JUL), 1–11. https://doi.org/10.3389/fnhum.2013.00439

Meub, L., & Proeger, T. (2016). Can anchoring explain biased forecasts? Experimental evidence. Journal of Behavioral and Experimental Finance, 12, 1-13. https://doi.org/10.1016/j.jbef.2016.08.001

Montibeller, G., & Von Winterfeldt, D. (2015). Cognitive and Motivational Biases in Decision and Risk Analysis. *Risk Analysis*, *35*(7), 1230–1251. https://doi.org/10.1111/risa.12360

Morelli, S. A., Torre, J. B., & Eisenberger, N. I. (2014). The neural bases of feeling understood and not understood. *Social Cognitive and Affective Neuroscience*, *9*(12), 1890–1896. https://doi.org/10.1093/scan/nst191

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), 7144. https://doi.org/10.1038/s41467-022-34769-6

Mulder, M. J. (2014). The temporal dynamics of evidence accumulation in the brain. *Journal of Neuroscience*, *34*(42), 13870–13871. https://doi.org/10.1523/JNEUROSCI.3251-14.2014

Murray, S., Stanley, M., McPhetres, J., Pennycook, G., & Seli, P. (2020). *" I've said it before and I will say it again": Repeating statements made by Donald Trump increases perceived truthfulness for individuals across the political spectrum*. https://doi.org/10.31234/osf.io/9evzc

Nadkarni, A., & Hofmann, S. G. (2012). Why do people use Facebook? *Personality and Individual Differences*, *52*(3), 243–249. https://doi.org/10.1016/j.paid.2011.11.007

Nesse, R. M., Bhatnagar, S., & Ellis, B. (2016). Evolutionary origins and functions of the stress response system. *Stress: Concepts, Cognition, Emotion, and Behavior*, 95–101. https://doi.org/10.1016/B978-012373947-6.00150-1

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211. https://doi.org/10.1523/JNEUROSCI.08-06-02201.1988

Niederhoffer, K. G., & Pennebaker, J. W. (2009). 621 Sharing One's Story: On the Benefits of Writing or Talking About Emotional Experience. In S. J. Lopez & C. R. Snyder (Eds.), *The Oxford Handbook of Positive Psychology* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195187243.013.0059

Novotny, P., Colligan, R. C., Szydlo, D. W., Clark, M. M., Rausch, S., Wampfler, J., Sloan, J. A., & Yang, P. (2010). A Pessimistic Explanatory Style Is Prognostic for Poor Lung Cancer Survival. *Journal of Thoracic Oncology*, *5*(3), 326–332. https://doi.org/10.1097/JTO.0b013e3181ce70e8

Oster, E., Shoulson, I., & Dorsey, E. R. (2013). Optimal expectations and limited medical testing: Evidence from Huntington disease. *American Economic Review*, *103*(2), 804–830. https://doi.org/10.1257/aer.103.2.804

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*(May 2018), 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Paton, D. (2003). Disaster preparedness: A social-cognitive perspective. *Disaster Prevention and Management: An International Journal*, *12*(3), 210–216. https://doi.org/10.1108/09653560310480686

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. 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., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

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. (2022a). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), 1–12. https://doi.org/10.1038/s41467-022-30073-5

Pennycook, G., & Rand, D. G. (2022b). Nudging Social Media toward Accuracy. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 152–164. https://doi.org/10.1177/00027162221092342

Pew Research Center. (2021). *Social Media Fact Sheet. Pew Research Center: Internet, Science & Tech.* https://www.pewresearch.org/internet/fact-sheet/social-media/

Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & Van Der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, *9*(8). https://doi.org/10.1098/rsos.211953

Piotrkowski, C. S. (1979). *Work and the family system: A naturalistic study of working-class and lower-middle-class families*. Free Press New York; WorldCat.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. https://doi.org/10.1038/22268

Polychronis, C., & Kogan, M. (2023). Working Together (to Undermine Democratic Institutions): Challenging the Social Bot Paradigm in SSIO Research. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–30. https://doi.org/10.1145/3610033

Porcelli, A. J., & Delgado, M. R. (2009). Acute Stress Modulates Risk Taking in Financial Decision Making. *Psychological Science*, *20*(3), 278–283. https://doi.org/10.1111/j.1467-9280.2009.02288.x

Porcelli, A. J., & Delgado, M. R. (2017). Stress and decision making: Effects on valuation, learning, and risk-taking. *Current Opinion in Behavioral Sciences*, *14*, 33–39. https://doi.org/10.1016/j.cobeha.2016.11.015

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2023). *The Misleading count: An identity-based intervention to mitigate the spread of partisan misinformation*. Philosophical Transactions B. https://doi.org/10.1098/rstb.2023.0040

Prior, M., Sood, G., & Khanna, K. (2015). You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions. *Quarterly Journal of Political Science*, *10*(4), 489–518. https://doi.org/10.1561/100.00014127

Raio, C. M., Orederu, T. A., Palazzolo, L., Shurick, A. A., & Phelps, E. A. (2013). Cognitive emotion regulation fails the stress test. *Proceedings of the National Academy of Sciences*, *110*(37), 15139–15144. https://doi.org/10.1073/pnas.1305706110

Rapp, D. N., & Salovich, N. A. (2018). Can’t We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from the Behavioral and Brain Sciences*, *5*(2), 232–239. https://doi.org/10.1177/2372732218785193

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://psycnet.apa.org/doi/10.1016/j.tics.2016.01.007

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. https://doi.org/10.3758/BF03196302

Ratcliff, Roger., & McKoon, Gail. (2008). Drift Diffusion Decision Model:Theory and data. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information.

Nature Human Behaviour, *7*(6), 892–903. https://doi.org/10.1038/s41562-023-01540-w

Ren, Z. (Bella), Dimant, E., & Schweitzer, M. E. (2021). Social Motives for Sharing Conspiracy Theories. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3919364

Macatee, R. J., Albanese, B. J., Schmidt, N. B., & Cougle, J. R. (2017). Attention bias towards negative emotional information and its relationship with daily worry in the context of acute stress: An eye-tracking study. *Behaviour research and therapy*, *90*, 96-110. https://doi.org/10.1016/j.brat.2016.12.013

Rimé, B. (2009). Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review. *Emotion Review*, *1*(1), 60–85. https://doi.org/10.1177/1754073908097189

Roberts, I. D., & Hutcherson, C. A. (2019). Affect and decision making: Insights and predictions from computational models. *Trends in Cognitive Sciences*, *23*(7), 602–614. https://doi.org/10.1016/j.tics.2019.04.005

Robinson, O. J., Charney, D. R., Overstreet, C., Vytal, K., & Grillon, C. (2012). The adaptive threat bias in anxiety: Amygdala–dorsomedial prefrontal cortex coupling and aversive amplification. *Neuroimage*, *60*(1), 523–529. https://doi.org/10.1016/j.neuroimage.2011.11.096

Robinson, O. J., Krimsky, M., & Grillon, C. (2013). The impact of induced anxiety on response inhibition. *Frontiers in Human Neuroscience*, *7*, 69. https://doi.org/10.3389/fnhum.2013.00069

Robinson, O. J., Letkiewicz, A. M., Overstreet, C., Ernst, M., & Grillon, C. (2011). The effect of induced anxiety on cognition: Threat of shock enhances aversive processing in healthy individuals. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(2), 217–227. https://doi.org/10.3758/s13415-011-0030-5

Robinson, O. J., Overstreet, C., Charney, D. R., Vytal, K., & Grillon, C. (2013). Stress increases aversive prediction error signal in the ventral striatum. *Proceedings of the National Academy of Sciences*, *110*(10), 4129–4133. https://doi.org/10.1073/pnas.1213923110

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: Perspectives from human threat of shock studies. *Frontiers in Human Neuroscience*, *7*, 203. https://doi.org/10.3389%2Ffnhum.2013.00203

Rode, H. (2016). To share or not to share: The effects of extrinsic and intrinsic motivations on knowledge-sharing in enterprise social media platforms. *Journal of Information Technology*, *31*, 152–165. https://doi.org/10.1057/jit.2016.8

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., Fleming, S. M., Loosen, A., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 299–302. https://doi.org/10.1038/s41467-020-16278-6

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Rosenthal-von der Pütten, A. M., Hastall, M. R., Köcher, S., Meske, C., Heinrich, T., Labrenz, F., & Ocklenburg, S. (2019). “Likes” as social rewards: Their role in online social comparison and decisions to like other People’s selfies. *Computers in Human Behavior*, *92*(April 2018), 76–86. https://doi.org/10.1016/j.chb.2018.10.017

Roxin, A. (2019). Drift–diffusion models for multiple-alternative forced-choice decision making. *The Journal of Mathematical Neuroscience*, *9*(1), 5. https://doi.org/10.1186/s13408-019-0073-4

Rubin, D. B., & Gelman, A. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.3389%2Ffnhum.2013.00069

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature reviews. Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Rygula, R., Papciak, J., & Popik, P. (2013). Trait Pessimism Predicts Vulnerability to Stress-Induced Anhedonia in Rats. *Neuropsychopharmacology*, *38*(11), 2188–2196. https://doi.org/10.1038/npp.2013.116

Saltz, E., Barari, S., Leibowicz, C., & Wardle, C. (2021). Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(5). https://doi.org/10.37016/mr-2020-81

Sánchez-Fuenzalida, N., Van Gaal, S., Fleming, S. M., Haaf, J. M., & Fahrenfort, J. J. (2023). Predictions and rewards affect decision-making but not subjective experience. *Proceedings of the National Academy of Sciences*, *120*(44), e2220749120. https://doi.org/10.1073/pnas.2220749120

Scheier, M. F., Matthews, K. A., Owens, J. F., Schulz, R., Bridges, M. W., Magovern, G. J., Sr, & Carver, C. S. (1999). Optimism and Rehospitalization After Coronary Artery Bypass Graft Surgery. *Archives of Internal Medicine*, *159*(8), 829–835. https://doi.org/10.1001/archinte.159.8.829

Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences*, *114*(11), 2881–2886. https://doi.org/10.1073/pnas.1615259114

Scholz, C., C. Baek, E., & Falk, E. B. (2023). Invoking self-related and social thoughts impacts online information sharing. *Social Cognitive and Affective Neuroscience*, *18*(1). https://doi.org/10.1093/scan/nsad013

Scholz, C., Jovanova, M., Baek, E. C., & Falk, E. B. (2020). Media content sharing as a value-based decision. *Current Opinion in Psychology*, *31*, 83–88. https://doi.org/10.1016/j.copsyc.2019.08.004

Scissors, L., Burke, M., & Wengrovitz, S. (2016). What's in a Like Attitudes and behaviors around receiving likes on facebook. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *27*, 1501–1510. https://doi.org/10.1145/2818048.2820066

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

Sharot, T. (2021). To quell misinformation, use carrots—Not just sticks. *Nature*, *591*(7850), 347. https://doi.org/10.1038/d41586-021-00657-0

