Understanding and supporting belief accuracy in a digital world

Understanding and Supporting Belief Accuracy in a Digital World

Jason W. Burton

Department of Psychological Sciences
Birkbeck, University of London

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

March 2, 2022

I, Jason W. Burton, 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 work.

The empirical research presented in Chapter 2 of this thesis is the result of a collaboration between myself, Adam J. L. Harris, Punit Shah, and Ulrike Hahn, which is now published in *Cognition* (Burton et al., 2022). The empirical research presented in Chapter 3 of this thesis is the result of a collaboration between myself, Nicole Cruz, and Ulrike Hahn, which is now published in *Nature Human Behaviour* (Burton, Cruz, et al., 2021). The empirical research presented in Chapter 4 of this thesis is the result of a collaboration between myself, Ulrike Hahn, Abdullah Almaatouq, and M. Amin Rahimian, which has been presented at the 9th ACM Collective Intelligence Conference (Burton, Hahn, et al., 2021) and the 43rd Annual Meeting of the Cognitive Science Society (Burton, Almaatouq, et al., 2021).

Acknowledgements

To my parents, Ron and Margie, whose unwavering love, support, and encouragement provided me with a platform to pursue my passions and find purpose, even far from home. To my partner, Sara, who rode with me through my anxieties and through a pandemic with heroic patience and compassion. And to my supervisor, Ulrike, whose advice and insights challenged me to think deeper and more critically than ever before.

Abstract

Advances in computing capacities have given rise to a "digital world" in which information can be accessed and shared at a faster pace, larger scale, and lower cost than what was previously possible. While this new digital world has promised a more informed public, research over the past decade has raised major concerns about the accuracy of people's beliefs, pointing to increasing polarisation, anti-intellectualism, and conspiratorial thinking. Efforts to understand why the promise of the digital world has not been realised often follow one of two perspectives. On one hand, psychological studies argue that humans process information irrationally to believe what they want to believe. On the other hand, studies of new digital media argue that structural features of the digital world present distorted information to users. In this thesis, I challenge these literatures by highlighting the limitations of widely-accepted research methods, and provide initial evidence that the same technologies denounced for undermining the integrity of our beliefs can be re-designed to promote accurate decision making. Using Herbert Simon's theory of bounded rationality as an organising framework, I present three studies examining (1) optimistic belief updating as a psychological account of belief inaccuracy "in the mind," (2) moral contagion as a structural account of belief inaccuracy "in the world," and (3) rewiring algorithms as a novel digital tool to support belief accuracy online. Theoretical, methodological, and practical implications are discussed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How do you know what you know to be true? Some things, like that fire is hot and lemons are sour, can be easily observed directly. But for many things, we depend on the testimony of others. Be it a belief about anthropogenic climate change, the existence of a god, or the hygiene standards of your favourite restaurant, much of what we "know" about the world is derived from evidence communicated to us by peers and media. In such scenarios the accuracy of your beliefs is determined not only by your ability — as an individual — to evaluate evidence communicated to you and update your priors, but also by the information environment you find yourself in. That is, our beliefs are shaped by both cognitive capacities "in the mind" and access to information "in the world."

Over the past two decades, our information environments have undergone fundamental changes at the hands of digitalisation. Backed by advances in computing capacities (Figure 1.1), the advent of social media and internet-based communications has allowed for information to be disseminated online at a faster pace, larger scale, and lower price than what is possible in the analog world (see, e.g., Hilbert & López, 2011; Holst, 2021; Rosa, 2013). On one hand, the new "digital world"[1] we find ourselves in promises a more engaged, more informed public. For instance, a primary concern of the past was information-scarcity, and the fact that information was monopolised by only a few media institutions that would "serve, and propagandise on behalf of, the powerful societal interests that control and finance them" (Herman & Chomsky, 2010, p. xi). In such a centralised information environment the beliefs of many are influenced by the testimony of few, which can lead, and has led, to pervasive and persistent inaccurate beliefs about empirical facts among the public (e.g., beliefs in phantom weapons of mass destruction

[1] The term "digital world" hereinafter refers to the societal setting in which the internet, social media, and other communication technologies are readily available. Loosely, from the year 2000 to the present-day.

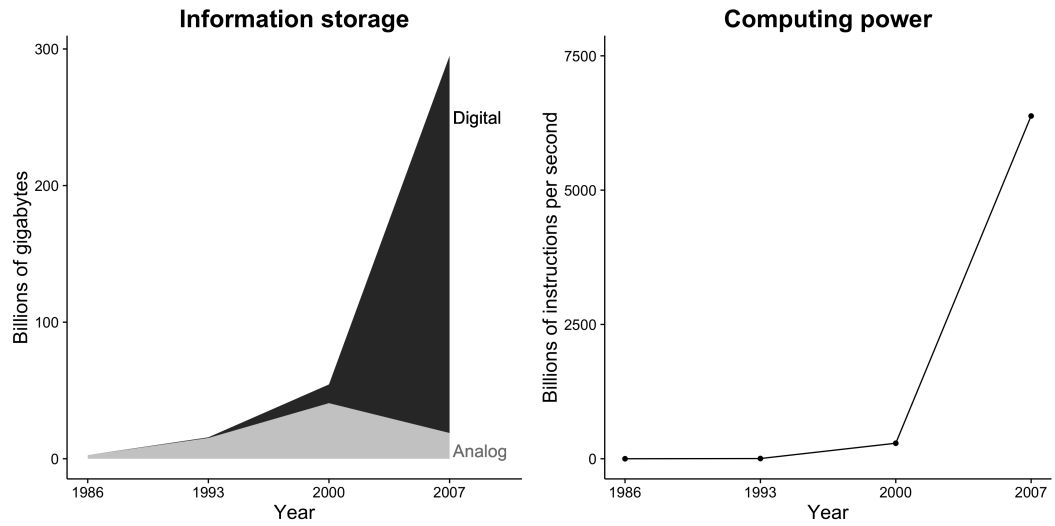

Figure 1.1: The increase of computing capacities from 1986 to 2007. Figure taken from Salganik (2017, p. 4); adapted from Hilbert and López (2011).

and the US invasion of Iraq; Lewandowsky et al., 2009). However, worries about a monolithic, gatekept information environment seem to be dissolved by new digital media and citizen journalism. People are no longer constrained to the information and viewpoints that media institutions deem newsworthy; the disadvantaged and the marginalised now have the ability to make their voices heard across the globe, instantaneously, so long as they have an internet connection. As of 2021, across all ages globally, just 25% say that the main way they access news online is directly via a news website, while 73%[2] say they do so via a "side door" — namely, social media (26%), a search engine (25%), mobile alerts (9%), a news aggregator (8%), or email (5%) (N. Newman et al., 2021, p. 25). By decentralising the information environment in this way, it would seem that the digital world grants people access to a higher quantity and wider diversity of viewpoints to be integrated into their beliefs — an observation that is supported by survey data showing that 87% of American internet users feel the web helps them to learn new things (Pew Research Center, 2014a).

On the other hand, the digital world introduces new epistemic challenges. For one, the information-richness online has been shown to have a darkside, because

> "...the wealth of information means a dearth of something else — a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention" (H. A. Simon, 1971, p. 41).

[2]This figure rises to 81% for under-35s, with 34% reporting social media to be their main access point for news online (N. Newman et al., 2021).

Indeed, it has long been demonstrated in psychology that we are limited in how much information we can attend to and process (Hills, 2019; Lanham, 2006; Lorenz-Spreen et al., 2020). For instance, limits on working memory capacity lead individuals to selectively attend to auditory stimuli in noisy environments in the so-called "cocktail party phenomenon" (Conway et al., 2001; Moray, 1959; N. Wood & Cowan, 1995), and limits on visual perception lead individuals to display an obliviousness towards salient but unexpected stimuli in "inattentional blindness" (Mack & Rock, 1998; Neisser & Becklen, 1975; Simons & Chabris, 1999; Simons & Levin, 1998). This psychological reality suggests that more information will not necessarily translate into a more informed public, as is further evidenced by a 2019 global survey finding that 32% of people actively avoid the news (a 3-point increase from 2017) with 28% of respondents in agreement "that there is too much news these days" (N. Newman et al., 2019, p. 26). Moreover, it has been pointed out that today's information-overload means that the *experience* of information has changed, vis-à-vis the intentionality of our information search and the salience of source characteristics (Seifert, 2017). That is to say, in the analog world "we did not have to be warned to 'consider the source' — the sources were distinct, intentionally accessed, and literally in our (face to) faces" (Seifert, 2017, p. 398); but in the digital world, information is experienced passively as it scrolls by and often looks the same regardless of the source's credibility. As a result, accurately evaluating the reliability of information encountered online may be particularly difficult.

Only a generation ago, it may have been considered common sense that increasing people's access to information would lead them to become more informed. Yet, in the contemporary context of digitalisation, this claim seems far more contentious. Is it cognitively realistic to expect people to effectively navigate and accurately evaluate information online? What characteristics of the digital world might affect people's ability to form accurate beliefs? How can people's belief accuracy be supported online? For the early promises of the digital world to be realised, it is necessary to develop a base of empirical evidence on such questions — and proper methods for doing so.

1.1 The emergence of the "post-truth" society

The impetus for understanding and supporting belief accuracy in the digital world is driven by recent socio-political trends that have spread across the globe. Events like the 2011 Tahrir Square protests, the 2017 #metoo movement, and the 2020 Black Lives Matter

marches for George Floyd demonstrate how decentralised information environments provide new means for social awareness and speaking truth to power (Tufekci, 2017). But at the same time, it has been argued that aspects of the digital world have brought about the "end of good faith" (Seifert, 2017, p. 398) and contributed to the emergence of a "post-truth" society "in which a large share of the populace is living in an epistemic space that has abandoned conventional criteria of evidence, internal consistency, and fact-seeking" (Lewandowsky et al., 2017, p. 360). Consider, for example, three conspicuous trends in the present-day that implicate the accuracy of our beliefs: polarisation, anti-intellectualism, and conspiratorial thinking.

Polarisation is defined as "a state in which the opinions, beliefs, or interests of a group or society no longer range along a continuum but become concentrated at opposing extremes" (Merriam-Webster, n.d.). For democratic societies, polarisation and the lack of common ground among social groups threatens the ability to collectively reason with and act on pressing issues (e.g., climate change, Dunlap et al., 2016; and COVID-19, Allcott et al., 2020). While the universality of the trend is subject to much debate (e.g., across countries, Boxell et al., 2021; or across sub-populations, Boxell et al., 2017), several studies have provided evidence suggesting that polarisation has been increasing in parts of the developed world, particularly in the United States (e.g., Abramowitz & Saunders, 2008; Abramowitz & Webster, 2016; Bishop, 2009; Iyengar & Westwood, 2015). For instance, in 2004, 70% of Republicans were more conservative on policy issues than the median Democrat and 68% of Democrats were more liberal than the median Republican, but by 2014, those figures were 92% and 94%, respectively (Pew Research Center, 2014b). And in an affective sense, the proportion of Democrats and Republicans holding "very unfavourable" opinions of the opposing party has more than doubled since 1994 (Pew Research Center, 2014b). Insofar as polarisation is increasing, a conventional narrative both in and outside of academia is that new digital media is to blame. This, it is argued, is because the decentralised nature of the digital world has led to fragmentation, such that the online information environment is divided into a variety of niches with very little interaction between them (Bright, 2018; Gentzkow, 2016; Lewandowsky et al., 2017). In turn, this fragmentation means individuals often find themselves in "echo chambers" where they encounter only information that reinforces their pre-existing beliefs, biases, and prejudices, both as the result of self-selection (homophily) and algorithmic personalisation (Pariser, 2017; Sunstein, 2018). This narrative has grown popular to the point that

United States President Barack Obama addressed it as a key threat to democracy in his 2017 farewell address, when he stated that

> "For too many of us, it's become safer to retreat into our own bubbles, whether in our neighborhoods or on college campuses, or places of worship, or especially our social media feeds, surrounded by people who look like us and share the same political outlook and never challenge our assumptions... the splintering of our media into a channel for every taste — all this makes this great sorting seem natural, even inevitable. And increasingly, we become so secure in our bubbles that we start accepting only information, whether it's true or not, that fits our opinions, instead of basing our opinions on the evidence that is out there" (Obama, 2017).

Yet despite the intuitive allure of the echo chamber hypothesis, its empirical merits are contested (see, e.g., Boxell et al., 2017; Dubois & Blank, 2018; R. K. Garrett, 2017; Gentzkow & Shapiro, 2011; A. Guess et al., 2018; Möller et al., 2018), and analyses of United States Congressional voting patterns suggest that increasing polarisation can be traced back the 1970s and 1980s (DeSilver, 2020), before the digital world as we know it had established itself.

Coinciding with increasing polarisation, there are also concerns about growing anti-intellectualism and anti-science attitudes (Hotez, 2020; Rutjens et al., 2018). Evidence of this trend can be noted in recent political messaging by elected officials in several countries: the United Kingdom's Justice Secretary, Michael Gove, infamously claimed in the run up to the 2016 EU Referendum that "the people of this country have had enough of experts" (Mance, 2016); Brazilian President Jair Bolsonaro responded to reports of environmental harm due to deforestation by suggesting people should "poop every other day" (BBC, 2019); Deputy Prime Minister of Italy, Mateo Salvini, publicly dismissed a mandatory vaccine policy for school children as "useless, and in many cases dangerous" (Politi, 2018); United States President Donald Trump explained that "One day — it's like a miracle — [COVID-19] will disappear... Nobody really knows" (Wolfe & Dale, 2020), only for his administration to later manipulate data to justify the relaxation of emergency restrictions (House Select Subcommittee on the Coronavirus Crisis, 2021). Public support for such figures has led researchers to search for correlates of anti-science attitudes (Hu et al., 2021; Lewandowsky et al., 2013; Rutjens et al., 2021) and for academic publications to uncharacteristically voice endorsements for a "pro-science" political candidate (Nature,

2020; Scientific American, 2020). Why has it come to this, and why now? As with trends in polarisation, blame has been directed at new digital media's effect on the information environment, and in particular, its distortion of expertise. In the digital world, publishing costs are eliminated and long-held journalistic reputations are discounted as any given individual can now "publish" their work alongside the likes of *BBC*, *Reuters*, or *The Associated Press* online. While this feature of the digital world is not without value (e.g., as a safeguard to information access and for dissenting voices to be heard), both the journalistic reputation and cost of a publication have long served as indicators of source quality (Seifert, 2017). Put simply, the decentralised nature of online information environments has allowed for anyone to become an "expert," and people's tendency to interpret vague, meaningless statements as profound suggests discourse online can be hijacked by influencers without any actual authoritative knowledge (Llewellyn, 2020; Pennycook et al., 2015; cf. *networked microcelebrity*, Tufekci, 2013). Nevertheless, evidence establishing a causal relationship between the digital world and anti-intellectualism has proven elusive, and the presence of such attitudes has long pre-dated the internet. As Woodrow Wilson stated in 1912: "What I fear is a government of experts. God forbid that, in a democratic country, we should resign the task and give the government over to experts" (as quoted in Fisher & Shapiro, 2020).

Amidst the trends of polarisation and anti-intellectualism, it is perhaps not surprising to note that beliefs in conspiracy theories have also become a concern. Notable examples of this are unfortunately in abundance, ranging from the seemingly innocuous (e.g., Finland doesn't exist; UFOs) to the bigoted and dangerous (e.g., QAnon; Holocaust denial). Although it is difficult to assess the absolute prevalence of conspiratorial beliefs given the variety, dynamism, and occasionally blurred boundaries of conspiracy theories, survey data from YouGov (2020) suggests that significant proportions of the global population hold such beliefs. For example, among more than 1,000 respondents per country, 30% of the Japanese believe it is either probably or definitely true that the truth about harmful effects of vaccines is deliberately hidden from the public, 46% of Spaniards believe it is either probably or definitely true that there is a single group of people who secretly control the world, 33% of Egyptians believe it is either probably or definitely true that the AIDS virus was created and deliberately spread by a secret group or organisation, and 20% of Americans believe it is either probably or definitely true that their own government assisted the 9/11 terrorist attacks (YouGov, 2020). Such beliefs, which have no

basis in fact, can have insidious effects on society when they are allowed to spread. As psychological studies have shown, exposure to conspiratorial discourse can reduce people's trust in authorities and willingness to participate in civic engagement (Einstein & Glick, 2015; Jolley & Douglas, 2014), even when people perceive the claims as implausible (Raab et al., 2013). This is particularly concerning given that beliefs in conspiracy theories have historically been stimulated in times of crisis (Van Prooijen & Douglas, 2017), when trust in authorities and civic engagement is most needed. A recent example of this occurred in the United Kingdom, where the announcement of the first national lockdown to curb COVID-19 in 2020 coincided with not only a steep increase in Google searches for the term "lockdown," but also for the term "5G" — the mobile network technology that was at the centre of a conspiracy theory that claimed it was linked to the spread of the virus (Figure 1.2). As there has been a flurry of disturbing conspiracy-related events in recent

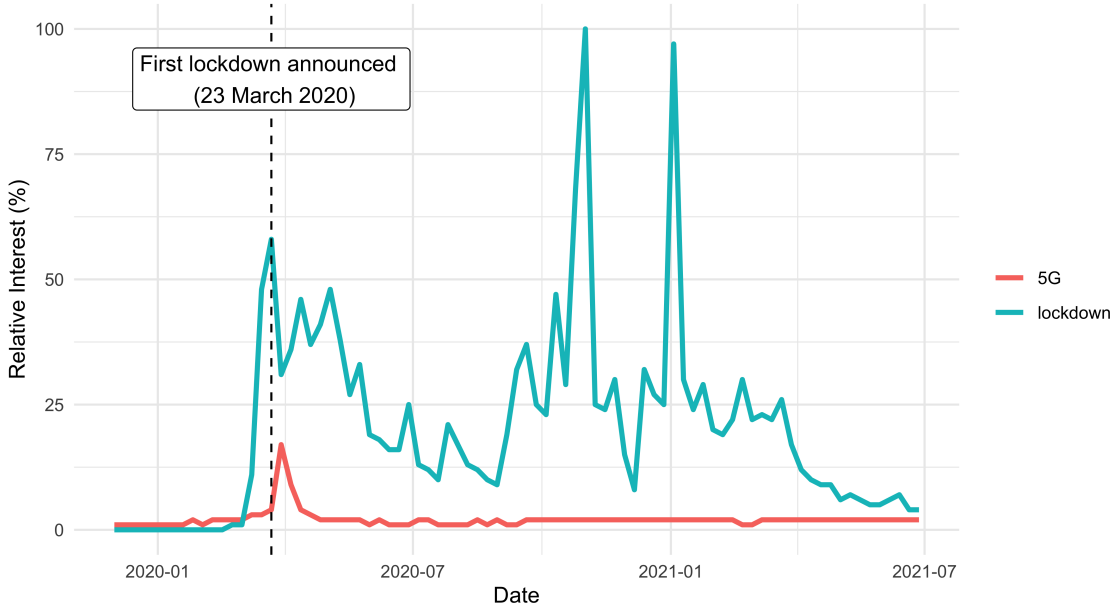

Figure 1.2: Google search trends for the terms "5G" and "lockdown" in the United Kingdom between December 2019 and June 2021, when the COVID-19 pandemic unfolded. Relative interest represents search interest relative to its highest point within the time range (i.e., a value of 100 is the peak popularity for the term between December 2019 and June 2021).

times (e.g., the storming of the United States Capitol in the name of *QAnon* and unsubstantiated claims of voter fraud in the 2020 Presidential Election), some have argued that the spread of conspiratorial thinking is being driven by new digital media. Here, it is said that the unbridled access to (fragmented) information and communication in the digital world allows individuals, who would otherwise be restricted to the fringes in the analog world, to coalesce in online forums and harden one another's inaccurate beliefs

(Klein et al., 2019; Lewandowsky et al., 2017). For instance, the growth of the *QAnon* conspiracy theory — a theory built on claims that a cabal of blood-drinking, pedophilic politicians and media personalities are working to take over the world — originated on the anonymous *4chan* forum in the "politically incorrect" (*/pol/*) imageboard where racist, misogynistic, and anti-Semitic tropes have been normalised through repeated sharing (Colley & Moore, 2020). Yet, despite the prevalence of conspiratorial beliefs today, empirical evidence suggests that conspiracy theories are not unique to the digital world and may not actually be on the rise (Van Prooijen & Douglas, 2017). In an analysis of more that 100,000 letters from United States citizens to the *New York Times* and *Chicago Tribune* between 1890 and 2010, for example, Uscinski and Parent (2014) found that conspiratorial content has varied over time but has not substantially increased. Thinking about the role of the internet in conspiratorial thinking, they explain that

> "there are plenty of sites for people to find information supporting conspiracy theories, but people not predisposed to believing in conspiracy theories will likely not seek them out. Technology could just as easily decrease conspiracy theorizing because it increases access to anti-conspiratorial information... The data show that technology is unrelated to the level of conspiracy theorizing" (Uscinski & Parent, 2014, p. 122).

Indeed, further support for this conclusion is the observation that conspiracy theories have punctuated much of human history, from speculation that Emperor Nero deliberately had Rome burnt down whilst singing in AD 64, to the Salem Witch Trials in 1692-1693, to the "Red Scare" of the 1940s and 1950s (Van Prooijen & Douglas, 2017).

As terms like *echo chamber*, *disinformation*, and *post-truth* have seeped into our everyday vocabulary, it is difficult not to associate the evolution of the digital world with polarisation, anti-intellectualism, conspiratorial thinking, and threats to belief accuracy at large. But, in the absence of a counterfactual world in which the introduction of new digital media does not coincide with the emergence of the so-called post-truth society, the empirical evidence of causal connections between the two remains tenuous.

1.2 Bounded rationality as an organising framework

Scholarly efforts to reconcile the emergence of the post-truth society with the evolution of the digital world have largely followed one of two streams. Along the first stream are

studies that point to psychological problems "in the mind" (reviewed further in Chapter 2). The basic premise here is that people are irrational, motivated reasoners whose self-enhancing biases and motivational deficiencies lead them to "believe what they want to believe" (Kunda, 1990, p. 480). Thus, in the context of the digital world, it is presumed that users of new digital media will naturally navigate and examine information online in a biased manner: they will search for information that confirms their prior beliefs (e.g., Wason, 1960, 1968), interpret information in a way that affirms their worldview (e.g., Kahan, 2013; Lord et al., 1979; Van Bavel & Pereira, 2018), and update their beliefs more in response to desirable information than undesirable information (e.g., Sharot et al., 2011; Sunstein et al., 2016). Altogether, this suggests that the potential benefits of the digital world's accessible, decentralised information are seemingly undermined by human cognition, resulting in people forming inaccurate beliefs despite accurate information available to them.

Alternatively, a second stream of research focuses on structural problems "in the (digital) world" (reviewed further in Chapter 3). Here, it is argued that proprietary design features of online information environments make it so that users are unable to reasonably navigate information and form accurate beliefs. For instance, studies have shown that the decentralised topology of online social networks provides individuals with inaccurate cues of consensus on important issues (e.g., Lerman et al., 2016; A. J. Stewart et al., 2019), the allowance of anonymous, automated accounts (bots) enables bad actors to amplify misinformation (e.g., Ferrara et al., 2016; Shao et al., 2018), opaque content-curating algorithms manipulate what information users encounter (e.g., Lazer, 2015), and perverse incentives reward the sharing of low-quality content (e.g., Brady et al., 2017; Vosoughi et al., 2018). From this perspective, the potential benefits afforded by the digital world are seemingly undermined not by the users, but by new digital media itself, which results in rational people forming inaccurate beliefs because of distorted information exposed to them.

A reader of these literatures seeking to understand why inaccurate beliefs might arise in the digital world often finds themself caught in a false dichotomy of *is it the technology or is it the user*. But taking one perspective or the other means missing crucial parts of the substantive picture. Relatedly, narrowing oneself to either perspective alone also means missing crucial methodological opportunities. Where studies of psychological shortcomings in the mind have typically valued explanation but lacked ecological validity and statistical

power, computational studies of online information environments have typically valued prediction but lacked the types of experimental controls needed for causal inference and generalisability (Hofman et al., 2017). In order to better understand threats to belief accuracy in the digital world, and how to overcome them, an appropriate conceptual framework is needed to organise and integrate research findings. Here, H. A. Simon's theory of bounded rationality provides a helpful lens.

Posited as an alternative to the neoclassical model of the utility maximising "economic man" upheld by rational choice theory, bounded rationality refers to a model of human behaviour and decision making that takes into account limitations imposed by both cognitive capacities and access to information in the environment (H. A. Simon, 1955, 1957). At its crux, bounded rationality dispels with the idea that we follow some general-purpose axioms of logic or optimisation. Rather, as boundedly rational agents, we are adaptive — constantly managing trade-offs and changing our utility functions depending on the environment in which we find ourselves (H. A. Simon, 1957, 2000; also referred to as "ecological rationality," Gigerenzer & Selten, 2002; Todd & Gigerenzer, 2007; Wheeler, 2018). Through this conceptual framework, the false dichotomisation of the user versus the technology is substituted for a more integrated perspective where one account does not explain away the other. Put simply, seeking to understand and support belief accuracy in the digital world through the lens of bounded rationality means taking into account the "bounds" imposed by both psychological limitations "in the mind" of the user and the structural characteristics of information environments "in the (digital) world," and most importantly, identifying interactions between the two.

1.3 Thesis structure

In this thesis, I adopt the framework of bounded rationality and take aim at the substantive issue of belief accuracy in the context of a digital world. Specifically, I consider (1) a psychological mechanism that may jeopardise our ability to accurately utilise online information access, (2) a structural feature of online environments that may distort the information to which we are exposed, and (3) how such environments might be (re)designed to actively support belief accuracy. In doing so, I take stock of methodological approaches that presently shape the discourse on this topic, making special note of various pitfalls that may undercut our knowledge.