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

Sharot, T., & Sunstein, C. R. (2020). How people decide what they want to know. *Nature Human Behaviour*, *4*(1), 14–19. https://doi.org/10.1038/s41562-019-0793-1

Shefrin, H. M. (2015). How Psychological Pitfalls Generated the Global Financial Crisis. In The Routledge companion to strategic risk management (pp. 289-315). Routledge

Shevlin, B. R. K., Smith, S. M., Hausfeld, J., & Krajbich, I. (2022). High-value decisions are fast and accurate, inconsistent with diminishing value sensitivity. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6). https://doi.org/10.1073/pnas.2101508119

Skinner, B. F. (1966). The behavior of organisms: An experimental analysis. 1938. *New York: Appleton-Century-Crofts*.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*(2), 245–261. https://doi.org/10.1111/j.1465-7295.1993.tb00881.x

Sousa, N., & Almeida, O. F. (2012). Disconnection and reconnection: The morphological basis of (mal) adaptation to stress. *Trends in Neurosciences*, *35*(12), 742–751. https://doi.org/10.1016/j.tins.2012.08.006

Speckmann, F., & Unkelbach, C. (2022). Monetary incentives do not reduce the repetition-induced truth effect. *Psychonomic Bulletin & Review*, *29*(3), 1045–1052. https://doi.org/10.3758/s13423-021-02046-0

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal*

Statistical Society. Series B: Statistical Methodology, *64*(4), 583–616. https://doi.org/10.1111/1467-9868.00353

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). STAI manual for the state-trait anxiety inventory. Self-Evaluation Questionnaire. In *MANUAL*. https://doi.org/10.1037/t06496-000

Stafford, T., Pirrone, A., Croucher, M., & Krystalli, A. (2020). Quantifying the benefits of using decision models with response time and accuracy data. *Behavior Research Methods*, *52*(5), 2142–2155. https://doi.org/10.3758/s13428-020-01372-w

Starcke, K., & Brand, M. (2012). Decision making under stress: A selective review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1228–1248. https://doi.org/10.1016/j.neubiorev.2012.02.003

Statista. (2022). *Number of social network users worldwide from 2017 to 2025. Retrieved January, 4, 2024.* https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Statista. (2023). *Internet and social media users in the world 2023*. *Retrieved January, 4, 2024.* https://www.statista.com/statistics/617136/digital-population-worldwide/

Steptoe, A., Wright, C., Kunz-Ebrecht, S. R., & Iliffe, S. (2006). Dispositional optimism and health behaviour in community-dwelling older people: Associations with healthy ageing. *British Journal of Health Psychology*, *11*(1), 71–84. https://doi.org/10.1348/135910705X42850

Stopczynski, A., Pentland, A. ‘Sandy,’ & Lehmann, S. (2018). How physical proximity shapes complex social networks. *Scientific Reports*, *8*(1), 17722. https://doi.org/10.1038/s41598-018-36116-6

Strunk, D. R., & Adler, A. D. (2009). Cognitive biases in three prediction tasks: A test of the cognitive model of depression. *Behaviour Research and Therapy*, *47*(1), 34–40. https://doi.org/10.1016/j.brat.2008.10.008

Strunk, D. R., Lopez, H., & DeRubeis, R. J. (2006). Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behaviour Research and Therapy*, *44*(6), 861–882. https://doi.org/10.1016/j.brat.2005.07.001

Tamir, D. I., & Hughes, B. L. (2018). Social Rewards: From Basic Social Building Blocks to Complex Social Behavior. *Perspectives on Psychological Science*, *13*(6), 700–717. https://doi.org/10.1177/1745691618776263

Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8038–8043. https://doi.org/10.1073/pnas.1202129109

Tamir, D. I., Zaki, J., & Mitchell, J. P. (2015). Informing others is associated with behavioral and neural signatures of value. *Journal of Experimental Psychology: General*, *144*(6), 1114. https://psycnet.apa.org/doi/10.1037/xge0000122

Taylor, S. E., & Brown, J. D. (1994). Positive Illusions and Well-Being Revisited: Separating Fact From Fiction. *Psychological Bulletin*, *116*(1), 21–27. https://doi.org/10.1037/0033-2909.116.1.21

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., & Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *American Psychologist*, *55*(1), 99–109. https://doi.org/10.1037/0003-066X.55.1.99

Tindle, H. A., Chang, Y.-F., Kuller, L. H., Manson, J. E., Robinson, J. G., Rosal, M. C., Siegle, G. J., & Matthews, K. A. (2009). Optimism, Cynical Hostility, and Incident Coronary Heart Disease and Mortality in the Women’s Health Initiative. *Circulation*, *120*(8), 656–662. https://doi.org/10.1161/CIRCULATIONAHA.108.827642

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 136–151. https://doi.org/10.1177/00027162221087936

Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: Literature review and synthesis. *Annals of the International Communication Association*, *44*(2), 157–173. https://doi.org/10.1080/23808985.2020.1759443

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*(11), 913–916. https://doi.org/10.1016/j.tics.2021.07.013

Van den Bos, R., Harteveld, M., & Stoop, H. (2009). Stress and decision-making in humans: Performance is related to cortisol reactivity, albeit differently in men and women. *Psychoneuroendocrinology*, *34*(10), 1449–1458. https://doi.org/10.1016/j.tics.2018.01.004

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

Vellani, V., Zheng, S., Ercelik, D., & Sharot, T. (2023). The illusory truth effect leads to the spread of misinformation. *Cognition*, *236*, 105421. https://doi.org/10.1016/j.cognition.2023.105421

Verkuil, B., Brosschot, J. F., de Beurs, D. P., & Thayer, J. F. (2009). Effects of explicit and implicit perseverative cognition on cardiac recovery after cognitive stress. *International Journal of Psychophysiology*, *74*(3), 220–228. https://doi.org/10.1016/j.ijpsycho.2009.09.003

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10), 1558–1564. https://doi.org/10.1093/scan/nsx092

Walsh, R. M., Forest, A. L., & Orehek, E. (2020). Self-disclosure on social media: The role of perceived network responsiveness. *Computers in*

Human Behavior, *104*, 106162. https://doi.org/10.1016/j.chb.2019.106162

Webb, T., Joseph, J., Yardley, L., & Michie, S. (2010). Using the internet to promote health behavior change: A systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of Medical Internet Research*, *12*(1), e1376. https://doi.org/10.2196/jmir.1376

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Anxiety Enhances Threat Processing Without Competition Among Multiple Inputs: A Diffusion Model Analysis. *Emotion*, *10*(5), 662–677. https://doi.org/10.1037/a0019474

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, 1–10. https://doi.org/10.3389/fninf.2013.00014

Wien, A. H., & Olsen, S. O. (2014). Understanding the Relationship between Individualism and Word of Mouth: A Self-Enhancement Explanation. *Psychology & Marketing*, *31*(6), 416–425. https://doi.org/10.1002/mar.20704

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

Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, *44*(1), 68–82. https://doi.org/10.1016/0749-5978(89)90035-6

Yechiam, E., & Zeif, D. (2023). The effect of incentivization on the conjunction fallacy in judgments: A meta-analysis. *Psychological Research*, *87*(8), 2336–2344. https://doi.org/10.1007/s00426-023-01837-5

Youssef, F. F., Dookeeram, K., Basdeo, V., Francis, E., Doman, M., Mamed, D., Maloo, S., Degannes, J., Dobo, L., Ditshotlo, P., & Legall, G. (2012). Stress alters personal moral decision making.

Psychoneuroendocrinology, *37*(4), 491–498. https://doi.org/10.1016/j.psyneuen.2011.07.017

Yu, R. (2016). Stress potentiates decision biases: A stress induced deliberation-to-intuition (SIDI) model. *Neurobiology of Stress*, *3*, 83–95. https://doi.org/10.1016/j.ynstr.2015.12.006

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/s0140-6736(20)30461-x

Zhang, Y., & Rand, D. G. (2023). Sincere or motivated? Partisan bias in advice-taking. *Judgment and Decision Making*, *18*, e29. https://doi.org/10.1017/jdm.2023.28

Ziemer, C.-T., & Rothmund, T. (2024). Psychological Underpinnings of Misinformation Countermeasures: A Systematic Scoping Review. *Journal of Media Psychology*, 1864-1105. https://doi.org/10.1027/1864-1105/a000407

Zizzo, D. J., Stolarz-Fantino, S., Wen, J., & Fantino, E. (2000). A violation of the monotonicity axiom: Experimental evidence on the conjunction fallacy. *Journal of Economic Behavior & Organization*, *41*(3), 263–276. https://doi.org/10.1016/S0167-2681(99)00076-1

Chapter 7: Appendix

7.1 Chapter 2: Under threat weaker evidence is required to reach an undesirable conclusion

Supplementary Table 2.1. Mean difference in posterior distributions and 95% HDI Comparison (Experimental Data)

Estimate	Threat Minus Control
Distance between Decision Thresholds (α)	-0.191 [-0.426 0.05]
Non-Decision Time (t0)	-0.056 [-0.299, 0.192]
Starting Point (z)	0.028 [-0.001, 0.054]
Drift Rate (β0)	0.172 [-0.001, 0.304]
Drift Rate Bias (β1)	-0.255 [-0.412, -0.092]

Supplementary Table 2.2. Proportion of correctly identified factories (Simulated Data).

Proportion of correctly identified factories	df	F-value	p-value
Group	(1,78)	1.916	0.17
Valence of Factory	(1,78)	1.766	0.188
Group * Valence of Factory	(1,78)	13.113	<0.001

As in the experimental data, we observed a significant group by valence interaction (F(1,78) = 13.113, p < 0.001, partial η2 = 0.144) in the simulated data. In the simulated control group data, the proportion of correctly categorized desirable factories (M = 0.812, SE = 0.016) was larger than the proportion of correctly categorized undesirable factories (M = 0.735, SE = 0.017; t(42) = 3.552, p < 0.001, Cohen's d = 0.542). This difference was not observed in the simulated threat group data (proportion of correctly identified undesirable factories: M = 0.819, SE = 0.019; proportion of correctly identified desirable factories: M = 0.783, SE = 0.02; t(36) = 1.611, p = 0.116, Cohen's d = 0.265).

7.2 Chapter 3: Futile Rewards: Why Accuracy Incentives Fail to Reduce Biased Evidence Accumulation

7.2.1 Supplementary Methods

Sensitivity. To assess whether accuracy incentives influenced participants' discernment between signal and noise, we calculated each individual's dPrime (d') for each incentive level separately using the *psycho* R package (Makowski, 2018). d' is as a measure of sensitivity from signal detection theory (Pallier, 2002) to distinguish between correct and incorrect trials. It reflects the distance between the signal and signal + noise distributions and is calculated as follows:

$$d' = Z(Hit\ Rate) - Z(False\ Alarm\ Rate)$$

Here, a higher d' value indicates greater sensitivity in differentiating correct from incorrect trials. We then compared the average d' between the different incentive levels (Experiment 1 & Replication: $0 vs $5; Experiment 2: $0 vs $25) using paired t-tests. To substantiate findings that were not statistically significant, Bayes tests were computed.