In Chapter 2, I present an experimental study probing a prevailing account of belief

inaccuracy "in the mind." Namely, this entails examining the widely cited mechanism of optimistic belief updating, whereby people are said to revise their beliefs more in response to desirable information than undesirable information as the result of motivational bias (Sharot et al., 2011). By empirically demonstrating flaws in the conventional method used to support the optimistic belief updating account, I argue that there is insufficient evidence to conclude a motivational bias in belief updating exists. More broadly, I leverage the findings to speak to the difficulties of identifying bias — as a systematic deviation from accuracy — with artificial laboratory-style experimentation; thereby de-emphasising the narrative that the post-truth society has been brought about by individuals' motivated reasoning alone.

In Chapter 3, I present an observational study probing a high-profile account of belief inaccuracy "in the (digital) world." Specifically, I narrow in on the recent finding of "moral contagion" — an effect whereby the use of moral-emotional words (e.g., *kill*, *shame*, *compassion*) in messages increases their diffusion in online social networks, suggesting that social media platforms favour outrageous content over informational content (Brady et al., 2017). Through out-of-sample prediction tests, model comparisons, and specification curve analyses, I find the moral contagion effect to be inconsistent, hinge on arbitrary analytical decisions, and perform no better than an implausible XYZ contagion model (where it is hypothesised that the presence of the letters X, Y, and Z increases messages' diffusion). Based on these findings, I argue that the evidence of moral contagion may be little more than a spurious correlation, and further, that conventional analytic techniques for studying information diffusion online can support patently absurd claims, despite utilising large, naturalistic datasets. As such, this chapter de-emphasises the narrative that the post-truth society has been brought about by maliciously designed digital platforms alone.

In Chapter 4, I shift focus to explore how belief accuracy can be supported by engineering digital tools that dynamically manipulate online environments for *mind-world* interaction. Drawing from the literature on "wisdom of the crowd" effects and collective intelligence design, I show, through agent-based modelling and an online multiplayer experiment, that algorithms can be used to mediate communication between users so that they arrive at more accurate collective beliefs. Although exploratory, results presented in this chapter provide a proof of concept, showing that the same features of the digital world that have been decried as epistemically dangerous (in this case, content-curating

algorithms) can be used to support belief accuracy online, pointing towards new paths for a digital, democratic future.

Finally, Chapter 5 serves to integrate the findings presented in Chapters 2-4 and highlight their implications for the study and support of belief accuracy in the digital world, incorporating theoretical, methodological, and applied points of view.

Chapter 2

A psychological problem "in the mind"

> Because of the psychological limits of the organism, actual human rationality-striving can at best be an extremely crude and simplified approximation to the kind of global rationality that is implied
>
> H. A. Simon, 1955, p. 101

For people's belief accuracy to benefit from the increased information access afforded by the digital world it is assumed that they can effectively navigate, reason with, and respond to the information they encounter. However, several longstanding studies in psychology have pointed out that people's information processing may be systematically biased, such that their ability to form accurate beliefs is jeopardised by psychological mechanisms "in the mind" (e.g., Adams, 1961; Festinger, 1957; Lord et al., 1979; Macdougall, 1906; Wason, 1960; Weinstein, 1980). Underlying these works and their contemporary advocates (e.g., Kahan, 2013, 2016; Sharot et al., 2011; Van Bavel & Pereira, 2018) is the basic premise that people's motivation to mitigate cognitive dissonance can overpower their motivation for accuracy (Festinger, 1957; Kunda, 1990). As a result, people navigating information tend to display a *confirmation bias* (Nickerson, 1998; cf. *motivated reasoning,* Kunda, 1990; *my-side bias,* Baron, 1995; *congeniality bias,* Hart et al., 2009; *belief bias,* Klauer et al., 2000), whereby information that is concordant with pre-existing beliefs is favoured over information that is discordant. While evidence for such self-enhancing bias

appears under various guises and terminologies throughout the literature, understanding how inaccurate beliefs can arise due to psychological mechanisms requires disentangling three distinct processes involved: (1) information sampling, (2) evaluating or reasoning with information, and (3) belief updating.

The first process, information sampling, refers the the cognitive-behavioural task of choosing which information to attend to and process in the environment. For an individual seeking to form an accurate belief on a given hypothesis, the normative expectation is that their information sampling is guided by the intention to maximise the usefulness of each query they make according to an information-theoretic utility function, such as Kullback-Leibler divergence (Kullback & Leibler, 1951), information gain (Lindley, 1956), or Bayesian diagnosticity (Good, 1950). Descriptively, however, people's information sampling strategies have been shown to deviate from such ideals. In a classic study commonly cited as original evidence for confirmation bias (despite the terminology not appearing in the actual text), Wason (1960) challenged participants to infer the rule used to generate a sequence of three numbers, 2-4-6 (the correct answer being "increasing in magnitude"), by submitting their own three numbers for the experimenter to identify as conforming to the rule or not. What he found was that a proportion of participants only or primarily sought out positive evidence to confirm whatever rule they believed to be the answer, rather than efficiently seeking to falsify their hypotheses (Wason, 1960; but see Oaksford & Chater, 1994). Reflecting on the 1960 study, Wason (1968) explained that the experiment demonstrates, "on a miniature scale, how dogmatic thinking and the refusal to entertain the possibility of alternatives can easily result in error" (p. 174). Indeed, the confirmatory approach to information sampling identified by Wason (1960, 1968) (cf. *positive test strategy*, Klayman & Ha, 1987) has important implications for belief accuracy, and perhaps particularly so in the context of an information-rich, digital world. Since the massive volumes of information online make it infeasible for any individual to process all information relevant to seemingly any given topic at hand, the ability to efficiently sample information becomes highly consequential. Yet, as is evidenced by recent studies of *selective exposure* and news consumption, people often follow strategies similar to that identified by Wason (1960, 1968) and sample information that conforms to their prior beliefs (e.g., Bolin & Hamilton, 2018; Hart et al., 2009; T. P. Newman et al., 2018; Pariser, 2017; Schmidt et al., 2017; Seargeant & Tagg, 2019; Stroud, 2008).

The second process involved in belief formation is reasoning, or, how one evaluates

information encountered. Whereas biased sampling of information can lead individuals to develop discrepant beliefs based on the skewed bodies of evidence to which they each attend, individuals who engage in biased, motivated reasoning may develop discrepant beliefs even when they share the same evidence. A famous example of this comes in a seminal study of so-called *biased assimilation*, where Lord et al. (1979) presented proponents and opponents of capital punishment with an array of mixed evidence on the topic and found that participants tended to accept evidence supporting their initial belief at face value, whilst scrutinising counter evidence hypercritically. Consequently, Lord et al. (1979) observed participants' beliefs become more polarised after being provided with information, despite the information being neutral on the whole and posing as what would intuitively be understood as "common ground" (also see Corner et al., 2012; Dandekar et al., 2013). Along these same lines, developments in the subfield of political psychology have seen the production of several studies showing that people's reasoning may be distorted by motivation to protect their social identities in the face of policy-relevant facts — referred to as *cultural cognition* and the *politically motivated reasoning paradigm* by Kahan and colleagues. These include, for example, experiments where participants are seen to arrive at identity-consistent interpretations of statistical data on crime (Kahan et al., 2017), video footage of protesters (Kahan, Hoffman, et al., 2012), climate change evidence (Corner et al., 2012; Kahan, Peters, Wittlin, et al., 2012), arguments on the risks/benefits of nanotechnology (Kahan et al., 2009), and economic time series data (Caddick & Rottman, 2019), arguably because the utility of maintaining one's identity as favourable to their ingroup outweighs the utility of being accurate (Kahan, 2016; Van Bavel & Pereira, 2018). With respect to the promises of a digital world, the implication of biased reasoning is self-evident: if people rationalise their way around inconvenient facts then increasing their access to information will not necessarily lead them towards the truth.

The third process is belief updating, which refers to people's ability to effectively revise their beliefs upon the receipt of new information. Here, there are two well-documented phenomena that figure into the discussion of general confirmation bias and motivated reasoning: *conservatism* and *optimistic belief updating*. With regard to conservatism, the basic result in the large, longstanding literature is that when people encounter new information they tend to revise their initial belief less than is prescribed by the normative standard of Bayes' theorem (e.g., Edwards, 1968; Peterson & Miller, 1965; Phillips et al., 1966; Phillips & Edwards, 1966). The paradigmatic example proceeds by challenging

participants to guess the probability that there is one of two compositions of coloured chips in a bag (e.g., does your bag contain 70% red and 30% blue chips, or 30% red and 70% blue chips?) after seeing a number of chips drawn. Although participants update their guess in the correct direction from the starting point of complete uncertainty, 50% either way, the magnitude of their updating tends to be insufficient given the evidence provided (Edwards, 1968). While conservatism undoubtedly inhibits effective belief updating, it alone seems an inadequate explanation for how inaccurate beliefs arise in the context of a digital world, where it could be presumed that the abundance of information online would, eventually, lead people towards the truth. On top of conservatism, however, more recent studies have identified the phenomenon of optimistic belief updating, which is of special interest to this thesis. Building on early demonstrations of unrealistic comparative optimism by Weinstein (1980) — showing that people rate their own chances of experiencing positive life events to be greater than average and their chances of experiencing negative life events to be lower than average — a substantial body of multidisciplinary research has observed that people tend to revise their beliefs more (i.e., less conservative, but still conservative) in response to desirable information than undesirable information (e.g., Chowdhury et al., 2014; Eil & Rao, 2011; Marks & Baines, 2017; Möbius et al., 2014; Moutsiana et al., 2013; Sharot et al., 2011). This asymmetry, where "good news" is more readily integrated into individuals' beliefs than "bad news," has rightly been pointed out as cause for concern in high-stakes domains such as lay people's views on anthropogenic climate change, for which the discounting of information that is discordant with one's beliefs may have dire consequences (Sunstein et al., 2016). With respect to the broader context of a digital world, the implication of optimistic belief updating is that people faced with swathes of information online may find themselves with increasingly skewed beliefs, because they are more influenced by encounters with desirable information than undesirable information.

As outlined above, the beliefs people form may be determined by psychological biases when sampling information, evaluating information, and updating their beliefs. However, the central issue at hand for this thesis is belief *accuracy*. That is to say, the only "bias" that is of interest is bias that incurs systematic accuracy costs (for a historical overview of how bias has been defined in research, see U. Hahn & Harris, 2014). Defining bias in this way quickly reframes much of the foundational evidence for a general confirmation bias and motivated reasoning. For example, despite Wason's (1960) participants predominantly sampling information with confirmatory tests instead of falsifying hypotheses, 21

of the total 29 were able to correctly identify the rule governing his number sequence. Yet unfortunately, it is not always possible to straightforwardly assess bias as systematic deviation from accuracy (i.e., are descriptive processes incurring accuracy costs *on average*?). While this is obvious for certain topics, such as matters of subjective preference, it also maligns many seminal paradigms in psychology. In the classic evidence of biased assimilation, for instance, Lord et al. (1979) themselves point out that

> "there can be no real quarrel with a willingness to infer that studies supporting one's theory-based expectations are more probative than, or methodologically superior to, studies that contradict one's expectations... The same bias leads most of us to be skeptical about reports of miraculous virgin births or herbal cures for cancer, and despite the risk that such theory-based and experience-based skepticism may render us unable to recognize a miraculous event when it occurs, overall we are surely well served by our bias" (p. 2106).

Similarly, the original proponent of unrealistic optimism has clarified how,

> "a woman who says that her risk of heart disease is only 20%... may be perfectly correct when her family history, diet, exercise, and cholesterol level are taken into consideration, despite the fact that the risk for women in general is much higher" (Weinstein & Klein, 1996, p. 2).

In experimental set-ups like these, the true value against which accuracy could be measured is unknown to both the participants and the experimenters. For Lord et al. (1979) this would require knowledge of the precise net societal effect of capital punishment; for Weinstein (1980) this would require knowing each individual participant's precise personal risk of experiencing life events. Where such controlled experimentation, which is indeed a cornerstone of modern psychology, can enable causal inferences about the biasing role of prior beliefs or social identity in information processing, it lacks ecological validity vis-à-vis real-world consequences on belief accuracy (e.g., Jarvstad et al., 2013). How then can bias, as systematic deviation from accuracy, be operationalised? Here, normative models that define "optimal" reasoning and decision-making provide a valuable tool. Aside from the referenced studies pertaining to conservatism in belief updating, which typically root their analyses to comparisons between participants' actual behaviour and Bayesian prescriptions, neglect of normative models renders many experimental results inapplicable or uninterpretable with respect to real-world accuracy costs. Of course, a reader familiar

with the work of Herbert Simon might point out that this emphasis on optimal, normative models is at odds with the theoretical framework of bounded rationality adopted by this thesis. However, while it may be that deviations from normative models indicate adaptive cognition (H. A. Simon, 1956, 2000), it is first necessary to establish whether or not individuals deviate from such models to begin with.

In the remainder of this chapter I present an empirical investigation of the optimistic belief updating phenomenon so as to demonstrate the challenges of identifying true bias, as a systematic deviation from accuracy. Following this, I conclude the chapter by linking it back to our overarching objective of understanding belief accuracy in a digital world.[1]

2.1 Asymmetric belief updating in response to neutral stimuli

Over the past decade, research has argued that people are optimistically biased when updating their beliefs in light of new information, such that desirable information elicits greater updates than undesirable information. Given the grim implications of such motivational distortion for the accuracy of our beliefs, the phenomenon has attracted considerable cognitive and neuroscientific interest (Chowdhury et al., 2014; N. Garrett et al., 2018; N. Garrett et al., 2014; Kappes et al., 2018; Kuzmanovic et al., 2016; Kuzmanovic & Rigoux, 2017; Kuzmanovic et al., 2018; Marks & Baines, 2017; Moutsiana et al., 2013; Moutsiana et al., 2015; Sharot, 2011; Sharot, Guitart-Masip, et al., 2012; Sharot, Kanai, et al., 2012; Sharot et al., 2011; Sharot et al., 2007). However, recent research has highlighted limitations of existing results and demonstrated that what appears to be optimistically asymmetric belief updating may in fact be attributable to a statistical artefact arising from the experimental design, rather than a self-enhancing motivational bias (Harris et al., 2013; Shah et al., 2016; but see N. Garrett & Sharot, 2017, for rebuttal). In this research, we further investigate this "statistical artefact hypothesis" with three preregistered experiments. All three demonstrate asymmetric belief updating in response to neutral information in our main analysis, and we further observe uninterpretable variability across samples and across various analytic techniques. Given there is no desirability in these trials, this asymmetry cannot be attributed to optimism.

[1] The following studies are based on a collaboration between myself, Adam J. L. Harris, Punit Shah, and Ulrike Hahn, which is now published in *Cognition* (Burton et al., 2022). The relevant preregistration and all data and code has been made available on an OSF project page.

The update method

Evidence for optimistic belief updating has primarily been obtained from "the update method" (Sharot et al., 2011). The simplest instantiation of the method proceeds as follows: Participants first estimate their chance of experiencing a negative life event (E1), then they are provided with the base rate statistic for experiencing that event (BR), and finally, they are asked to re-estimate their chance of experiencing said event (E2). For example, a participant, Sam, might be asked to estimate their chance of experiencing a negative life event, like getting divorced, to which they reply 5%. Sam is then presented with the BR of divorce (the actual proportion of the general population that gets divorced in their lifetime), which is 45%. Since the BR in this instance is greater than Sam's E1 for this negative event, this belief updating trial would be classed as one with undesirable information. Trials on which participants receive desirable information typically elicit greater updates than trials with undesirable information, which is interpreted as evidence of optimism in belief updating (e.g., Sharot et al., 2011).

The statistical artefact hypothesis

While the update method has been accepted as the basis for a range of high-profile cognitive and neuroscientific work (e.g., N. Garrett et al., 2018; Moutsiana et al., 2015; Sharot, Guitart-Masip, et al., 2012; Sharot et al., 2011; Sharot et al., 2007), it has also been subjected to critique. As the update method requires people to update a probabilistic belief, the appropriate normative standard against which to evaluate behaviour so as to identify bias is Bayes' theorem (Hardman, 2009; Kahneman & Tversky, 1973; Phillips et al., 1966; Phillips & Edwards, 1966). Continuing our example, Sam is asked to report their risk of divorce. Normatively, this risk is comprised of two types of information: knowledge about the base rate of divorce and any individuating information Sam has to differentiate their risk from the average person's. For instance, Sam might intend never to marry in the first place, which suggests that they should differentiate themself from the average person about marriage-related base rates (and this has long been understood to be the central challenge in assessing the accuracy of people's personal risk estimates for future life events, e.g., Weinstein, 1980). These two types of information — the base rate and individuating information — combine multiplicatively in Bayes' theorem, yet the update method neglects the influence of individuating information on normatively rational belief updating (Shah et al., 2016).

One way in which individuating information poses a difficulty for the update method is in the classification of trials with desirable versus undesirable information. For example, the primary component of Sam's belief may be their intention to never marry (whilst recognising this may change later in life, and, as such, is greater than zero), but they estimate the base rate of divorce in the general population is 50%, combining to a personal risk estimate of 5%. Upon learning the BR (45%) they should, in fact, slightly decrease their estimate of their own risk, not increase it as the classification of the desirability of BR (based on its comparison with E1) would suggest.

Even more worryingly, basic aspects of the probability scale — namely, that it is multiplicative and bounded — mean that over- and under-estimates relative to a given BR should not give rise to equal amounts of belief change once individuating knowledge comes into play. As Shah et al. (2016) demonstrate, it is mathematically impossible to equate the amount of BR error (the difference between one's estimate of the BR and the actual BR), the amount of individuating knowledge, and the normatively necessary degree of belief change across desirable and undesirable trials on the bounded, zero to 100 probability scale. Matching any two of these three will necessarily give rise to a divergence on the third (displayed later via a numerical simulation; Figure 2.8).

As a result of both the issues above, entirely rational Bayesian agents will display seemingly optimistic belief updating (Shah et al., 2016). While Kuzmanovic and Rigoux (2017) have recently sought to remedy these issues (see "further analyses" for critique), the Shah et al. (2016) critique suggests that the vast majority of evidence for optimistic belief updating may be nothing more than a statistical artefact.

Why (neutral) valence matters

Valence — whether information is good, bad, or neutral — is essential to the claims supporting the existence of optimistic belief updating. As Sharot and Garrett (2016) explain, optimistic belief updating is "a valence-dependent asymmetry in how people use favourable and unfavourable information" (p. 31). That is, optimistic belief updating is motivated by the perceived valence of encountered information. Therefore, establishing a causal relationship between asymmetries in belief updating (such as those recorded with the update method) and information valence is necessary to distinguish true, unrealistically optimistic belief updating from non-motivational, confounding effects.

Typically, belief updating studies use negative life events when asking participants for

probability estimates with the update method (e.g., how likely is it that you will contract liver disease?). Negative life events are convenient because their perceived valence is considered to be universal, and one can readily gather accurate base rates from sources like the Office for National Statistics to use in experimental materials (e.g., Sharot et al., 2011). Given the statistical confounds associated with the update method, however, Shah et al. (2016) argued for the need to include both positive and negative life events in studies using the update method to evaluate directional, valence-driven effects (as in N. Garrett & Sharot, 2017; Marks & Baines, 2017). The present work goes further by including neutral events as our focal stimuli. With negative events, evidence for optimistic updating is claimed where participants update their beliefs more in trials where BR is less than E1 than where BR is greater than E1. If such an asymmetry can be observed with neutral events, this cannot be attributed to optimism, since there is no valence from which to ascribe (un)desirability. Thus, by parsimony, one should not appeal to a motivational explanation to explain such a pattern in situations where desirability is present.

In this research, we present evidence for a non-motivational account of asymmetric belief updating by eliciting the statistical artefact with neutral life events and replicating it in three variations of the update method, and displaying its variability across samples and alternative analytic techniques. As inherently non-valenced, an asymmetry in belief updating with neutral events suggests that what has been previously interpreted as evidence of optimistic belief updating is the result of statistical patterns rooted in methodological constraints, rather than a motivational bias. That this artefactual asymmetry varies so unpredictably further undercuts the interpretability of past and future results returned by the update method.

2.1.1 Method

Since studies utilising the update method typically involve more than one stimulus (life event), there are variations in the way this experiment can be and has been run: Participants can provide E2s immediately after receiving the BR for each event (Kuzmanovic et al., 2016; Marks & Baines, 2017), or they can provide all these at the end, after having provided E1s and received BRs for all events in the study (N. Garrett & Sharot, 2014, 2017). Such procedural changes have been made in the past without justification and it is unknown how they affect the results of the update task. We subsequently replicated our experiment in three ways for robustness (Figure 2.1). While no predictions of different

results between these studies were made, this project represents a direct replication of the same methodology using these three slightly different procedures. All data otherwise followed the same scoring plan.

Participants

One hundred participants were recruited for each study ($N = 300$). This sample size was deemed sufficient by a power analysis conducted with a tool for calculating statistical power for mixed effects models provided by Judd et al. (2017). However, due to the quasi-experimental nature of this study, and in fact any study using the update method, the validity of such a power analysis is limited. Since we do not know how many life events participants will rate as neutral, we cannot know how many stimuli will be included in the main analysis, which focuses specifically on neutral life events. While we made a conscious effort to compile seemingly neutral life events, it is also plausible that some participant(s) will not rate any life events as neutral, and thereby be excluded from the main analysis. With these unknowns in mind, we arrived at the sample size of 100 because a hypothetical power analysis with 51 stimuli (the number of events used), an effect d of 0.5, and 100 participants returns a high power of 0.84. This is also roughly double the sample size used in Marks and Baines (2017), which is the only other optimistic belief updating study that uses the same main analysis (a linear mixed effects model).

The Prolific Academic online research platform was used for recruitment, and the participant pool location was restricted to the United Kingdom because the stimuli (the life events and base rate statistics) were compiled to suit this specific population, and certain events and base rates may not be relevant to participants living elsewhere. The sample for each study were independent of one another. In Study 1, participants' ages ranged from 18 to 81 ($M = 32.82$, $SD = 11.52$) with 73 females; in Study 2, ages ranged from 18 to 67 ($M = 33.29$, $SD = 10.31$) with 78 females; and in Study 3, ages ranged from 19 to 73 ($M = 36.09$, $SD = 12.06$) with 80 females.

Design

A 3 (event valence: negative, neutral, positive) x 2 (direction of error: upwards, downwards), quasi-experimental within-subjects design was implemented with the Qualtrics web-based survey software. Event valence (negative, neutral, or positive) signifies the self-reported desirability of experiencing a given life event. While the set of life events used

(Table A.1) was intended to be comprised of neutrally-valenced events, event valence was coded trial-by-trial with participants' self-reports, meaning that an event had to have been rated as "neither positive nor negative" (three on a five-point Likert scale) to be classed as a neutral trial for a given participant. Direction of estimation error (henceforth referred to as "direction of error") indicates whether a participant's initial self-estimate was less than or greater than the presented base rate on a given trial (whilst previous studies would refer to these trials as desirable vs. undesirable, such a categorisation cannot be made for neutral events). Our dependent variable — the magnitude of belief updating — is the absolute difference between E1 and E2, signed as positive if the update is in the predicted direction (according to the direction of error), or negative if it is in the opposite direction. See the subsection on "measures" below for further definition of these variables.

Stimuli

A total of 51 life events, each with an accompanying base rate statistic ($M_{BR} = 38.39$, $SD_{BR} = 21.58$), was presented to each participant (Table A.1). This set of life events is comprised of both previously used and novel material. Two life events were taken from Shah et al.'s (2016) materials that were deemed to be neutral at face value ("Be exactly the same weight in 10 years' time" and "Last the whole of winter without catching a minor cold"). Twenty-one life events were taken from the materials used by N. Garrett and Sharot (2017), which were found to be rated as neutral by pilot participants [3.00 $\pm$ 0.99 on a 1 (extremely negative) to 5 (extremely positive) scale]. Finally, 28 novel life events were compiled by gathering or calculating base rate estimates based on external sources (e.g., Ofcom, BBC) that were associated with seemingly regular, mundane life events [e.g., "How likely is it that the next store you visit is air conditioned?" ($BR = 30\%$); "...that you use more than 3.7GB of mobile data over the next four weeks?" ($BR = 17\%$)].

Procedures

Study 1. Study 1 involved two parts (Figure 2.1). In Part 1, each participant was presented with a life event and asked to estimate the likelihood of this event happening to them in the future ($E1$); participants were also asked to give a second estimate of the likelihood of the event happening to an average person (eBR; an estimate of the base rate). These two estimates were made on a 0% to 100% scale and there were no restrictions on participants' response time. The order of these two estimates was counterbalanced between

subjects by randomly assigning each participant to one of two conditions: E1 followed by eBR, or eBR followed by E1. After these two initial estimates were recorded, participants were presented with the actual base rate statistic (BR). On the next page they were required to write down the BR correctly ($BR\ Recall$). If they incorrectly recalled the statistic, they were provided with the correct statistic on screen and required to enter that value. This ensured that the correct BR was attended to by all participants. Participants were then asked to re-estimate the likelihood of the event happening to them personally ($E2$). Once this sequence was completed for all 51 life events, in a randomised order, participants moved on to Part 2.

In Part 2, they were presented with each of the 51 life events again in a randomised order and asked to indicate the valence of each event, "How would you feel about experiencing this event?" on a five-point Likert scale (1 = extremely negative, 2 = somewhat negative, 3 = neither positive nor negative, 4 = somewhat positive, 5 = extremely positive) ($Valence\ Rating$). This rating would later be used to classify the trials as negatively, neutrally, or positively valenced according to each individual participant's subjective rating. Trials rated as 3 were considered neutral.