7.2.2 Supplementary Results

Accuracy incentives do not alter discernment between signal and noise. Thus far we have shown that accuracy incentives do not reduce the bias towards desirable responses, it then follows that they may also be ineffective as a tool to increase accuracy and therefore might not improve participants' ability to discriminate between signal and noise. To test this, we calculated each participant's d', representing their sensitivity in distinguishing correct from incorrect trials, for each incentive level. We entered the d' for each incentive level into a paired t-tests. This revealed that providing incentives for correct responses did not significantly enhance participants' discrimination accuracy (Experiment 1: t(68) = 1.055, p = 0.295, Cohen's d = 0.127, Replication: t(71) = 0.387, p = 0.7, Cohen's d = 0.046, Experiment 2: t(91) = 0.834, p = 0.407, Cohen's d = 0.087; **see Figure S3.1**). Specifically, the d' scores in trials without a reward for correct responses ($0 accuracy incentives - Experiment 1: M = 0.81, SE = 0.078, Replication: M = 0.779, SE = 0.09, Experiment 2: M = 0.731, SE = 0.072) were not significantly different from those in trials where participants were rewarded for accuracy, regardless of the reward amount ($25 accuracy incentives - Experiment 1: M = 0.763, SE = 0.078, Replication: M = 0.762, SE = 0.095, $25 accuracy incentives - Experiment 2: M = 0.767, SE = 0.072). These results hold when controlling for noise (**see Supplementary Tables 3.8 & 3.9 & 3.15**). Again, Bayes tests provide moderate to strong support in favour of the null hypothesis (Experiment 1: BF_{01} = 6.125, Replication: BF_{01} = 10.009, Experiment 2: BF_{01} = 8.623).

Taken together these results illustrate that accuracy incentives do not improve participants' discernment between signal and noise.

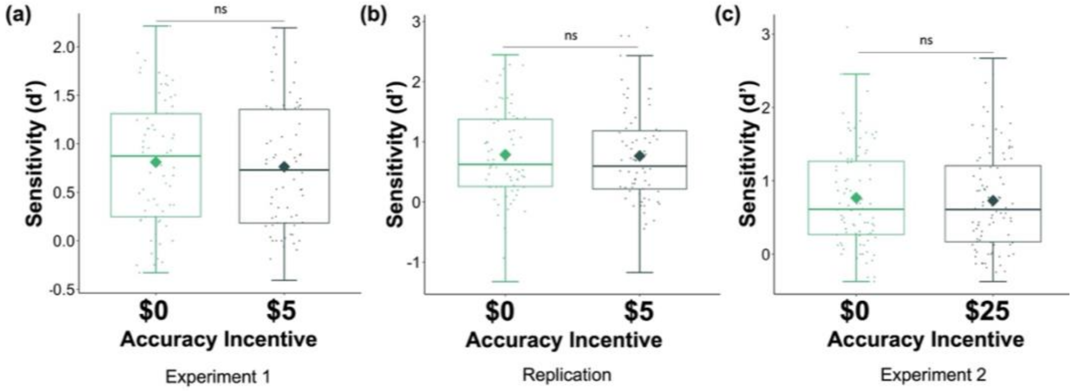

Figure S3.1. Accuracy Incentives do not improve discernment between signal and noise. In **(a)** Experiment 1, **(b) its** replication, and **(c)** Experiment 2 participants' sensitivity (d') did not differ between trials in which they were incentivized for accuracy and those in which they were not incentivized. Y axis shows Sensitivity (d'). X axis shows accuracy incentive level. Data are plotted as boxplots for each incentive level, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean d' score per incentive level. Individuals' d' scores are shown separately as dots. Symbols above each boxplot indicate significance level compared to 0. ns = not significant.

7.2.3 Supplementary Tables

Supplementary Table 3.1. Model Fits (Experiment 1)

Model	Starting Point	Distance between Decision Thresholds	Drift Rate	DIC	BPIC
Model 0	z = 0.5	a ~ 1	v ~ 1	34303	34488
Model 1	0<z<1	a ~ 1	v ~ 1	34315	34509
Model 2	z = 0.5	a ~ 1 + noise	v ~ 1	34232	34434
Model 3	0<z<1	a ~ 1 + noise	v ~ 1	34239	34446
Model 4	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1	34295	34494
Model 5	0<z<1	a ~ 1 + accuracy incentives	v ~ 1	34067	34322
Model 6	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1	34251	34473
Model 7	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1	34015	34293
Model 8	z = 0.5	a ~ 1	v ~ 1 + noise	33761	33999
Model 9	0<z<1	a ~ 1	v ~ 1 + noise	33525	33820
Model 10	z = 0.5	a ~ 1 + noise	v ~ 1 + noise	33764	34007
Model 11	0<z<1	a ~ 1 + noise	v ~ 1 + noise	33528	33829
Model 12	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise	33757	34002
Model 13	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise	33519	34008
Model 14	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	33758	34008
Model 15	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	33522	33827
Model 16	z = 0.5	a ~ 1	v ~ 1 + desirability	33676	33934
Model 17	0<z<1	a ~ 1	v ~ 1 + desirability	33652	33941
Model 18	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability	33625	33906
Model 19	0<z<1	a ~ 1 + noise	v ~ 1 + desirability	33603	33915
Model 20	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability	33669	33932
Model 21	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability	33645	33939
Model 22	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	33618	33906
Model 23	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	33595	33912
Model 24	z = 0.5	a ~ 1	v ~ 1 + noise + desirability	33131	33446
Model 25	0<z<1	a ~ 1	v ~ 1 + noise + desirability	33109	33451
Model 26	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability	33132	33447
Model 27	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability	33111	33458
Model 28	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	33125	33446

Model 29	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	33102	33441
Model 30	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	33130	33453
Model 31	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	33108	33461
Model 32	z = 0.5	a ~ 1	v ~ 1 + accuracy incentives	34304	34502
Model 33	0<z<1	a ~ 1	v ~ 1 + accuracy incentives	34077	34331
Model 34	z = 0.5	a ~ 1 + noise	v ~ 1 + accuracy incentives	34257	34477
Model 35	0<z<1	a ~ 1 + noise	v ~ 1 + accuracy incentives	34024	34302
Model 36	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	34300	34504
Model 37	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	34073	34335
Model 38	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	34251	34478
Model 39	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	34022	34304
Model 40	z = 0.5	a ~ 1	v ~ 1 + noise + accuracy incentives	33762	34005
Model 41	0<z<1	a ~ 1	v ~ 1 + noise + accuracy incentives	33525	33829
Model 42	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	33764	34012
Model 43	0<z<1	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	33528	33833
Model 44	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	33759	34008
Model 45	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	33521	33825
Model 46	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	33759	34014
Model 47	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	33525	33829
Model 48	z = 0.5	a ~ 1	v ~ 1 + desirability + accuracy incentives	33679	33944
Model 49	0<z<1	a ~ 1	v ~ 1 + desirability + accuracy incentives	33655	33949

Model 50	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	33625	33911
Model 51	0<z<1	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	33604	33921
Model 52	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33672	33941
Model 53	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33948	33948
Model 54	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33620	33910
Model 55	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33598	33920
Model 56	z = 0.5	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	33134	33449
Model 57	0<z<1	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	33111	33457
Model 58	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	33138	33460
Model 59	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	33110	33459
Model 60	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33131	33456
Model 61	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33105	33457
Model 62	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33130	33457
Model 63	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33111	33472
Model 64	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + noise*desirability	33105	33457

Supplementary Table 3.2. Model Fits (Replication)

Model	Starting Point	Distance between Decision Thresholds	Drift Rate	DIC	BPIC
Model 0	z = 0.5	a ~ 1	v ~ 1	34852	35057
Model 1	0<z<1	a ~ 1	v ~ 1	34678	34942
Model 2	z = 0.5	a ~ 1 + noise	v ~ 1	34789	35016
Model 3	0<z<1	a ~ 1 + noise	v ~ 1	34613	34899
Model 4	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1	34852	35061
Model 5	0<z<1	a ~ 1 + accuracy incentives	v ~ 1	34679	34946
Model 6	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1	34789	35022
Model 7	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1	34614	34903
Model 8	z = 0.5	a ~ 1	v ~ 1 + noise	34224	34485
Model 9	0<z<1	a ~ 1	v ~ 1 + noise	34041	34359
Model 10	z = 0.5	a ~ 1 + noise	v ~ 1 + noise	34224	34489
Model 11	0<z<1	a ~ 1 + noise	v ~ 1 + noise	34044	34367
Model 12	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise	34225	34492
Model 13	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise	34042	34363
Model 14	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	34222	34491
Model 15	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	34045	34372
Model 16	z = 0.5	a ~ 1	v ~ 1 + desirability	33895	34172
Model 17	0<z<1	a ~ 1	v ~ 1 + desirability	33859	34171
Model 18	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability	33829	34128
Model 19	0<z<1	a ~ 1 + noise	v ~ 1 + desirability	33798	34132
Model 20	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability	33896	34177
Model 21	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability	33859	34173
Model 22	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	33829	34129
Model 23	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	33798	34130
Model 24	z = 0.5	a ~ 1	v ~ 1 + noise + desirability	33269	33607

Model 25	0<z<1	a ~ 1	v ~ 1 + noise + desirability	33235	33610
Model 26	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability	33269	33610
Model 27	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability	33234	33609
Model 28	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	33271	33611
Model 29	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	33233	33604
Model 30	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	33271	33616
Model 31	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	33235	33613
Model 32	z = 0.5	a ~ 1	v ~ 1 + accuracy incentives	34855	35064
Model 33	0<z<1	a ~ 1	v ~ 1 + accuracy incentives	34682	34950
Model 34	z = 0.5	a ~ 1 + noise	v ~ 1 + accuracy incentives	34793	35026
Model 35	0<z<1	a ~ 1 + noise	v ~ 1 + accuracy incentives	34621	34912
Model 36	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	34858	35074
Model 37	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	34682	34953
Model 38	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	34791	35029
Model 39	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	34618	34912
Model 40	z = 0.5	a ~ 1	v ~ 1 + noise + accuracy incentives	34229	34495
Model 41	0<z<1	a ~ 1	v ~ 1 + noise + accuracy incentives	34046	34370
Model 42	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	34226	34494
Model 43	0<z<1	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	34045	34372
Model 44	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	34227	34496
Model 45	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	34048	34375

Model 46	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	34228	34502
Model 47	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	34046	34374
Model 48	z = 0.5	a ~ 1	v ~ 1 + desirability + accuracy incentives	33897	34177
Model 49	0<z<1	a ~ 1	v ~ 1 + desirability + accuracy incentives	33861	34177
Model 50	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	33832	34496
Model 51	0<z<1	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	33797	34133
Model 52	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33898	34182
Model 53	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33865	34183
Model 54	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33830	34134
Model 55	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	33800	34139
Model 56	z = 0.5	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	33273	33615
Model 57	0<z<1	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	33237	33613
Model 58	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	33270	33614
Model 59	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	33234	33614
Model 60	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33273	33618
Model 61	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability +	33237	33615

			accuracy incentives		
Model 62	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33272	33620
Model 63	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	33239	33622
Model 64	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + noise*desirability	33239	33617