Study 2. Study 2 involved three parts (Figure 2.1). The procedure followed that of Study 1, except the E2s for all events were recorded in Part 2, which was followed by the recording of all valence ratings in Part 3.

Study 3. Study 3 involved two parts (Figure 2.1) and mirrored the procedure of Study 1, except the ordering of E2 and Valence Rating were swapped such that the updating occurred across the two parts (i.e., E1 in part one, E2 in Part 2). This procedure follows that used in Experiments 3A and 3B in Shah et al. (2016).

Exclusion criteria

Following on from Shah et al. (2016), a preregistered exclusion criterion was employed prior to analysing the data from each study. Mean updates in each of the six conditions (i.e., negative/neutral/positive by upwards/downwards) were calculated and outliers were removed ($\pm 3 \times$ the interquartile range)[2]. However, all of our results hold regardless of this

[2]We note that on page 7 of the preregistration this exclusion criterion is written as "±3 the interquartile range." This discrepancy is due to a clerical error whereby the multiplication sign was inadvertently removed when pasting the text into the OSF registration form and converting the Word file to PDF format. In addition, the mention of removing trials "in which a derived probability cannot be applied" is also erroneous as it is not applicable to the methodology presently used because we provided static, non-derived base rates.

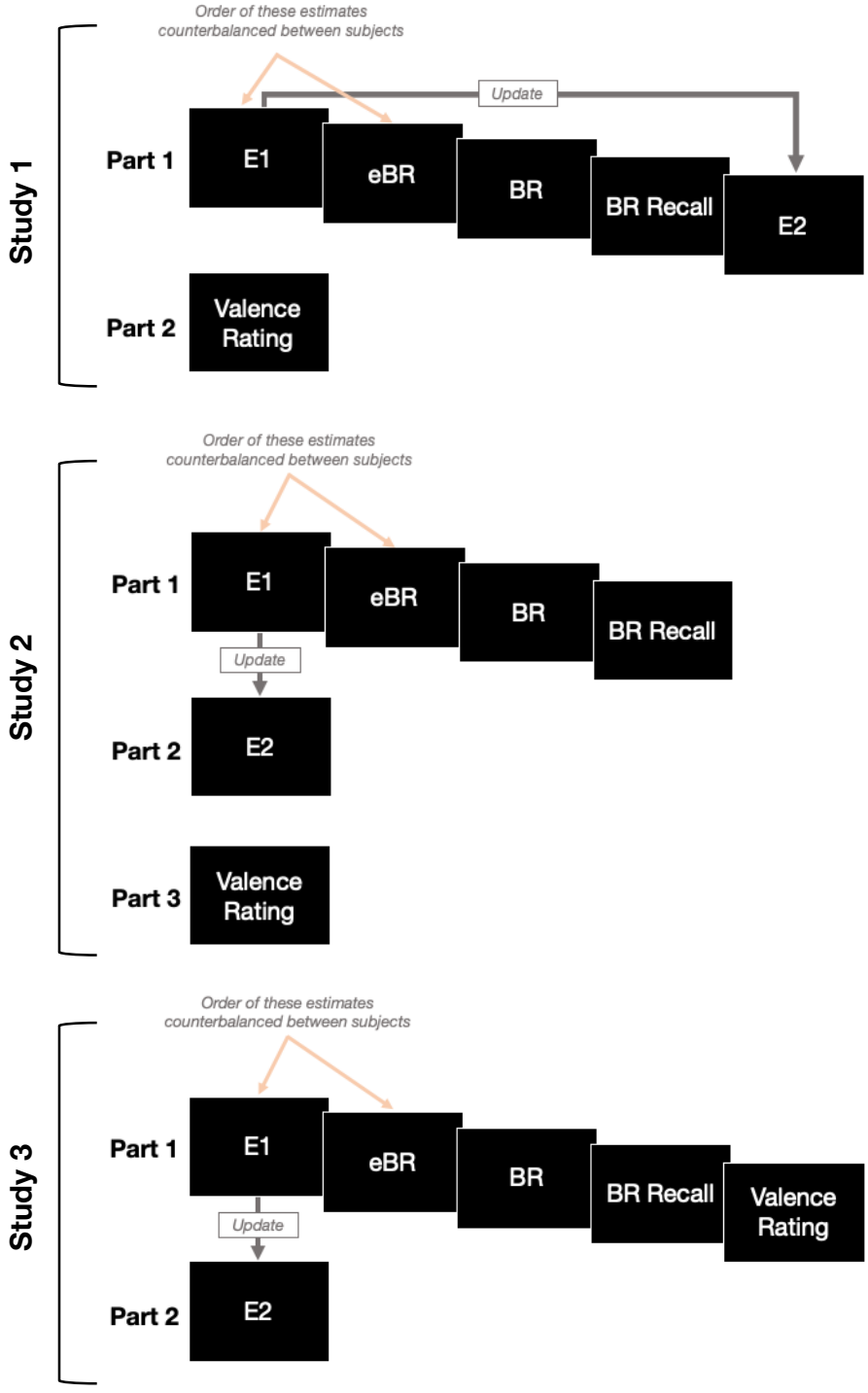

Figure 2.1: Schematics of the procedure used in each study. Images depict the task sequence for a single trial, differing slightly across studies: participants make an initial self-estimate ($E1$) and an estimate of the base rate (eBR), view the base rate (BR), write down the base rate correctly ($BR\ Recall$), rate the event's valence on a five-point Likert scale where a rating of 3 indicates the event is perceived as neutral ($Valence\ Rating$), and make a revised self-estimate ($E2$). The same stimuli were used in each study. We have followed the abbreviations of Kuzmanovic and Rigoux (2017). To aid cross-referencing, in Shah et al. (2016) E1 and E2 are SE1 and SE2, eBR is BR1, BR is actualBR.

criterion, except for a single supplementary result (noted below in the "further analyses" subsection). In addition, trials where E1 equalled the BR were necessarily excluded in our analyses, as they are in all analyses of the data from the update method, because the central, quasi-experimental classification of trials as desirable or undesirable — or in our case, upwards or downwards — cannot be applied. From the 5,100 trials recorded in each study, these criteria resulted in 339 (6.65%), 414 (8.12%), and 431 (8.45%) trials being excluded from studies 1, 2, and 3, respectively.

Measures

There are three measures that are central to interpreting results from the update method: direction of error, update value, and event valence. First, each trial's direction of error is determined by calculating the difference between BR and E1. If the difference is positive (i.e., $BR > E1$) then the direction of error is classed as upwards, and vice versa for a downwards direction of error.

The update value in each trial, which indicates the magnitude of belief updating, is then determined by calculating the absolute difference between participants' re-estimates and initial estimates ($|E2 - E1|$). This is then labelled as positive if updating goes towards the BR (i.e., in accordance with the direction of error) or negative if updating goes away from the BR (i.e., not in accordance with the direction of error). For example, if E1 is 40, BR is 50, and E2 is 30, the update would be coded as -10; whereas if the BR was 10 then the update would be coded as +10. Additional preregistered analyses in which direction of error is calculated in the normatively appropriate manner on the basis of participants' estimates of the BR (eBR) are presented in the "further analyses" subsection, whereby $BR > eBR$ indicates an upwards direction of error and vice versa.

Finally, event valence is determined trial-by-trial with the self-report question, "How would you feel about experiencing this event?", measured on a five-point Likert scale from "extremely negative" (1) to "neither positive nor negative" (3) to "extremely positive" (5). Following N. Garrett and Sharot (2017), we code ratings of 1 and 2 as negative, 3 as neutral, and 4 and 5 as positive. While our set of life events was compiled with neutral valence in mind, this procedure grants each participant the opportunity to provide updates for negative, positive, and neutral events, depending on their personal preferences.

2.1.2 Results

Main analysis

To test our central hypothesis that participants will update asymmetrically in response to neutral information, we conducted three studies with independent samples in which 100 participants proceeded through the update method with 51 life events. Following our preregistered analysis plan, we applied a linear mixed effects model (LMM) — as in Marks and Baines (2017) — to trials in which the stimulus (life event) was rated as neutral by the participant. Update value was entered as the dependent variable, direction of error (upwards/downwards) as a fixed factor, and participant as a random factor.

To select a model specification, we first fit the specification with the maximally complex random effects structure and then iteratively reduced model complexity until all degenerate random effects parameters were removed and the model was not singular (Bates et al., 2018). This procedure led us to a specification with only random intercepts by participant; however, the results hold across model specifications and we report the statistics of the maximally complex model specification in the supplementary information (Table A.2). Finally, we used Type III tests and Satterthwaite's approximation for degrees of freedom to calculate the statistical significance of the fixed effects. In all three studies, asymmetric belief updating was observed with neutral life events (Figure 2.2).

In Study 1, there were 1,521 trials in which participants updated in response to a rated-as-neutral life event, with 801 trials with an upwards direction of error ($M = 2.11$, $SD = 4.56$) and 720 with a downwards direction of error ($M = 11.33$, $SD = 15.91$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 1515) = 244.47$, $p < 0.001$), such that an upwards direction of error (i.e., $BR > E1$) decreased update scores by approximately 9.13 percentage points (fixed effect estimate) $\pm$ 0.58 (standard error), as compared to downwards direction of error.

In Study 2, there were 1,699 trials in which participants updated in response to a rated-as-neutral life event, with 831 trials with an upwards direction of error ($M = 4.16$, $SD = 10.01$) and 868 with a downwards direction of error ($M = 10.08$, $SD = 18.06$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 1694) = 77.09$, $p < 0.001$), such that an upwards direction of error (i.e., $BR > E1$) decreased update scores by about 6.24 percentage points (fixed effect estimate) $\pm$ 0.71 (standard error), as compared to downwards direction of error.

In Study 3, there were 1,667 trials in which participants updated in response to a rated-

as-neutral life event, with 828 trials with an upwards direction of error ($M = 4.13$, $SD = 8.82$) and 839 with a downwards direction of error ($M = 10.56$, $SD = 20.80$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 1662) = 68.48$, $p < 0.001$), such that an upwards direction of error (i.e., $BR > E1$) decreased update scores by about 6.51 percentage points (fixed effect estimate) $\pm$ 0.79 (standard error) as compared to downwards direction of error.

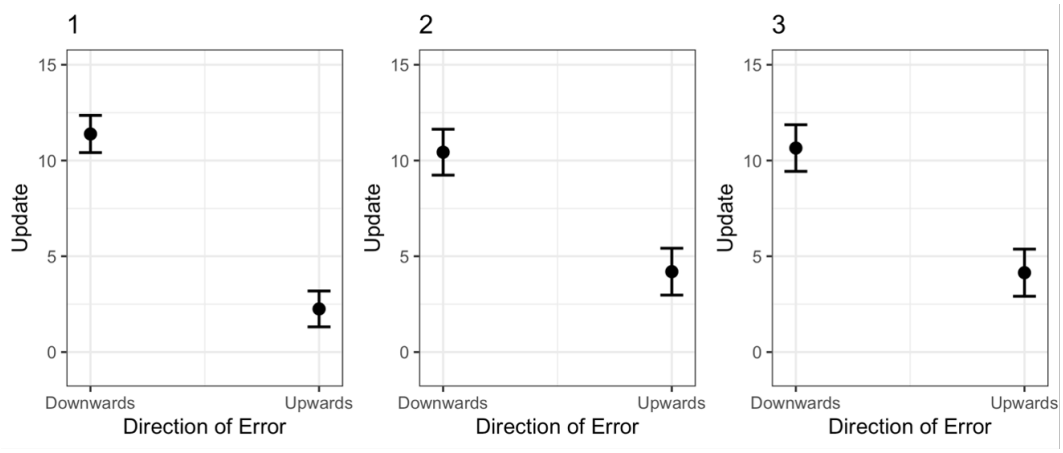

Figure 2.2: Asymmetries in belief updating with neutral life events in each of the three studies. Points indicate the magnitude of belief updating predicted by the linear mixed effects model with bars representing 95% confidence intervals. Numbering of plots corresponds to the study.

Secondary analysis

In addition to our central hypothesis testing, we also conducted a preregistered analysis of each study's data with LMMs that included event valence as a second fixed factor in interaction with direction of error. This analysis parallels that of previous related work (N. Garrett & Sharot, 2017) and allows us to assess whether the supposed desirability of information influenced belief updating, as well as view that effect in comparison with the asymmetry in neutral trials described above (Figure 2.3). However, because we used generally neutral events, so we are not well-powered to speak about positive and negative valence categories, given that there are relatively fewer events in these categories for each participant. To select a model specification, we followed the same procedure as in the main analysis, which led us to select a specification that includes only random slopes and intercepts by participant for direction of error and no correlation parameters. The results once again hold across model specifications, and we report the statistics of the maximally complex model in the supplementary information (Table A.3).

In Study 1, there were significant main effects of both direction of error ($F(1, 98) =$

233.80, $p < 0.001$) and event valence ($F(2, 4713) = 47.98$, $p < 0.001$), plus a significant interaction term ($F(2, 4706) = 61.80$, $p < 0.001$), whereby the updating asymmetry was smallest for the positive events. Despite this, participants updated more in response to a downwards direction of error across both negative and positive events. In 1,662 trials with negative life events, participants updated more in response to a downwards direction of error ($n = 683$, $M = 10.73$, $SD = 14.23$) than upwards ($n = 979$, $M = 2.86$, $SD = 6.31$). Equally, in the 1,578 trials with positive life events, participants updated more in response to a downwards direction of error ($n = 882$, $M = 4.97$, $SD = 9.41$) than upwards ($n = 696$, $M = 3.10$, $SD = 6.65$).

In Study 2, there were again significant main effects of direction of error ($F(1, 98) = 112.85$, $p < 0.001$) and event valence ($F(2, 4656) = 35.49$, $p < 0.001$), as well as a significant interaction term ($F(2, 4610) = 43.86$, $p < 0.001$). Among the 1,582 trials with negative events, participants updated more in response to a downwards direction of error ($n = 692$, $M = 13.08$, $SD = 21.33$) than upwards ($n = 890$, $M = 3.76$, $SD = 10.18$), in a similar fashion as Study 1. However, update values in the 1,405 trials with positive events displayed no asymmetry between downwards ($n = 835$, $M = 4.05$, $SD = 10.98$) and upwards direction of error ($n = 570$, $M = 4.48$, $SD = 10.62$).

In Study 3, there were once again significant main effects of direction of error ($F(1, 98) = 42.35$, $p < 0.001$) and event valence ($F(2, 4636) = 18.51$, $p < 0.001$), as well as a significant interaction term ($F(2, 4586) = 60.09$, $p < 0.001$). Interestingly, a flip in belief updating asymmetry was observed, which has previously been considered to be characteristic of optimistic belief updating. In 1,616 trials with negative life events, participants updated more in response to a downwards direction of error ($n = 639$, $M = 12.41$, $SD = 22.17$) than upwards ($n = 977$, $M = 3.62$, $SD = 9.51$); whereas participants updated less in response to a downwards direction of error ($n = 723$, $M = 3.45$, $SD = 10.42$) than upwards ($n = 663$, $M = 6.57$, $SD = 14.16$) in the 1,386 trials with positive life events.

In addition to this analysis with study-by-study LMMs, we can apply this same model specification to an aggregation of the data from Studies 1-3 so as to get a general overview of asymmetries across all trial categories among our 300 total participants. Here we find that the same significant main effects and interaction term remain: direction of error, $F(1, 459) = 482.39$, $p < 0.001$, event valence, $F(2, 14042) = 85.62$, $p < 0.001$); interaction term, $F(2, 13814) = 149.48.09$, $p < 0.001$. By applying Tukey's post-hoc test to examine pairwise differences among trial types, there is a significant asymmetry in downwards vs.

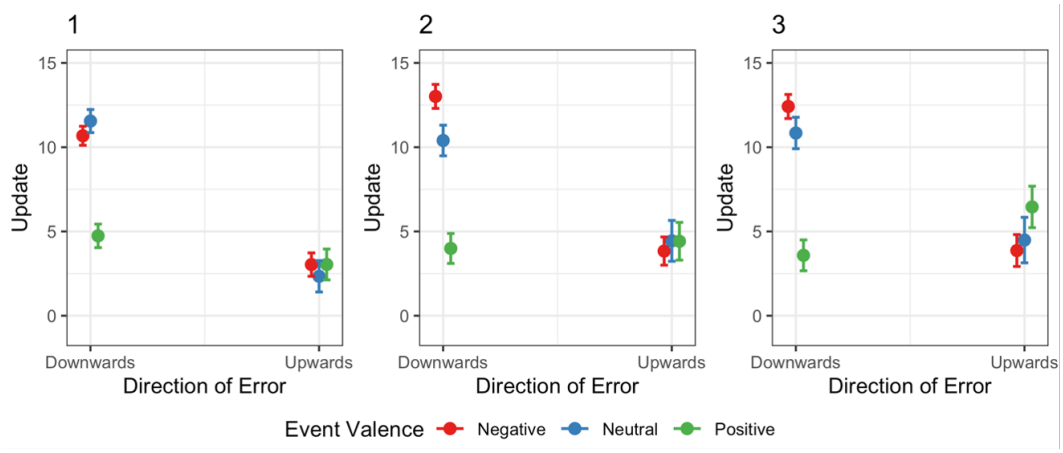

Figure 2.3: Asymmetries in belief updating with event valence and direction of error as fixed factors. Points indicate the magnitude of belief updating predicted by the linear mixed effects model with bars representing 95% confidence intervals. Numbering of plots corresponds to the study.

upwards updating with neutral events ($M = 10.78$ vs. $M = 3.70$, $M_{diff} = 7.12$, $p < 0.001$) of nearly the same magnitude as that in updating with negative events ($M = 11.96$ vs. $M = 3.53$, $M_{diff} = 8.47$, $p < 0.001$), and no significant asymmetry in updating with positive events ($M = 4.15$ vs. $M = 4.61$, $M_{diff} = -1.09$, $p < 0.887$).

Further analyses

Beyond the reported LMMs, there are a number of other analyses that have been applied to data produced by the update method. Relating to the previously outlined problems of misclassification and the bounded probability scale, Shah et al. (2016) analyse the data with an alternative classification scheme, and compare participants' updating with rational Bayesian predictions. Kuzmanovic and Rigoux (2017) seek to account for the influence of individuating information with a computational modelling technique based on reinforcement learning. And, finally, one might argue that we should consider the analysis used in the original work of Sharot et al. (2011), which uses regressions to assess correlations between "estimation error" (i.e., $|E1–BR|$) and belief updating. In the following paragraphs, we apply each of these techniques to the data from our three studies. Despite different attempts to remedy the limitations of the update method, the results of these techniques display unexplained variability in updating asymmetries with neutral events, which cannot be explained by a motivational optimism account. Each of these further analyses was preregistered, except for the regression analysis of updating behaviour, which was conducted in response to a peer review.

Accounting for direction of error misclassification. A central limitation of the standard update method is its neglect of individuating information. As pointed out by Shah et al. (2016), participants may hold one estimate of their personal likelihood of experiencing each event (E1), which is influenced by individuating information, and another estimate of the likelihood of an average person experiencing each event (an estimate of the base rate; eBR), which is not influenced by individuating information. This means that by classifying direction of error (or, in the case of previous studies of optimistic belief updating, the desirability) in each trial on the basis of E1 instead of eBR, trials can be misclassified and subsequently muddle the results. To assess the empirical consequence of misclassification, we re-analysed the data with an alternative direction of error assigned by comparing eBR to BR in each trial (i.e., $eBR > BR$ is downwards, and vice versa). Across all of the data collected, 25.30% ($n = 3,571$) of trials were misclassified. In this section we report the reduction in the fixed effect estimates produced by LMMs that exclusively analysed neutral trials (as in the main analysis), as well as the altered effects produced by LMMs that included event valence as a second fixed factor. In each instance we followed the procedure for the LMMs reported in the main analysis, whereby we first fit the model specification with the maximally complex random effects structure and then iteratively reduce the random effects structure until all degenerate random effects parameters are removed and the model is not singular. For the LMMs with only neutral trials, this left us with only random intercepts by participant; and for the LMMs with event valence as a second fixed factor this left us with random slopes and intercepts for direction of error by participant, and no correlation parameters. While the maximally complex random effects specifications were singular, we also report the results of these specification in Tables A.4 and A.5 for comparison.[3]

In Study 1, accounting for the misclassification of direction of error reduced the fixed effect estimate by 78%, from 9.13 ($SE = 0.58$, $p < 0.001$) to 1.98 ($SE = 0.44$, $p < 0.001$), but still, an LMM exclusively testing trials with neutral events displayed a significant asymmetry ($F(1, 1436) = 20.73$, $p < 0.001$). An LMM including event valence as a

[3]This analysis differs slightly from the preregistered analysis plan in which we stated that we would test for an interaction between these classification schemes. Since the reclassification of direction of error also means that update values can change (i.e., an update of -10 would change to +10 if the direction of error is reclassified), different observations were identified as outliers by our exclusion criteria (i.e., ± 3 $\times$ the interquartile range for a given condition). This in turn results in two distinct datasets: one where the standard classification scheme is applied and one where misclassification is accounted for. For this reason, it was not possible to test for an interaction by adding the classification scheme as a fixed factor in our LMMs that tests for effects within a dataset, and we instead provide qualitative comparisons between analyses.

second fixed factor produced significant but reduced main effects of direction of error ($F(1, 96) = 71.52$, $p < 0.001$) and event valence ($F(2, 4541) = 26.95$, $p < 0.001$), and a weakened interaction term ($F(2, 4509) = 3.29$, $p = 0.037$). While the same asymmetries are displayed (i.e., downwards direction of error elicited significantly greater updates than upwards for neutral, negative, and positive trials), there is a prominent reduction in the magnitude of these asymmetries (Figure 2.4).

Accounting for misclassification in Study 2 reduced the fixed effect estimate by 63%, from 6.24 ($SE = 0.71$, $p < 0.001$) to 2.32 ($SE = 0.71$, $p = 0.001$), but an LMM exclusively testing trials with neutral events again displayed an asymmetry ($F(1, 1639) = 10.63$, $p = 0.001$). Reduced effects were also observed once event valence was included as a second fixed factor in the LMM, but nevertheless, there were still significant main effects of direction of error ($F(1, 93) = 31.81$, $p < 0.001$) and event valence ($F(2, 4534) = 9.55$, $p < 0.001$), and an interaction ($F(2, 4410) = 8.33$, $p < 0.001$). Once again, the same significant asymmetries persisted but their magnitudes were heavily reduced (Figure 2.4). In Study 3, misclassification accounted for 67% of the fixed effect estimate produced by an LMM exclusively testing trials with neutral events, reducing 6.51 ($SE = 0.79$, $p < 0.001$) to 2.15 ($SE = 0.71$, $p = 0.002$). But, again, the asymmetry in trials with neutral events remained ($F(1, 1565) = 9.23$, $p = 0.002$). However, once event valence is included in the LMM as a second fixed factor, the "flip" in asymmetries commonly interpreted as a result of valence-dependent updating disappears. While the effect of direction of error ($F(1, 101) = 5.28$, $p = 0.024$), event valence ($F(2, 4400) = 21.10$, $p < 0.001$), and the interaction remained significant ($F(1, 4149) = 5.91$, $p = 0.003$), there is another notable reduction in the magnitude (Figure 2.4).

These results suggest that a non-negligible amount of trials in the update method may be missclassified based on E1 (rather than eBR) as having an upwards vs. downwards direction of error; or, in the conventional use of the update method with valenced life events, missclassified as involving desirable vs. undesirable information. Upon correcting for this issue and using the normatively appropriate classification scheme on the basis of eBR, the asymmetries in belief updating were greatly attenuated. Still, statistically significant asymmetries with neutral events were observed, thereby suggesting that this correction alone is not enough to remedy the update method. This is because, while this analysis alleviates the issue of misclassification, it does not address the issue of the bounded probability scale. As prescribed by Bayes' theorem, updates of the same absolute

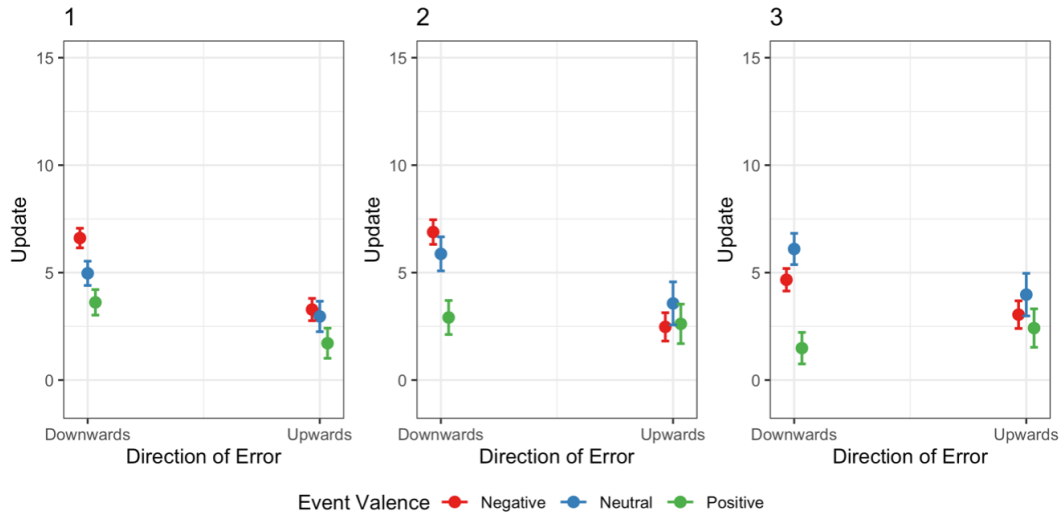

Figure 2.4: Plots of the observed asymmetries in belief updating once the misclassification of direction of error is accounted for in each study. Points indicate the estimated marginal means of belief updating as predicted by the linear mixed effects model with bars representing 95% confidence intervals. Numbering of plots corresponds to the study.

magnitude cannot be equated if they are on different parts of the probability scale or made in opposite directions. For example, updating from 20% to 30% is not equivalent to updating from 40% to 30%, yet this analysis leaves this unaccounted for (see Figure 2.8).