Supplementary Table 3.3. Model Fits for Experiment 2

Model	**Starting Point**	**Distance between Decision Thresholds**	**Drift Rate**	**DIC**	**BPIC**
Model 0	z = 0.5	a ~ 1	v ~ 1	46411	46667
Model 1	0<z<1	a ~ 1	v ~ 1	46110	46441
Model 2	z = 0.5	a ~ 1 + noise	v ~ 1	46387	46647
Model 3	0<z<1	a ~ 1 + noise	v ~ 1	46087	46424
Model 4	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1	46388	46674
Model 5	0<z<1	a ~ 1 + accuracy incentives	v ~ 1	46105	46444
Model 6	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1	46366	46661
Model 7	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1	46084	46429
Model 8	z = 0.5	a ~ 1	v ~ 1 + noise	45890	46164
Model 9	0<z<1	a ~ 1	v ~ 1 + noise	45587	45932
Model 10	z = 0.5	a ~ 1 + noise	v ~ 1 + noise	45906	46184
Model 11	0<z<1	a ~ 1 + noise	v ~ 1 + noise	45588	45938
Model 12	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise	45869	46178
Model 13	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise	45586	45945
Model 14	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	45884	46195

Model 15	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise	45586	45950
Model 16	z = 0.5	a ~ 1	v ~ 1 + desirability	45854	46190
Model 17	0<z<1	a ~ 1	v ~ 1 + desirability	45749	46141
Model 18	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability	45846	46189
Model 19	0<z<1	a ~ 1 + noise	v ~ 1 + desirability	45725	46124
Model 20	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability	45844	46201
Model 21	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability	45745	46144
Model 22	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	45832	46192
Model 23	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability	45722	46128
Model 24	z = 0.5	a ~ 1	v ~ 1 + noise + desirability	45364	45749
Model 25	0<z<1	a ~ 1	v ~ 1 + noise + desirability	45256	45701
Model 26	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability	45380	45769
Model 27	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability	45263	45705
Model 28	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	45351	45760
Model 29	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability	45254	45693
Model 30	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	45370	45778
Model 31	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability	45268	45717
Model 32	z = 0.5	a ~ 1	v ~ 1 + accuracy incentives	46410	46673
Model 33	0<z<1	a ~ 1	v ~ 1 + accuracy incentives	46109	46445
Model 34	z = 0.5	a ~ 1 + noise	v ~ 1 + accuracy incentives	46392	46664
Model 35	0<z<1	a ~ 1 + noise	v ~ 1 + accuracy incentives	46103	46448
Model 36	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	46403	45700
Model 37	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + accuracy incentives	46121	46467

Model 38	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	46382	46687
Model 39	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + accuracy incentives	46100	46456
Model 40	z = 0.5	a ~ 1	v ~ 1 + noise + accuracy incentives	45903	46181
Model 41	0<z<1	a ~ 1	v ~ 1 + noise + accuracy incentives	45603	45958
Model 42	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	45904	46186
Model 43	0<z<1	a ~ 1 + noise	v ~ 1 + noise + accuracy incentives	45604	45963
Model 44	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	45885	46199
Model 45	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + accuracy incentives	45595	45959
Model 46	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	45888	46209
Model 47	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + accuracy incentives	45601	45971
Model 48	z = 0.5	a ~ 1	v ~ 1 + desirability + accuracy incentives	45867	46211
Model 49	0<z<1	a ~ 1	v ~ 1 + desirability + accuracy incentives	45764	46165
Model 50	z = 0.5	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	45834	46184
Model 51	0<z<1	a ~ 1 + noise	v ~ 1 + desirability + accuracy incentives	45740	46135
Model 52	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	45859	46207
Model 53	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + desirability + accuracy incentives	45755	46156
Model 54	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	45822	46193
Model 55	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + desirability + accuracy incentives	45725	46127
Model 56	z = 0.5	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	45376	45769
Model 57	0<z<1	a ~ 1	v ~ 1 + noise + desirability + accuracy incentives	45277	45722

Model 58	z = 0.5	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	45381	45779
Model 59	0<z<1	a ~ 1 + noise	v ~ 1 + noise + desirability + accuracy incentives	45275	45727
Model 60	z = 0.5	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	45356	45771
Model 61	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	45257	45709
Model 62	z = 0.5	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	45357	45777
Model 63	0<z<1	a ~ 1 + noise + accuracy incentives	v ~ 1 + noise + desirability + accuracy incentives	45263	45725
Model 64	0<z<1	a ~ 1 + accuracy incentives	v ~ 1 + noise + desirability + noise*desirability	45271	45731

Supplementary Table 3.4. Repeated Measures ANOVA log Response Times (Experiment 1).

Log-transformed RT	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,68)	12.059	<0.001	0.151
Noise	(1,68)	41.101	<0.001	0.377
Noise x Accuracy Incentives	(1,68)	0.075	0.785	0.001

Supplementary Table 3.5. Repeated Measures ANOVA log Response Times (Replication).

Log-transformed RT	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,72)	7.908	0.006	0.099
Noise	(1,72)	66.572	<0.001	0.48
Noise x Accuracy Incentives	(1,72)	0.52	0.4730	0.007

Supplementary Table 3.6. Repeated Measures ANOVA Response Bias (Experiment 1).

Response Bias	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,68)	0.299	0.586	0.004

Noise	(1,68)	0.151	0.699	0.002
Noise x Accuracy Incentives	(1,68)	0.218	0.642	0.003

Supplementary Table 3.7. Repeated Measures ANOVA Response Bias (Replication).

Response Bias	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,72)	0.585	0.447	0.008
Noise	(1,72)	0.033	0.857	0.000
Noise x Accuracy Incentives	(1,72)	0.361	0.550	0.005

Supplementary Table 3.8. Repeated Measures ANOVA dPrime (Experiment 1).

dPrime	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,68)	0.739	0.393	0.011
Noise	(1,68)	107.076	<0.001	0.612
Noise x Accuracy Incentives	(1,68)	0.313	0.578	0.005

Supplementary Table 3.9. Repeated Measures ANOVA dPrime (Replication).

dPrime	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,71)	0.257	0.614	0.004
Noise	(1,71)	66.766	<0.001	0.485
Noise x Accuracy Incentives	(1,71)	0.491	0.486	0.007

Supplementary Table 3.10. 95% HDI Comparisons (Experiment 1 & Replication).

Estimate	**Experiment 1**	**Experiment 2**
Distance between Decision Thresholds (α)	2.29 [2.18, 2.4]	2.32 [2.18, 2.46]
βAccuracy Incentives Distance between Decision Thresholds	0.058 [0.019, 0.092]	0.032 [0.001, 0.072]
Non-decision Time (t0)	6.1 [5.98, 6.23]	6.18 [6.05, 6.31]
Starting Point (z)	0.504 [0.493, 0.515]	0.49 [0.478, 0.501]
inter-trial Starting Point (sz)	0.053 [-0.009, 0.116]	0.061 [-0.003, 0.124]

Drift Rate (β0)	0.093 [-0.009, 0.196]	0.032 [-0.072, 0.146]
βDesirability Drift Rate	0.125 [0.002, 0.264]	0.278 [0.111, 0.449]
βNoise Drift Rate	0.383 [0.294, 0.47]	0.396 [0.29, 0.5]

Supplementary Table 3.11. Repeated Measures ANOVA Response Bias based on simulated data (Experiment 1).

Response Bias	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,68)	0.193	0.662	0.003
Noise	(1,68)	1.736	0.192	0.025
Noise x Accuracy Incentives	(1,68)	0.99	0.323	0.014

Supplementary Table 3.12. Repeated Measures ANOVA Response Bias based on simulated data (Replication).

Response Bias	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,72)	0.087	0.769	0.001
Noise	(1,72)	0.297	0.587	0.004
Noise x Accuracy Incentives	(1,72)	1.255	0.266	0.017

Supplementary Table 3.13. Repeated Measures ANOVA log Response Times (Experiment 2).

Log-transformed RT	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,91)	7.157	0.009	0.073
Noise	(6.77,616.22)	11.842	<0.001	0.115
Noise x Accuracy Incentives	(7.6,691.72)	0.61	0.761	0.007

Supplementary Table 3.14. Repeated Measures ANOVA Response Bias (Experiment 2).

Response Bias	**df**	**F-value**	**p-value**	**partial η2**
Accuracy Incentives	(1,91)	2.254	0.899	0.005
Noise	(9,819)	0.464	0.137	0.024
Noise x Accuracy Incentives	(7.87,716.28)	0.743	0.651	0.008

Supplementary Table 3.15. Repeated Measures ANOVA dPrime (Experiment 2).

dPrime	df	F-value	p-value	partial η2
Accuracy Incentives	(1,91)	1.398	0.240	0.015
Noise	(7.04,640.272)	20.261	<0.001	0.182
Noise x Accuracy Incentives	(9,819)	0.864	0.557	0.009

Supplementary Table 3.16. 95% HDI Comparison in Experiment 3.

Estimate	Experiment 1
Distance between Decision Thresholds (α)	2.3 [2.22, 2.38]
βAccuracy Incentives Distance between Decision Thresholds	0.058 [0.02, 0.094]
Non-decision Time (t0)	6.04 [5.92, 6.15]
Starting Point (z)	0.503 [0.492, 0.514]
inter-trial Starting Point (sz)	0.051 [-0.006, 0.108]
Drift Rate (β0)	-0.018 [-0.08, 0.044]
βDesirability Drift Rate	0.124 [0.034, 0.215]
βNoise Drift Rate	1.68 [1.22, 2.16]

Supplementary Table 3.17. Repeated Measures ANOVA Response Bias based on simulated data (Experiment 2).

Response Bias	df	F-value	p-value	partial η2
Accuracy Incentives	(1,92)	0.001	0.975	0.000
Noise	(9,828)	0.749	0.664	0.08
Noise x Accuracy Incentives	(9,828)	0.811	0.606	0.009

7.3 Chapter 4: Changing the Incentive Structure of Social Media Platforms to halt the Spread of Misinformation

7.3.1 Replication Studies

Methods (Experiment 4-6)

Participants (Experiment 4). Fifty participants residing in the US completed the task on *Prolific Academic* (25 Democrats, 8 Republican, 17 Other, M_{age} = 33.16, SD_{age} ± 9.804; female = 24, male = 25, other = 1, Non-White = 15, White = 35). No participants failed the attention checks. Participants received £7.50 per hour for their participation in addition to a memory test performance-related bonus. For all experiments in this study, ethical approval has been provided by the Research Ethics Committee at University College London and all participants gave informed consent.

Participants (Experiment 5).