Comparisons with rational Bayesian predictions. Given that updates in different parts of the scale cannot be mathematically equated to one another, a seemingly sensible analysis of data produced by the update method is to compare participants' actual updating behaviour to rational Bayesian predictions. As done in Shah et al. (2016), the collection of eBRs from participants in our studies allowed for the calculation of implied likelihood ratios (LHR) — a measure of participants' individuating information derived from Bayes' theorem — for each trial following the logic of Equation 2.1 and Equation 2.2:

$$PosteriorOdds = PriorOdds \times LHR \tag{2.1}$$

$$\frac{P(h|e)}{1-P(h|e)} = \frac{P(h)}{1-P(h)} \times LHR \tag{2.2}$$

If eBR and E1 are then divided by 100, the equation can be rewritten with the terminology of the present studies as follows in Equation 2.3:

$$LHR = \frac{E1}{1-E1} \div \frac{eBR}{1-eBR} \tag{2.3}$$

With the implied LHRs serving as a measure of individuating information participants believe they possess, we subsequently calculated the predicted posterior odds for each trial (Equation 2.4), which could then be used to indicate how much a rational Bayesian agent "should" update in each trial (Equation 2.5):

$$PosteriorOdds = \frac{BR}{1 - BR} \times LHR \tag{2.4}$$

$$BayesianUpdate = |\frac{E1 - PosteriorOdds}{1 + PosteriorOdds}| \tag{2.5}$$

From here, we tested for asymmetries in belief updating with two measures across conditions, within participants: a Bayesian difference measure (i.e., predicted belief change – observed belief change) and a Bayesian ratio measure (i.e., observed belief change ÷ predicted belief change).

Our results show that there is variability between studies (Tables 2.1 to 2.2). For instance, comparisons of upwards versus downwards updating with the Bayesian ratio measure indicates no asymmetry in trials with neutral events in Studies 2 and 3, but there is a statistically significant asymmetry observed in Study 1 and in the aggregated data of Studies 1-3, aggregate analyses indicate that the statistical artefact, whereby asymmetry persists in trials with neutral life events, remained when interpreting the data in this way.

Study	*Event Valence*	*Mean of difference measure for downwards trials*	*Mean of difference measure for upwards trials*	*t*	*p-value*
1	Positive	0.13	0.09	-4.31	<0.001
	Neutral	0.10	0.08	-2.00	0.049
	Negative	0.09	0.10	1.54	0.128
2	Positive	0.14	0.07	-4.57	<0.001
	Neutral	0.11	0.08	-2.04	0.044
	Negative	0.06	0.10	2.63	0.009
3	Positive	0.12	0.07	-2.79	0.006
	Neutral	0.10	0.07	-2.52	0.013
	Negative	0.06	0.11	2.84	0.006
Aggregate	Positive	0.13	0.08	-6.47	<0.001
	Neutral	0.10	0.07	-3.81	<0.001
	Negative	0.07	0.10	4.14	<0.001

Table 2.1: Results of paired t-tests comparing Bayesian difference measures that compare participants' updating to rational Bayesian predictions.

These results highlight the inherited flaws of this analysis — although the comparisons

Study	*Event Valence*	*Median ratio measure for downwards trials*	*Median ratio measure for upwards trials*	*Z*	*p-value*
1	Positive	0.00	0.00	1.41	0.921
	Neutral	0.57	0.04	-3.73	<0.001
	Negative	0.49	0.00	-4.92	<0.001
2	Positive	0.00	0.17	-1.19	0.116
	Neutral	0.53	0.28	0.79	0.784
	Negative	0.51	0.07	-2.26	0.012
3	Positive	0.00	0.46	-2.14	0.016
	Neutral	0.53	0.28	-0.13	0.449
	Negative	0.51	0.09	-3.71	<0.001
Aggregate	Positive	0.00	0.10	-2.36	0.009
	Neutral	0.53	0.25	-2.34	0.010
	Negative	0.51	0.00	-6.25	<0.001

Table 2.2: Results of Wilcoxon signed rank tests comparing Bayesian ratio measures that compare participants' updating to rational Bayesian predictions.

are normatively appropriate, both the difference and ratio measures are susceptible to artefacts produced by the bounded probability scale and uneven effects of response noise (Shah et al., 2016). When a participant is required to translate a perceived personal risk estimate onto the probability scale, response noise will arise where a participant's non-integer estimates are forcibly rounded, where a participant misinterprets his or her internal state, or where a participant simply mis-types (e.g., entering "15" instead of "14"). The influence of such response noise will depend on where updating is taking place on the probability scale. For instance, as one approaches either end of the scale, response noise will constitute different proportions of the probability estimate. This issue is in turn reflected in the Bayesian comparison measures, deeming them insufficient to address the statistical artefact.

Analysis of learning rates derived from Kuzmanovic and Rigoux's (2017) computational model. A recent paper by Kuzmanovic and Rigoux (2017) proposed two new modelling techniques as analytic remedies for update method. First, they present a Bayesian model in which they fit a scaling (S) and asymmetry (A) parameter to model participants subjective updates:

$$SubjectiveUpdate_{good} = BayesianUpdate \times (S + A) \tag{2.6}$$

$$SubjectiveUpdate_{bad} = BayesianUpdate \times (S - A) \tag{2.7}$$

The logic of these equations is that if participants update equally on desirable and undesirable information, the asymmetry parameter, A, will equal zero, and hence the right-hand bracketed expression will be constant across both equations. Interestingly, the modelling represented in these equations can be considered computationally equivalent to determining whether $\frac{SubjectiveUpdate_{good}}{BayesianUpdate} = \frac{SubjectiveUpdate_{bad}}{BayesianUpdate}$, which Shah et al. (2016) and the present work already addressed as the Bayesian ratio measure.

Beyond this, Kuzmanovic and Rigoux (2017) also propose a reinforcement learning model. The full model presented is:

$$BeliefUpdate = LearningRate \times PredictionError \times (1 - rP \times W) \tag{2.8}$$

where $BeliefUpdate$ represents the update value, $PredictionError$ represents the difference between eBR and BR, and rP represents "relative personal knowledge," and W is a free parameter to account for participants' individual variability in their sensitivity to rP (W is thus irrelevant when considering rational Bayesian agents) (Kuzmanovic & Rigoux, 2017). To address the implications of the reinforcement learning model we consider the crux of the argument: do learning rates differ across conditions? To do so, we simply rearranged Equation 2.8 to permit a trial-by-trial calculation of learning rates:

$$LearningRate = \frac{BeliefUpdate}{PredictionError \times (1 - rP)} \tag{2.9}$$

Using Wilcoxon signed rank tests, we then compared learning rates across conditions, within participants. Once again, we observed unexplained variability in the results of each study — statistically significant asymmetries were observed in Study 1 and in the aggregated data, but not in Studies 2 and 3 — suggesting that the statistical artefact pervades this approach too with asymmetrical learning rates capable of being seen in trials with valence-neutral events (Table 2.3).

As identified in separate ongoing work (Harris & Hahn, 2021), the reason this approach fails to address the artefact can be traced back to its use of rP as a substitute for LHR, the normatively appropriate representation of individuating knowledge. As Kuzmanovic and Rigoux (2017) explain, rP represents the "the difference between eBR and E1 relative to the maximal possible difference in each trial" (p. 5). The calculation proceeds as follows:

Study	*Event Valence*	*Median learning rate for downwards trials*	*Median learning rate for upwards trials*	*Z*	*p-value*
1	Positive	0.00	0.00	0.28	0.610
	Neutral	0.61	0.04	-3.04	0.001
	Negative	0.60	0.00	-4.83	<0.001
2	Positive	0.00	0.19	-0.89	0.187
	Neutral	0.68	0.45	0.53	0.704
	Negative	0.56	0.14	-1.86	0.032
3	Positive	0.00	0.55	-2.45	0.007
	Neutral	0.64	0.30	-0.46	0.324
	Negative	0.57	0.12	-4.02	<0.001
Aggregate	Positive	0.00	0.15	-2.15	0.016
	Neutral	0.66	0.28	-2.14	0.016
	Negative	0.58	0.00	-6.12	<0.001

Table 2.3: Results of Wilcoxon signed rank tests comparing the learning rate measure derived from the reinforcement learning model presented by Kuzmanovic and Rigoux (2017).

$$rP = \begin{cases} (eBR - E1)/(eBR - 1) & \text{if } E1 < eBR \\ (E1 - eBR)/(99 - eBR) & \text{if } E1 > eBR \\ 0 & \text{if } E1 = eBR \end{cases} \tag{2.10}$$

When participants estimate their own risk (E1) as equal to their estimate of the base rate (eBR), $rP = 0$, appropriately indicating that they have no individuating knowledge. However, if rP is to accurately represent a participant's relative personal knowledge, it should be sensitive only to the LHR. Figure 2.5 plots rP against the LHR for seven different estimates of the base rate. Appropriately, there is a monotonic relationship between rP and the amount of individuating information one possesses: rP increases as the absolute distance between E1 and eBR increases. However, for the same LHR, rP differs across different estimates of the base rate. Given that the normative amount of personal knowledge remains the same within each LHR, this demonstrates that rP is not appropriately representing the amount of individuating information possessed. Consequently, the model and its learning rate measure are left vulnerable to potential statistical artefacts, as is reflected in our results (Table 2.3).

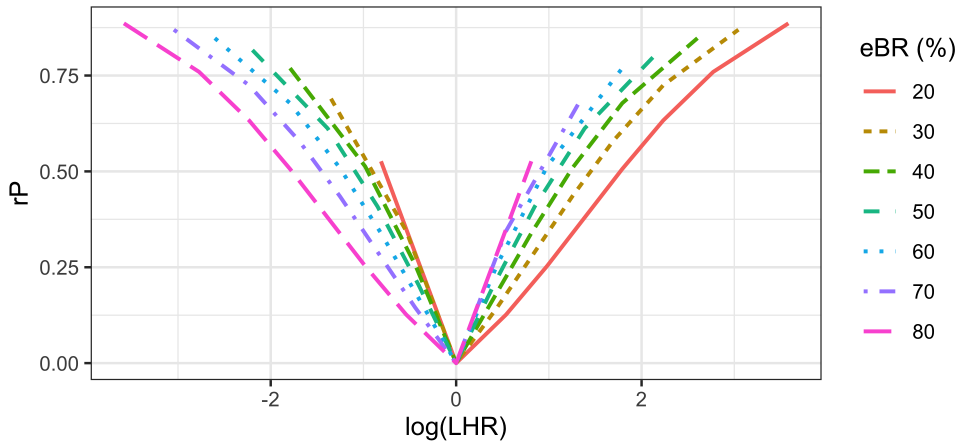

Figure 2.5: Numerical simulation comparing Kuzmanovic and Rigoux's (2017) rP parameter to the Bayesian LHR parameter. rP and LHR are calculated for the listed eBRs and $E1 = \{10, 20, 30, \ldots 90\}$. Figure adapted from Harris and Hahn (2021).

Regression analysis of updating behaviour. Beyond the more recent techniques outlined above from Shah et al. (2016) and Kuzmanovic and Rigoux (2017), one might also suggest that we apply the procedure used in the original work of Sharot et al. (2011). In their analysis, Sharot et al. (2011) use regressions to assess correlations between "estimation error" (i.e., $|E1–BR|$) and belief updating. Since estimation error is presumably correlated with the magnitude of updates, comparisons of regression coefficients across conditions, within participants is expected to display potential asymmetries while naturally controlling for the magnitude of estimation error. In other words, if desirable trials have a larger regression coefficient than undesirable trials within participants, it would seem that participants are more conservative in belief updating when faced with bad news as compared to good news.

While it is meaningful to control for estimation error to ensure that observed asymmetries do not merely reflect an uneven distribution of errors, this particular analysis was not preregistered as it is conceptually flawed. As outlined in the introduction above it is meaningless to subtract the base rate (BR) from Sam's individual estimate (E1), because normatively, Sam's individual estimate should be a function of both of these quantities, and, normatively, Sam could even be required to revise the individual estimate in the opposite direction of that "error," depending on what Sam took the base rate to be (Shah et al., 2016).

Nevertheless, we followed this regression analysis procedure in a further, unregistered examination of our data. Similar to the other alternative analyses outlined above, this

analysis returns a statistically significant asymmetry with neutral events in Study 1, but there is neither a significant asymmetry in Studies 2 and 3, nor in the aggregated data (Table 2.4). Further, the statistically significant asymmetry in Study 1 does not hold when our preregistered exclusion criterion is not applied. It is difficult to interpret these results because this regression analysis falsely equivocates upwards and downwards updating on the compressed probability scale. Once again, the degree to which one should normatively update their beliefs is the product of individuating information and the base rate. This means that even if two individuals are faced with the same BR, have identical likelihood ratios, and provide E1s that are equal absolute distances from the BR — but one agent's E1 is above the BR and the other's is below the BR — their prescribed Bayesian updates will differ (see Figure 2.8). Yet, the regression analysis cannot account for this because it only considers the raw belief change and estimation error, while neglecting the influence of individuating information.