Two-hundred and sixty-one participants completed the task on *Prolific Academic* (132 Democrats, 90 Republican, 39 Other, M_{age} = 34.824, SD_{age} ± 12.632; female = 122, male = 131, others = 8, Non-White = 84, White = 177). Participants received £7.50 per hour for their participation in addition to a memory test performance-related bonus.

Participants (Experiment 6).

One-hundred and fifty participants completed the task on *Prolific Academic* (74 Democrats, 14 Republican, 62 Other, M_{age} = 34.2, SD_{age} ± 12.489; female = 70, male = 77, other = 3, Non-White = 39, White = 150). Participants received £7.50 per hour for their participation in addition to a memory test performance-related bonus.

Tasks and Statistical Analysis.

The tasks and analysis in Experiment 4-6 were identical to those used in Experiment 1-3 except for the following differences:

1) In Experiment 4 a 'repost' button was included in addition to 'skip', '(Dis)Like' and '(Dis)Trust' options.
2) In Experiment 5 feedback symbols were coloured – 'distrusts' and 'dislikes' in red and 'trusts' and 'likes' in green, instead of black and white.
3) Experiment 6 contained all 100 posts instead of a selection of 40 posts and did not contain final questions to assess whether participants believed the feedback was real.
4) The samples were not politically balanced (**see Participants Experiment 4-6**), as such analysis did not take into account political orientation.

Results (Experiments 4-6)

Participants use '(Dis)Trust' buttons to discern true from false information (Experiment 4).

Experiment 4 is a replication of Experiment 1, in which participants observe posts (half true half false) and could respond by clicking all, none or some of the following buttons: 'like', 'dislike', 'trust', 'distrust' (**see Methods for details**). Participant's use of '(Dis)Trust' (M = 0.111; SE = 0.01) was more discerning than their use of '(Dis)Like' (M = 0.03; SE = 0.006; F(1,49) = 51.996, $p < 0.001$, partial η2 = 0.51**, Figure S4.1**). They also used negative reactions (M = 0.082; SE = 0.011) in a more discerning manner than positive reactions (M = 0.06, SE = 0.006; F(1,49) = 7.147, p = 0.01, partial η2 = 0.13). Participants' used all reaction buttons, except 'dislike', to discern between true and false posts ('like': M = 0.053; SE = 0.008; t(49) = 6.982, $p < 0.001$, Cohen's d = 0.987; 'trust': M = 0.066; SE = 0.01; t(49) = 6.641, $p < 0.001$, Cohen's d = 0.939; 'dislike': M = 0.007; SE = 0.008; t(49) = 0.883, p = 0.381, Cohen's d = 0.125; 'distrust': M = 0.157; SE = 0.014; t(49) = 11.312, $p < 0.001$, Cohen's d = 1.6). These results hold when including an interaction of type of reaction and valence (**see Supplementary Table 4.18**).

Experiment 4 thus replicates the results of Experiment 1 to show that '(Dis)Trust' buttons are used to discern true from false information.

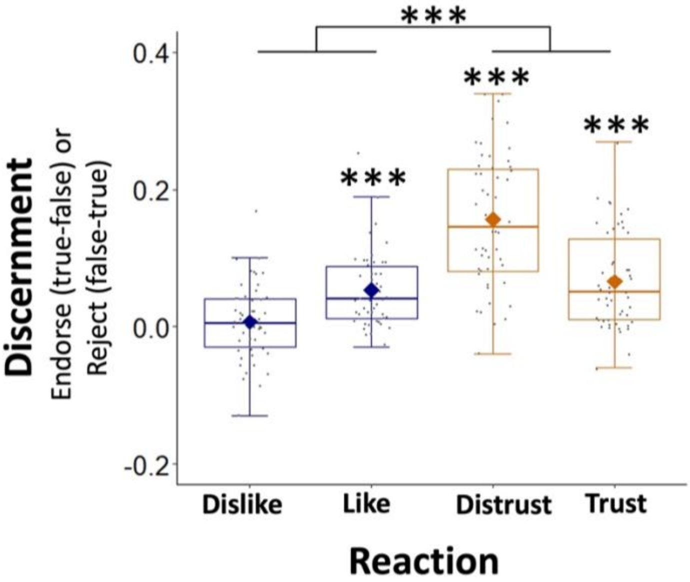

Figure S4.1. Participants' use '(Dis)Trust' to discern true from false information. 'Distrust' and 'trust' reactions were more discerning than 'like' and 'dislike' reactions. Y axis shows discernment between true and false posts. For positive reactions (e.g., *'likes'* and *'trusts'*) discernment is equal to the proportion of positive reactions for true information minus false information, and vice versa for negative reactions (*'dislikes'* and *'distrusts'*). X axis shows reaction buttons. Data are plotted as boxplots for each reaction, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean discernment per reaction. Individuals' mean discernment data are shown separately as grey dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001.

'(Dis)Trust' incentives improve discernment in sharing behaviour (Experiment 5 and Experiment 6). As in Experiment 2, we found an effect of type of feedback (F(1,257) = 8.112, p = 0.005, partial η2 = 0.031). In particular, participants reposted more true relative to false information in the '(Dis)Trust' conditions (M = 0.236, SE = 0.019) than the '(Dis)Like' conditions (M = 0.111, SE = 0.022; F(1,212) = 7.682, p = 0.006, partial η2 = 0.035, **Figure S4.2a**) and Baseline condition (M = 0.102, SE = 0.026; F(1,163) = 16.246, p < 0.001, partial η2 = 0.087). '(Dis)Like' feedback did not improve discernment in sharing behaviour compared to Baseline (F(1,142) = 2.184, p = 0.142, partial η2 =

0.015). No other effects were significant. These results hold when including an interaction of type of feedback and valence (**see Supplementary Table 4.19**).

In line with our results from Experiment 2, we observed an effect of feedback type (F(1,258) = 4.179, p = 0.042, partial η2 = 0.016), such that participants were more accurate (less errors) when they received '(Dis)Trust' feedback (M = 35.717, SE = 0.65) compared to '(Dis)Like' feedback (M = 37.63, SE = 0.767; F(1,212) = 3.955, p = 0.048, partial η2 = 0.018) and also more accurate than those who received no feedback (Baseline, M = 39.73, SE = 0.886; F(1,162) = 11.759, p < 0.001, partial η2 = 0.068). There was no difference in accuracy between participants in the '(Dis)Like' environment and those in the Baseline environment (F(1,143) = 2.746, p = 0.1, partial η2 = 0.019). No other effects were significant. These results hold when allowing for an interaction between type of feedback and valence (**see Supplementary Table 4.20**).

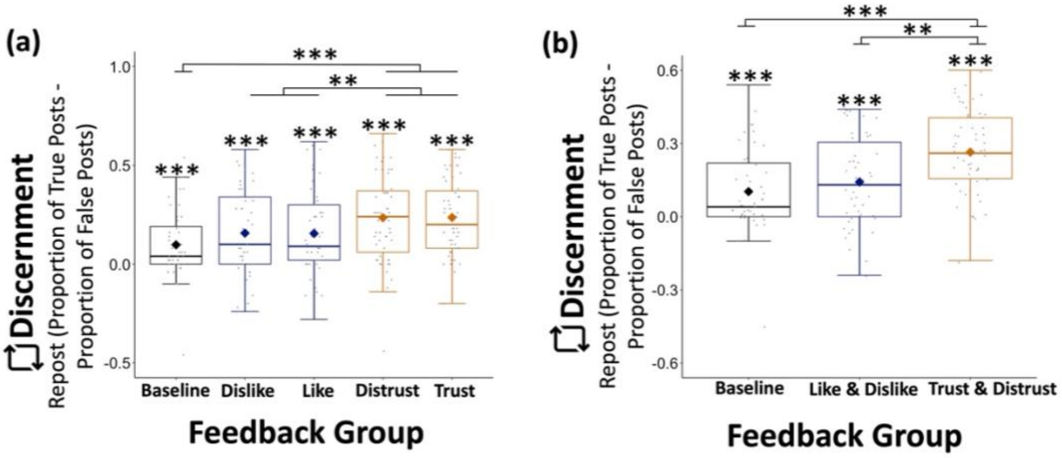

Figure S4.2. '(Dis)Trust' Feedback improves discernment in sharing behaviour. In both **(a)** Experiment 2 and **(b)** Experiment 3 participants who received '(Dis)Trust feedback shared more true relative to false information than participants in the '(Dis)Like' and Baseline conditions. '(Dis)Like' reactions are indicated in blue, '(Dis)Trust in orange and Baseline in black. Y axis shows proportion of true posts shared minus proportion of false posts shared. X axis shows feedback type. Data are plotted as boxplots for each reaction, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 × interquartile range. Diamond shape indicates the mean discernment per reaction. Individuals' mean discernment data are shown separately as grey dots. Symbols above each boxplot indicate significance level compared to 0. ***p < 0.001, **p < 0.01.

In Experiment 6, we again observed an effect of type of feedback (F(1,147) = 11.150, p < 0.001, partial η2 = 0.132, **Figure S2b**), with participants in the 'Trust and Distrust' feedback group posting more true relative to false information (M = 0.264, SE = 0.023) than those in the 'Like and Dislike' group (M = 0.147, SE = 0.027; F(1,101) = 11.122, p = 0.001, partial η2 = 0.099) or those who received no feedback at all (M = 0.106, SE = 0.026; F(1,101) = 21.141, p < 0.001, partial η2 = 0.173). By contrast there was no difference between the latter two groups (F(1,92) = 1.188, p = 0.279, partial η2 = 0.013).

There was also an effect of type of feedback on accuracy (F(1,147) = 4.596, p = 0.012, partial η2 = 0.059). Participants were more accurate when they received '(Dis)Trust' feedback (M = 35.464, SE = 0.596) compared to no feedback (Baseline, M = 39.7, SE = 0.89; F(1,101) = 10.694, p < 0.001, partial η2 = 0.096) with no difference between participants in the '(Dis)Like' environment and those in the Baseline environment (F(1,92) = 1.949, p = 0.11, partial η2 = 0.021) or (Dis)trust environment (M = 37.656, SE = 1.193; F(1,101) = 2.149, p = 0.146, partial η2 = 0.021)

The findings replicate those of Experiment 2 and 3 in suggesting that changing the incentive structure of social media platforms, such that 'carrots' and 'sticks' are partially contingent on accuracy, promotes discernment in sharing behaviour.

'(Dis)Trust' feedback increases the drift rate.

DDM was conducted as for Experiment 2 and 3. For both Experiment 5 (**see Table S4.1**) and Experiment 6 (**see Table S4.2**) we observed a significant difference in the drift rate. That is to say, '(Dis)Trust' feedback (Experiment 5: v = 0.287; 95% CI [0.242, 0.335]; Experiment 6: v = 0.321; 95% CI [0.265, 0.379]) increased the relative importance of the veracity of information when participants decided whether or not to share a post, compared to '(Dis)Like' feedback (Experiment 5: v = 0.198; 95% CI [0.147, 0.248], Experiment 6: 0.172; 95% CI [0.112, 0.233]) or no feedback at all (Experiment 5: v = 0.11; 95% CI [0.048, 0.171]; Experiment 6: v = 0.139; 95% CI [0.065, 0.215]). In Experiment

5 the drift rate in the ‘(Dis)Like’ condition was also higher than in the Baseline condition. While in Experiment 6 there was no difference between the Baseline condition and the ‘(Dis)Like’ condition. As shown in **Table S4.1** and **Table S4.2** the ‘(Dis)Like’ and ‘(Dis)Trust’ conditions did not differ on any other parameters, though some other parameters were different between the Baseline condition and the other conditions, but those were not consistent over all replications. 95% HDI comparisons corroborate this result **(see Supplementary Table 4.21 & 4.22 for HDI Comparisons).** In sum, Experiment 5 and 6 replicate the main results of interest in Experiment 2 and 3.

Table S4.1. Group estimates for DDM in Experiment 5.

Estimate	**Baseline** [95% CI]	**‘(Dis)Like’** [95% CI]	**‘(Dis)Trust’** [95% CI]
Distance between Decision Thresholds (α)	2.2 [2.11, 2.298]	2.451 [2.326, 2.58]	2.458 [2.35, 2.573]
Non-Decision Time (t0)	6.973 [6.799, 7.139]	6.685 [6.429, 6.949]	6.757 [6.541, 6.973]
Starting Point (z)	0.496 [0.484, 0.504]	0.477 [0.467, 0.487]	0.472 [0.463, 0.482]
Drift Rate (v)	0.11 [0.048, 0.171]	0.198 [0.147, 0.248]	0.287 [0.242, 0.335]

Table S4.2. Group estimates for DDM in Experiment 6.

Estimate	**Baseline** [95% CI]	**‘(Dis)Like’** [95% CI]	**‘(Dis)Trust’** [95% CI]
Distance between Decision Thresholds (α)	2.21 [2.117, 2.309]	2.489 [2.298, 2.691]	2.461 [2.312, 2.619]

Non-Decision Time (t0)	6.982 [6.812, 7.151]	6.476 [6.172, 6.769]	6.819 [6.567, 7.063]
Starting Point (z)	0.493 [0.482, 0.505]	0.483 [0.47, 0.497]	0.469 [0.458, 0.48]
Drift Rate (v)	0.139 [0.065, 0.215]	0.172 [0.112, 0.233]	0.321 [0.265, 0.379]

Effects on True and False Posts

Within the field, the gold standard is to measure discernment (that is endorsement of true relative to false posts) rather than measure endorsement of true and/or false posts separately (for a detailed explanation see Guay et al., 2023; Pennycook & Rand, 2021). In our manuscript we follow these recommendations and indeed find that the effects on discernment are very consistent across all experiments. Nevertheless, here we also report the effects of ‘(Dis)Trust’ relative to the effects of ‘(Dis)Like’ on true and false posts separately. We find that sometimes the effect is observed as a decrease in endorsing false posts, sometimes as an increase in endorsing true posts and sometimes as both, but always (as reported in the main text) as an increase in discernment. In particular, participants selected the ‘trust’ reaction button more for true (M = 37.906%, SE = 2.477) than false posts (M = 18.208%, SE = 1.568; t(105) = 9.744, p < 0.001, Cohen’s d = 0.946), and the ‘distrust’ reaction button more for false (M = 49.66%, SE = 2.179) than for true posts (M = 18.509%, SE = 1.476; t(105) = 15.872, p < 0.001, Cohen’s d = 1.542). Moreover, participants selected ‘trust’ (M = 37.906%, SE = 2.477) more than ‘like’ (M = 24.604%, SE = 1.601) for true posts (t(105) = 4.843, p < 0.001, Cohen’s d = 0.47), and selected ‘distrust’ (M = 49.66%, SE = 2.179) more than ‘dislike’ (M = 27.132%, SE = 1.897) for false posts (t(105) = 8.53, p < 0.001, Cohen’s d = 0.829). This latter effect is also observed in Experiment 4, where ‘distrust’ (M = 45.32%, SE = 3.317) is selected more often than ‘dislike’ (M = 13.88%, SE = 1.879) for false posts (t(49) = 8.637, p < 0.001, Cohen’s d = 1.221). With regards to sharing, in Experiment 2 participants shared fewer false posts in the ‘trust’ condition (M = 21.217%, SE = 2.423) than in the ‘like’ condition (M = 29.214%, SE = 2.233; t(133) = 0.244, p = 0.027, Cohen’s d = 0.407), and tended to do the same in the ‘distrust’ condition (M = 23.755%, SE = 2.7) than the ‘dislike’ condition (M =

31.289%, SE = 3.343; t(92) = 1.766, p = 0.081, Cohen's d = 0.365). In Experiment 3, participants shared more true posts in the 'Trust & Distrust' condition (M = 41.241%, SE = 2.162) than the 'Like & Dislike' condition (M = 32.5%, SE = 2.161; t(263) = 2.857, p = 0.005, Cohen's = 0.351). This was repeated in Experiment 6, with true posts shared more in 'Trust & Distrust' condition (M = 46.607%, SE = 3.197) compared to those in the 'Like & Dislike' condition (M = 33.702%, SE = 3.618; t(101) = 2.266, p = 0.026, Cohen's d = 0.448).

7.3.2 Supplementary Tables

Supplementary Tables for Experiment 1-3.

Supplementary Table 4.1. Individual Ratings per Stimulus in Experiment 1.

Stimuli	Veracity	Trusts	Distrusts	Likes	Dislikes	Skips
90% to 95% of those hospitalized for COVID-19 in the United States are unvaccinated.	TRUE	53	29	20	17	12
Over,the past two years, climate and weather disaster damage has cost the US over 400 billion dollars.	TRUE	63	12	22	28	18
The concentration of carbon dioxide in the earth’s atmosphere has climbed to a level last seen more than 3 million years ago.	TRUE	38	26	12	35	25
Flamingos dye their sun-faded feathers to attract mates.	TRUE	22	37	47	4	26
Climate change has made hurricanes more dangerous, but not more frequent.	TRUE	34	34	19	18	26
Babies born in 2020 may suffer up to 7 times as many extreme heatwaves as 1960s kids.	TRUE	47	25	9	42	22
Some dinosaurs may have lived in herds as early as 193 million years ago.	TRUE	48	11	53	4	21
Some birds learn to recognize calls while still in their eggs.	TRUE	44	8	75	1	13
Wild parsnips can cause skin blisters which are dangerous to humans.	TRUE	40	19	20	19	36
A canadian woman was nearly hit by a meteorite that crashed through her bedroom ceiling.	TRUE	41	19	36	12	24
A solar storm hit the earth and brought northern lights to New York.	TRUE	31	33	51	3	18

Climate change is making the earth dimmer.	TRUE	21	43	10	23	35
Rain fell at the normally snowy summit of Greenland for the first time on record.	TRUE	40	20	25	19	26
A third of Antarctic ice shelf risks collapse as our planet warms.	TRUE	63	15	14	42	15
The clothing industry contributes up to 10% to the pollution driving the climate crisis.	TRUE	57	11	13	36	20
Deforestation has made humans more vulnerable to pandemics.	TRUE	36	32	16	28	30
From 2010 to 2017, natural gas production decreased by nearly 70% in New York and increased almost 1000%.	TRUE	23	33	8	13	49
Each year, 324,000 pregnant women experience domestic violence during their pregnancy.	TRUE	57	7	5	64	15
San Francisco had twice as many drug overdose deaths as COVID deaths last year.	TRUE	38	26	8	39	21
Marijuana intake is significantly correlated to psychotic disorders, particularly in teenagers.	TRUE	30	43	11	34	19
Overdose deaths in West Virginia are up by 45% from the prior year.	TRUE	45	11	8	54	19
Tattoo ink isn’t approved by the U.S. Food and Drug Administration.	TRUE	22	32	7	26	39
The fortification of flour with folic acid can prevent certain birth defects.	TRUE	27	33	34	7	32
Replacing table salt with a low-sodium substitute lowers the risk of stroke and	TRUE	56	12	57	4	14

other cardiovascular diseases.						
Professional soccer defenders, who head the ball most often, are almost five times more likely to develop a brain disease than the average person.	TRUE	48	22	12	30	22
Some fast-food items contain plastics linked to serious health problems.	TRUE	45	19	18	32	23
In the last 10 years less than half of the adults in the U.S. received a flu shot.	TRUE	48	20	18	22	21
Burnt seeds show that people used tobacco 12,000 years ago.	TRUE	45	11	41	5	27
Twitter is banned in Iran.	TRUE	43	11	15	30	33
France sets a minimum book delivery fee in effort to protect independent stores from Amazon.	TRUE	37	13	56	4	28
82% of gun owners in the U.S. support requiring all gun buyers to pass a background check.	TRUE	49	16	68	4	12
The US is the only modern industrialized country that does not already have a paid family medical leave.	TRUE	55	15	11	48	19
Canada charges the U.S. a 270% tariff on dairy products.	TRUE	17	36	10	37	25
China imposes the death penalty on drug dealers.	TRUE	39	10	20	29	31
The Texas power grid is not part of the U.S. power grid because Texas wanted to avoid federal regulation.	TRUE	46	30	15	23	26
Black-Lives-Matter apparels and other political expressions were banned at the 2021 Olympics.	TRUE	23	30	22	32	26
Having higher testosterone levels	TRUE	64	13	39	9	17

generally provides an advantage in athletic performance.						
Astronomers may have spotted the first known exoplanet in another galaxy.	TRUE	44	9	66	1	21
Surgeons in New York City successfully attached a pig kidney to a human patient.	TRUE	50	17	39	5	29
China’s lunar rock samples show that lava flowed on the moon 2 billion years ago.	TRUE	27	27	40	4	31
All identical twins may share a common set of chemical markers on their DNA.	TRUE	62	4	48	0	23
Scientists used a tiny brain implant to help a blind teacher see letters again.	TRUE	42	17	69	2	13
30 years ago, the US. ranked sixth among advanced economies in the share of women in the workforce. Today they rank 23th.	TRUE	26	37	12	29	29
Approximately 60,000 Canadians currently live undocumented in the USA.	TRUE	29	29	17	18	38
Rwanda presently has the global highest number of women in parliament at 61.1%.	TRUE	36	9	63	3	26
More than a million Californians do not have direct access to clean water.	TRUE	33	33	5	40	21
Inflation is costing US households with medium incomes an extra $175 a month.	TRUE	70	8	16	54	9
People of color and those of not use marijuana at the same rate. People of color are 3 times more likely to get arrested and convicted.	TRUE	58	13	14	36	24

10% of children in the United States are Texans.	TRUE	21	30	16	7	43
More than a half of Afghanistan’s population faces an acute food crisis this winter.	TRUE	62	2	10	48	15
Reports show that babies whose mothers got the COVID-19 vaccine died after breastfeeding.	FALSE	3	92	2	47	11
In the United Kingdom, 81% of COVID- 19 deaths are among the vaccinated.	FALSE	15	75	5	34	14
United Airlines lost 40% of their staff because of vaccine requirements.	FALSE	46	31	13	42	13
COVID-19 vaccines contain graphene and living organisms made of aluminum.	FALSE	6	70	3	30	29
At least two children died in Australia after receiving COVID-19 vaccines in a mass vaccination program for students.	FALSE	31	37	2	45	18
Test results show that Covid-19 shots damage the immune system.	FALSE	14	80	4	36	12
Children are more likely to end up in the hospital because of the vaccine than they are because of COVID	FALSE	15	75	7	49	12
Vaccinated individuals are more susceptible to COVID variant infections than unvaccinated.	FALSE	13	80	5	38	12
Cancer increased twentyfold among COVID-19 vaccinated due to suppressed T cells.	FALSE	12	78	0	36	16
COVID-19 PCR tests cannot differentiate between flu and COVID-19.	FALSE	10	73	7	31	14

A wind turbine could never generate as much energy as was invested in building it.	FALSE	27	59	13	31	12
90 percent of the world's glaciers are growing.	FALSE	11	69	18	17	25
Forest fires are caused by poor management. Not by climate change.	FALSE	26	55	17	29	18
Ivermectin sterilizes the majority (85%) of the men who take it.	FALSE	7	79	6	24	20
The Centers for Disease Control and Prevention warn of a polio-like outbreak in children coming within the next four months.	FALSE	15	59	4	30	19
It would cost $20 billion to end homelessness in the U.S. and halting global warming would cost $300 billion.	FALSE	34	33	24	14	25
An air quality test under a mask proved that it is not healthy to wear one.	FALSE	19	69	15	35	13
Flu cases dropped by 379 million in one year.	FALSE	31	37	37	8	24
New York hospitals reported thousands of fungal lung infections from mask-wearing.	FALSE	15	70	5	40	15
The Impossible Burger contains more estrogen than transgender hormone therapy.	FALSE	13	73	4	27	20
Prenatal ultrasounds carry extreme risks, including miscarriage and genetic damage.	FALSE	5	89	2	32	14
For men, a positive pregnancy test equals testicular cancer.	FALSE	13	73	8	26	20
AIDS was cured in more than a dozen patients.	FALSE	26	37	46	7	22

Abortion increases the risk of breast cancer.	FALSE	12	72	6	41	17
It is possible to completely detox the body from chemicals.	FALSE	18	62	32	14	13
Two forms of fluoride in our drinking water are labeled as extremely toxic by the Centers for Disease Control and Prevention.	FALSE	29	48	9	37	20
Sophia Stewart wrote books in the 70s that were stolen from her by Warner Bros. She won Hollywood‘s biggest lawsuit.	FALSE	22	14	37	11	48
Gravestones in Japanese cemeteries have QR-codes which can be scanned to get a biography of the deceased person.	FALSE	21	33	44	13	24
52% of Metropolitan Police officers have been found guilty of sexual misconduct while wearing uniform in the line of duty.	FALSE	20	55	1	47	16
Most of the money made by the National Football League goes to the players.	FALSE	17	61	14	18	22
The Biden administration gifted the Taliban with over $80 billion worth of military grade weapons.	FALSE	31	53	3	52	11
Sweden is abolishing cash.	FALSE	18	29	21	27	35
Refugees get more in monthly benefits than social security recipients.	FALSE	24	47	9	48	12
Members of Congress and their families and staff are exempt from repaying student loans.	FALSE	15	38	7	58	23
A single immigrant can bring an unlimited number of relatives to the U.S.	FALSE	15	61	6	37	18

The Soviet Union took all of its equipment from Afghanistan.	FALSE	9	62	6	19	32
There wasn’t a single American casualty in Afghanistan in the last year and a half of the Trump administration.	FALSE	23	60	26	13	19
More than half of the human genes are identical to those of mice.	FALSE	43	30	23	7	29
The top 1% pays 90% of income taxes in the U.S.	FALSE	18	69	10	30	12
The American murder rate is 50 times that of any other developed nation.	FALSE	42	28	8	49	14
The U.S. poverty rate is the 4th highest in the world.	FALSE	37	33	5	45	20
In the United States, 50 percent of social services are provided by the Catholic church.	FALSE	13	52	10	17	35
Marine fossils found on the Mount Everest are evidence of a global flooding.	FALSE	35	23	41	6	32
Joe Biden’s climate plan includes cutting 90% of red meat from our diets by 2030.	FALSE	9	65	12	51	10
14,000 abandoned wind turbines litter the United States.	FALSE	29	29	9	49	22
Wildfires were worse in the early part of the 1900s than they are today.	FALSE	16	59	3	18	30
The U.S. corn crop, at its peak, produces 40% more oxygen than the Amazon rainforest.	FALSE	21	41	39	6	25
An electric car costs more than seven times as much as a gasoline powered car.	FALSE	22	57	7	32	21
The amount of coral on the Great Barrier Reef is at record high levels.	FALSE	56	35	10	19	22

There has not been a long-term distinctive change in sea level rise rates in the last 120 years.	FALSE	62	14	24	21	25

Supplementary Table 4.2. Discernment of reactions (Experiment 1).

Discernment	df	F-value	p-value
including demographics			
Type of Reaction	(1,104)	95.832	<0.001
Valence of Reaction	(1,105)	17.33	<0.001
Gender	(1,101)	8.698	0.004
Political Orientation	(1,101)	26.928	<0.001
Ethnicity	(1,101)	0.276	0.601
Age	(1,101)	0.884	0.349
Type of Reaction x Political Orientation	(1,104)	24.084	<0.001
including valence x reaction			
Type of Reaction	(1,106)	80.936	<0.001
Valence of Reaction	(1,106)	18.26	<0.001
Type of Reaction x Valence of Reaction	(1,106)	51.489	<0.001

The interaction of type and valence of reaction is characterized by participants using the ‘distrust’ reaction button (M = 0.157, SE = 0.008) in a more discerning manner than the ‘trust’ reaction button (M = 0.099, SE = 0.008; t(106) = 9.338, p < 0.001, Cohen’s d = 0.903), but the ‘like’ reaction button (M = 0.06, SE = 0.008) in a more discerning manner than the ‘dislike’ button (M = 0.034, SE = 0.008; t(106) = 3.474, p < 0.001, Cohen’s d = 0.336).

Supplementary Table 4.3. % Reactions out of all posts. (Experiment 1).

% Reactions	df	F-value	p-value
including demographics			
Type of Reaction	(1,104)	36.672	<0.001
Valence of Reaction	(1,105)	13.964	<0.001
Gender	(1,101)	0.891	0.347
Political Orientation	(1,101)	0.939	0.335
Ethnicity	(1,101)	0.139	0.71
Age	(1,101)	9.519	0.003
Type of Reaction x Political Orientation	(1,104)	0.062	0.803

including valence x reaction			
Type of Reaction	(1,106)	37.785	<0.001
Valence of Reaction	(1,106)	14.891	<0.001
Type of Reaction x Valence of Reaction	(1,106)	0.19	0.664

Supplementary Table 4.4. % true and false posts shared out of all true or false posts in that feedback condition (Experiment 2).

Feedback Condition	**%True Posts Shared out of all true posts (SE)**	**%False Posts Shared out of all false posts (SE)**	**% of True Posts Shared Minus % False Posts Shared**
Trust	40 (3.419)	21 (2.423)	18 (2.592)
Like	37 (2.536)	29 (2.233)	8 (2.083)
Distrust	41 (3.192)	24 (2.7)	18 (2.646)
Dislike	40 (3.715)	31 (3.343)	9 (3.92)
Baseline	32 (3.537)	23 (2.817)	8 (2.498)

Supplementary Table 4.5. Discernment of sharing behaviour (Experiment 2).

Discernment	**df**	**F-value**	**p-value**
including demographics			
Intercept	(1,278)	15.286	<0.001
Type of Feedback	(1,278)	14.908	<0.001
Valence of Feedback	(1,278)	0.105	0.746
Gender	(1,278)	2.977	0.086
Political Orientation	(1,278)	66.606	<0.001
Ethnicity	(1,278)	0.688	0.408
Age	(1,278)	0.071	0.791
Type of Feedback x Political Orientation	(1,278)	3.012	0.051
including valence x reaction			
Intercept	(1,311)	105.905	0.001
Type of Feedback	(1, 311)	12.238	<0.001
Valence of Feedback	(1, 311)	0.012	0.913
Type of Feedback x Valence of Feedback	(1,311)	0.199	0.656

Supplementary Table 4.6. % posts shared out of all posts (Experiment 2).

% posts shared	**df**	**F-value**	**p-value**
including demographics			
Intercept	(1,278)	78.161	<0.001
Type of Feedback	(1,278)	1.031	0.311
Valence of Feedback	(1,278)	1.219	0.270
Gender	(1,278)	4.479	0.035
Political Orientation	(1,278)	1.518	0.219
Ethnicity	(1,278)	2.341	0.127
Age	(1,278)	0.032	0.858
Type of Feedback x Political Orientation	(1,278)	0.117	0.890
including valence x reaction			
Intercept	(1,311)	701.419	<0.001
Type of Feedback	(1, 311)	1.533	0.217
Valence of Feedback	(1, 311)	0.741	0.39
Type of Feedback x Valence of Feedback	(1, 311)	0	0.986

Supplementary Table 4.7. Belief Accuracy (Experiment 2).

Belief Accuracy	**df**	**F-value**	**p-value**
including demographics			
Intercept	(1,278)	78.161	<0.001
Type of Feedback	(1,278)	7.679	0.006
Valence of Feedback	(1,278)	0.386	0.535
Gender	(1,278)	2.112	0.147
Political Orientation	(1,278)	7.593	0.006
Ethnicity	(1,278)	3.984	0.047
Age	(1,278)	0.038	0.845
Type of Feedback x Political Orientation	(1,278)	0.171	0.843
including valence x reaction			
Intercept	(1,311)	7536.676	<0.001
Type of Feedback	(1,311)	8.847	0.003
Valence of Feedback	(1,311)	0.948	0.331
Type of Feedback x Valence of Feedback	(1,311)	0.323	0.57

Supplementary Table 4.8. Discernment of sharing behaviour (Experiment 3).

Discernment	**df**	**F-value**	**p-value**

including demographics			
Intercept	(1,381)	1.231	0.268
Type of Feedback	(1,381)	11.028	<0.001
Gender	(1,381)	1.357	0.259
Political Orientation	(1,381)	6.233	0.013
Ethnicity	(1,381)	0.169	0.682
Age	(1,381)	0.002	0.968
Type of Feedback x Political Orientation	(1,381)	1.524	0.219
Intercept	(1,400)	42.658	<0.001
Type of Feedback	(1, 400)	11.416	<0.001

Supplementary Table 4.9. % posts shared out of all posts (Experiment 3).

% posts shared	**df**	**F-value**	**p-value**
including demographics			
Intercept	(1,381)	41.384	<0.001
Type of Feedback	(1,381)	8.97	<0.001
Gender	(1,381)	0.614	0.542
Political Orientation	(1,381)	3.98	0.047
Ethnicity	(1,381)	4.732	0.03
Age	(1,381)	1.146	0.285
Type of Feedback x Political Orientation	(1,381)	0.342	0.71
Intercept	(1,400)	932.702	<0.001
Type of Feedback	(1, 400)	8.897	<0.001

Supplementary Table 4.10. Belief Accuracy (Experiment 3).

Belief Accuracy	**df**	**F-value**	**p-value**
including demographics			
Intercept	(1,381)	1123.65	<0.001
Type of Feedback	(1,381)	3.248	0.04
Gender	(1,381)	14.164	<0.001
Political Orientation	(1,381)	14.749	<0.001
Ethnicity	(1,381)	3.486	0.063
Age	(1,381)	25.786	<0.001
Type of Feedback x Political Orientation	(1,381)	0.301	0.74
Intercept	(1,400)	18657.083	<0.001
Type of Feedback	(1,400)	1.043	0.353

Supplementary Table 4.11. Mean difference in posterior distributions and 95% HDI Comparison in Experiment 2.

Estimate	'(Dis)Trust' minus Baseline	'(Dis)Trust' minus '(Dis)Like'	'(Dis)Like' minus Baseline
Distance between Decision Thresholds (α)	0.25 [0.105, 0.398]	0.03 [-0.13, 0.188]	0.22 [0.102, 0.339]
Non-Decision Time (t0)	-0.34 [-0.644, -0.039]	-0.255 [-0.557,0.045]	-0.089 [-0.291, 0.11]
Starting Point (z)	-0.016 [-0.032, 0.001]	-0.011 [-0.025,0.003]	-0.005 [-0.02, 0.009]
Drift Rate (v)	0.118 [0.041, 0.195]	0.115 [0.048, 0.183]	0.002[-0.075, 0.08]

Supplementary Table 4.12. Mean difference in posterior distributions and 95% HDI Comparison in Experiment 3.

Estimate	'(Dis)Trust' minus Baseline	'(Dis)Trust' minus '(Dis)Like'	'(Dis)Like' minus Baseline
Distance between Decision Thresholds (α)	-0.029 [-0.15, 0.091]	0.002 [-0.114, 0.119]	-0.031 [-0.155, 0.095]
Non-Decision Time (t0)	0.176 [-0.03, 0.381]	0.025 [-0.177, 0.224]	0.151 [-0. 057, 0.354]
Starting Point (z)	-0.011 [-0.027, 0.005]	-0.012 [-0.028, 0.003]	0.001 [-0.015, 0.016]
Drift Rate (v)	0.114 [0.061, 0.167]	0.083 [0.032, 0.135]	0.031 [-0.0164, 0.079]

Supplementary Table 4.13. Recovered Group estimates for DDM in Experiment 2 based on simulated data.

Estimate	**Baseline** [95% CI]	**'(Dis)Like'** [95% CI]	**'(Dis)Trust'** [95% CI]
Distance between Decision Thresholds (α)	2.143 [2.11, 2.176]	2.336 [2.314, 2.358]	2.379 [2.35, 2.41]
Non-Decision Time (t0)	7.016 [6.986, 7.046]	6.945 [6.929, 6.962]	6.667 [6.648, 6.687]
Starting Point (z)	0.499 [0.479, 0.518]	0.501 [0.485, 0.517]	0.486 [0.468, 0.504]

Drift Rate (v)	0.107 [0.068, 0.146]	0.107 [0.081, 0.133]	0.24 [0.209, 0.27]

Supplementary Table 4.14. Recovered Group estimates for DDM in Experiment 3 based on simulated data.

Estimate	**Baseline** [95% CI]	**'(Dis)Like'** [95% CI]	**'(Dis)Trust'** [95% CI]
Distance between Decision Thresholds (α)	2.247 [2.209, 2.284]	2.213 [2.183, 2.243]	2.195 [2.163, 2.228]
Non-Decision Time (t0)	6.904 [6.884, 6.926]	7.04 [7.021, 7.059]	7.083 [7.061, 7.105]
Starting Point (z)	0.493 [0.475, 0.511]	0.515 [0.498, 0.531]	0.485 [0.468, 0.501]
Drift Rate (v)	-0.005 [-0.035, 0.024]	0.007 [-0.024, 0.036]	0.141 [0.1, 0.181]

Supplementary Table 4.15. Pairwise Comparisons for Discernment Experiment 2.

Pairwise Comparison	**Experimental Data**			**Simulated Data**		
	Mean 1 (SE)	**Mean 2 (SE)**	**Statistic**	**Mean 1 (SE)**	**Mean 2 (SE)**	**Statistic**
(Dis)Trust vs Baseline	0.18 (0.018)	0.109 (0.028)	t(152) = 3.112, p = 0.002, Cohen's d = 0.515	0.2 (0.027)	0.109 (0.028)	t(152) = 2.243, p = 0.026, Cohen's d = 0.372
(Dis)Trust vs (Dis)Like	0.18 (0.018)	0.085 (0.019)	t(227) = 3.464, p < 0.001, Cohen's d = 0.465	0.2 (0.027)	0.118 (0.022)	t(227) = 2.407, p = 0.017, Cohen's d = 0.323
(Dis)Like vs Baseline	0.085 (0.019)	0.084 (0.025)	t(191) = 0.007, p = 0.995, Cohen's d = 0.001	0.118 (0.022)	0.109 (0.028)	t(191) = 0.255, p = 0.822, Cohen's d = 0.035

Supplementary Table 4.16. Pairwise Comparisons for Discernment Experiment 3.

	Experimental Data	**Simulated Data**

Pairwise Comparison	**Mean 1 (SE)**	**Mean 2 (SE)**	**Statistic**	**Mean 1 (SE)**	**Mean 2 (SE)**	**Statistic**
(Dis)Trust vs Baseline	0.101 (0.015)	0.008 (0.014)	t(261) = 4.498, p < 0.001, Cohen's d = 0.555	0.116 (0.024)	-0.019 (0.025)	t(261) = 3.826, p < 0.001, Cohen's d = 0.285
(Dis)Trust vs (Dis)Like	0.101 (0.015)	0.042 (0.013)	t(263) = 2.958, p = 0.003, Cohen's d = 0.364	0.116 (0.024)	0.032 (0.024)	t(263) = 2.463, p = 0.014, Cohen's d = 0.303)
(Dis)Like vs Baseline	0.042 (0.013)	0.008 (0.014)	t(252) = 1.731, p = 0.085, Cohen's d = 0.217	0.032 (0.024)	-0.019 (0.025)	t(252) = 1.476, p = 0.141, Cohen's d = 0.185

Supplementary Table 4.17. Correlations between participants' real and recovered DDM estimates in Experiment 2 and Experiment 3.

Estimate	**Experiment 2**	**Experiment 3**
Distance between Decision Thresholds (α)	*r* = 0.926, *p* < 0.001	*r* = 0.886, *p* < *0.0*01
Non-Decision Time (t0)	*r* = 0.997, *p* < *0.0*01	*r* = 0.995, *p* < *0.001*
Starting Point (z)	*r* = 0.471, *p* < *0.0*01	*r* = 0.321, *p* < *0.001*
Drift Rate (v)	*r* = 0.869, *p* < *0.0*01	*r* = 0.877, *p* < *0.0*01

We estimated both group-level and individual-level parameters. We then used the individual-level parameter estimates to simulate data for each participant respectively in the dataset. We used the same number of trials as in the experiments. Simulated data from each participant were then combined and used to perform model recovery analysis. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. We then correlated the real and the recovered individual-level parameters.

Supplementary Tables for Experiment 4-6.

Supplementary Table 4.18. Discernment of reactions (Experiment 4, including type x valence of reaction interaction).

Discernment	df	F-value	p-value
Type of Reaction	(1,49)	51.996	<0.001
Valence of Reaction	(1,49)	7.147	0.01
Type of Reaction * Valence of Reaction	(1,49)	71.625	<0.001

The interaction is characterized by participants using the ‘distrust’ reaction button (M = 0.157, SE = 0.014) in a more discerning manner than the ‘trust’ reaction button (M = 0.066, SE = 0.01; t(49) = 7.192, p < 0.001, Cohen’s d = 1.017), but the ‘like’ reaction button (M = 0.053, SE = 0.008) in a more discerning manner than the ‘dislike’ button (M = 0.007, SE = 0.008; t(49) = 4.407, p < 0.001, Cohen’s d = 0.623).

Supplementary Table 4.19. Discernment of sharing behaviour (Experiment 5).

Discernment	df	F-value	p-value
Intercept	(1,256)	157.841	<0.001
Type of Feedback	(1,256)	8.08	0.005
Valence of Feedback	(1,256)	0.009	0.927
Type of Feedback * Valence of Feedback	(1,256)	0.001	0.982

Supplementary Table 4.20. Belief Accuracy (Experiment 5).

Belief Accuracy	df	F-value	p-value
Intercept	(1,257)	6727.546	<0.001
Type of Feedback	(1,257)	4.151	0.043
Valence of Feedback	(1,257)	2.591	0.109
Type of Feedback * Valence of Feedback	(1,257)	0.013	0.909

Supplementary Table 4.21. Mean difference in posterior distributions and 95% HDI Comparison in Experiment 5.

Estimate	'(Dis)Trust' minus Baseline	'(Dis)Trust' minus '(Dis)Like'	'(Dis)Like' minus Baseline
Distance between Decision Thresholds (α)	0.258 [0.101 0.415]	0.007 [-0.17, 0.186]	0.251 [0.085, 0.418]
Non-Decision Time (t0)	-0.216 [-0.502, 0.07]	0.072 [-0.293, 0.432]	-0.288 [-0. 622, 0.046]
Starting Point (z)	-0.024 [-0.04, -0.007]	-0.005 [-0.019, 0.009]	-0.019 [-0.036, -0.002]
Drift Rate (v)	0.177 [0.095, 0.259]	0.089 [0.018, 0.163]	0.088 [0.005, 0.173]

Supplementary Table 4.21. Mean difference in posterior distributions and 95% HDI Comparison in Experiment 6.

Estimate	'(Dis)Trust' minus Baseline	'(Dis)Trust' minus '(Dis)Like'	'(Dis)Like' minus Baseline
Distance between Decision Thresholds (α)	0.251 [0.059, 0.442]	-0.027 [-0.29, 0.227]	0.278 [0.048, 0.511]
Non-Decision Time (t0)	-0.163 [-0.491, 0.157]	0.343 [-0.059, 0.749]	-0.506 [-0. 859, -0.152]
Starting Point (z)	-0.025 [-0.041, -0.008]	-0.014 [-0.033, 0.005]	-0.011 [-0.029, 0.008]
Drift Rate (v)	0.182 [0.085, 0.28]	0.149 [0.058, 0.236]	0.033 [-0.068, 0.135]

7.4 Bibliography

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231-1233. https://doi.org/10.1038/s41562-023-01667-w

Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, 3(22), 470. https://doi.org/10.21105/joss.00470

Pallier, C. (2002). Computing discriminability and bias with the R software.