Study	*Event Valence*	*Mean coefficient for downwards trials*	*Mean coefficient for upwards trials*	*t*	*p-value*
1	Positive	-0.02	0.10	-1.90	0.060
	Neutral	0.20	0.00	3.57	<0.001
	Negative	0.20	0.20	2.23	0.028
2	Positive	0.04	0.30	-3.22	0.002
	Neutral	0.11	0.06	0.44	0.662
	Negative	0.23	0.11	1.40	0.165
3	Positive	-0.14	0.11	-2.06	0.042
	Neutral	0.13	0.08	0.62	0.540
	Negative	0.40	0.13	1.52	0.132
Aggregate	Positive	-0.04	0.17	-4.03	<0.001
	Neutral	0.15	0.05	1.75	0.081
	Negative	0.27	0.15	2.76	0.006

Table 2.4: Results of paired t-tests comparing regression coefficients whereby "estimation error" is used to predict update values.

Follow-up simulations. In an additional exploratory analysis, we used simulations to test the robustness of the regression analysis results, and to explore why there is considerable, seemingly uninterpretable variability in results across studies. The rationale for our simulations stems from the statistical artefact hypothesis' explanation that "the very nature of the artefacts that plague the update method mean that, given the right set of events, everything and anything could empirically be found, even in entirely unbiased agents" (Shah et al., 2016, p. 107). That is, the statistical artefact hypothesis predicts

that the results of the regression analysis could change if we were to have sampled stimuli (life events) with different statistical properties (e.g., the events' average base rate error), which, crucially, are not valence-dependent. By simulating 500 "experiments" where we sample participants and events from the aggregated data of Studies 1-3 we found that the results of the regression analysis do indeed seem to be driven by statistical properties of the events used: in response to events with low average base rate error (i.e., $|eBR-BR|$) participants display asymmetric updating when they rate those events to be neutrally-valenced, but not when they rate them to be positively- or negatively-valenced (Figure 2.6). This is of course a nonsensical result that cannot be attributed to motivational optimism.

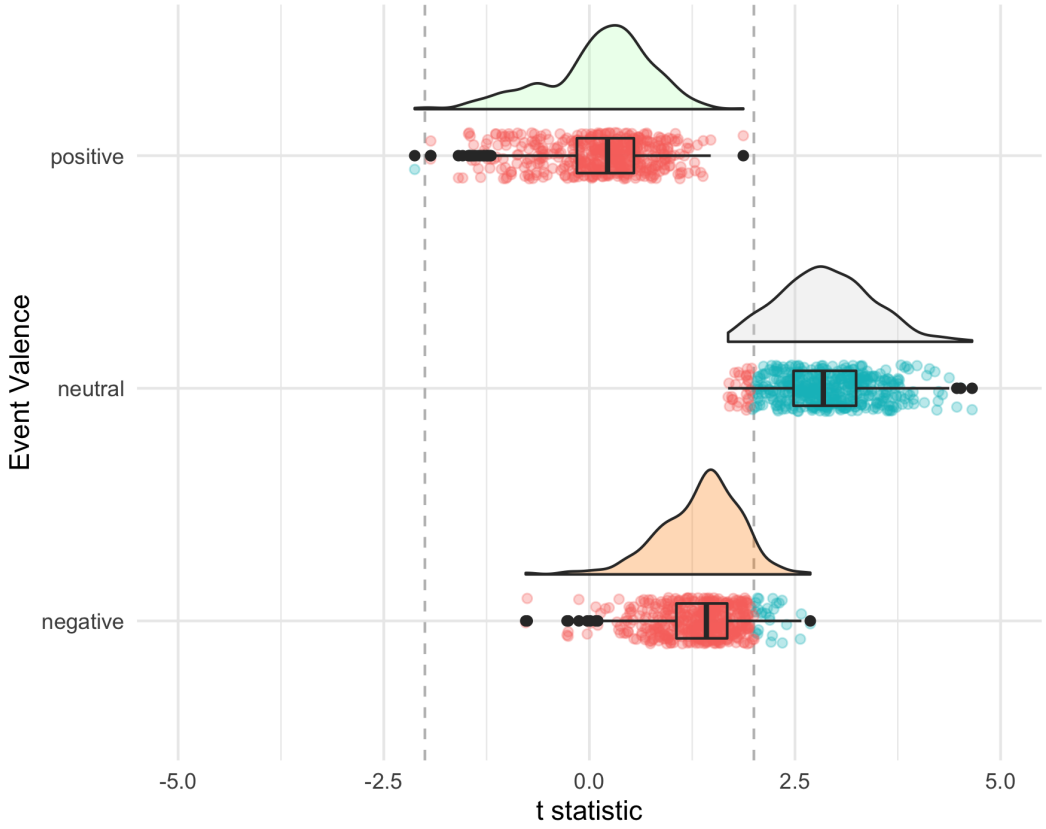

Figure 2.6: Results of the regression analysis using estimation error as the independent variable in 500 simulated "experiments." Each iteration, or experiment, involved sampling 200 participants and their responses to the 20 events with the lowest average base rate error ($|eBR-BR|$) from the aggregated data of Studies 1-3. Results are split out by event valence (y-axis). Blue points represent statistically significant ($p < 0.05$) asymmetries returned by paired samples t-tests comparing the regression coefficients in upwards vs. downwards updating. Red points represent non-significant asymmetries. The direction and magnitude of asymmetries is indicated by the t statistic (x-axis). Results suggest that, in response to events with low average base rate error, participants more frequently display asymmetric updating when they rate those events to be neutrally-valenced as compared to positively- or negatively-valenced.

Supplementary Study 4. Building on this, we subsequently preregistered and ran an additional Study 4 to empirically test the conclusions of our simulations. Specifically, we recruited 200 participants via the Prolific Academic platform ($M_{age} = 30.66$, $SD_{age} = 11.35$; 133 female, 63 male, 4 other) and presented them with the 20 life events that elicited the lowest average base rate error in Studies 1-3. We applied the same analyses that were used to analyse Studies 1-3: LMMs as in the main and secondary analyses, and the each of the further analyses described above. The LMMs replicated the results of Studies 1-3 and showed statistically significant asymmetries with neutral events, and held after accounting for misclassification (as well as after accounting for potential post-treatment bias and after adding stimuli as a random factor; for more details see the following section and supplementary information in Appendix A). However, the results of the supplementary analyses were inconclusive. Bayesian comparisons and the analysis of learning rates returned non-significant results for each event type: neutral, negative, and positive. The regression analysis returned non-significant results for neutral ($t(130) = 1.08$, $p = 0.282$) and negative events ($t(62) = 0.48$, $p = 0.635$), and a statistically significant asymmetry with positive events ($t(163) = -2.06$, $p = 0.041$).

In considering why the results of this additional study were inconclusive, we noted that there were noticeable departures in the distributions of how the participants in Study 4 estimated the relevant statistical properties of the new event sub-set vis-à-vis the data on which our selection had been based (Figure 2.7). This underscores further the limits of the current update methodology.

Further further analyses

Indeed, there are even more analytic approaches that may be applicable, such as including stimuli (life events) as a random factor to heed the statistical literature pointing out that the failure to do so can inflate Type I error rates on fixed effect estimates (Judd et al., 2012; Yarkoni, 2019). However, this does not resolve the asymmetry observed with the LMMs (Appendix A.1.2). Alternatively, one might argue that Bayesian modelling equivalents to the LMMs reported here should be used. But such an analysis introduces an additional "researcher degree of freedom" by requiring the specification of a prior distribution; and given that there is no precedent of this in the literature, we restricted our analyses to address our hypothesis.

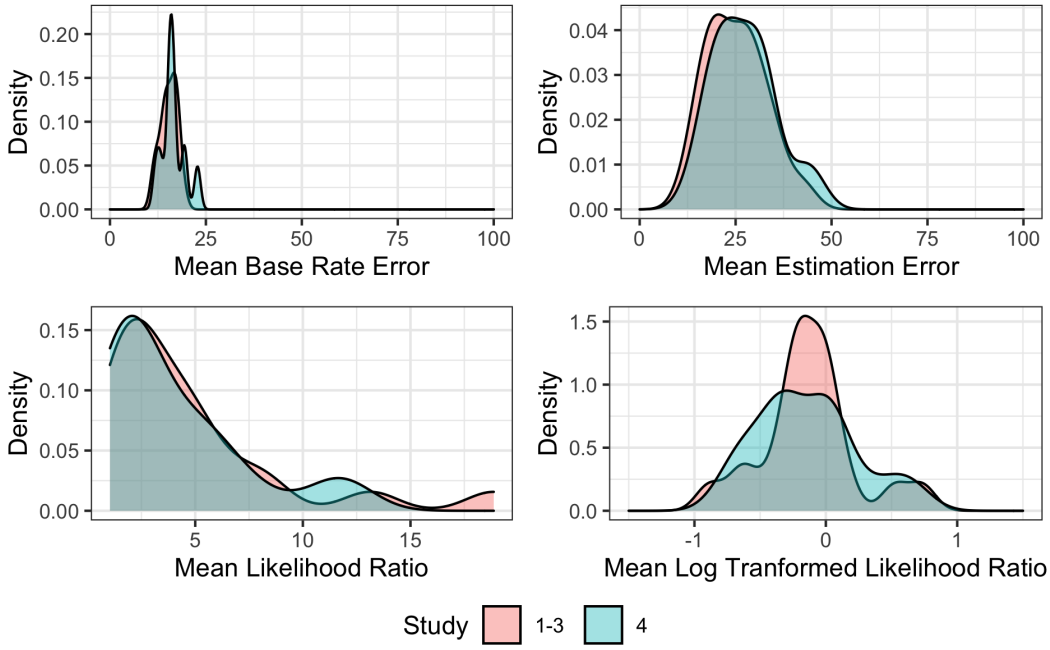

Figure 2.7: Density plots displaying the distributions of how participants estimated the statistical attributes of the 20 events used in Study 4 as compared to what was observed in Studies 1-3

2.1.3 Discussion

The present research provides the first targeted application of the update method to test for asymmetric belief updating with neutral stimuli (life events). In the main analysis, the central result obtained was the thrice replicated asymmetry in belief updating with neutral stimuli (Figure 2.2), which is coupled with variability among the results returned by various analytic techniques that have been proposed as fixes for the flaws of the update method (summarised in Table 2.5). Such findings are unpredicted and unexplainable by motivational accounts. Consequently, our results contribute to the debate concerning the status of the optimistic belief updating phenomenon (Sharot et al., 2011) by demonstrating the empirical consequence of the statistical confounds previously highlighted (Shah et al., 2016) — empirical consequences that subsequent rebuttals have questioned (e.g., N. Garrett & Sharot, 2017; Kuzmanovic & Rigoux, 2017). While seemingly optimistic asymmetries can arise in the data, the current results suggest that such effects are not valence-driven. Given that an asymmetry can be observed in the absence of valence, and thus an absence of motivation, the presence of an asymmetry with valenced events is insufficient evidence for one to conclude that a motivational bias is present.

Previous research has countered the statistical artefact hypothesis by including positive events in the deployment of the update method (e.g., N. Garrett & Sharot, 2017; Marks &

	Study 1	Study 2	Study 3	Aggregate
Accounting for misclassification (Shah et al., 2016)	<0.001	0.001	0.002	<0.001
Bayesian difference measure (Shah et al., 2016)	0.049	0.044	0.013	<0.001
Bayesian ratio measure (Shah et al., 2016)	<0.001	0.784	0.449	0.010
Learning rates (Kuzmanovic & Rigoux, 2017)	0.001	0.704	0.324	0.016
Regression analysis* (Sharot et al., 2011)	<0.001	0.662	0.540	0.081

Table 2.5: P values indicating the statistical significance of the asymmetries observed with neutral events according to five analytical techniques intended to resolve the flaws of the update method in each study and the aggregated data of Studies 1-3. Asymmetries under the alternative classification scheme were determined with linear mixed effects models. Comparisons of upwards vs. downwards updating using the Bayesian difference measure and the regression analysis (with "estimation error" as the independent variable) were made with t-tests; comparisons using the Bayesian ratio measure and learning rates were made with non-parametric Wilcoxon signed rank tests. The analysis marked with an asterisk was not preregistered.

Baines, 2017). While such studies succeed in demonstrating asymmetric belief updating with positive events in a manner that is consistent with motivational optimism, they fail to appreciate "that the statistical artefacts will necessarily vary in expression as a function of events" (Shah et al., 2016, p. 106). Consequently, "the very nature of the artefacts that plague the update method mean that, given the right set of events [with the right statistical attributes], everything and anything could empirically be found, even in entirely unbiased agents" (Shah et al., 2016, p. 107). This is because the statistical artefact hypothesis is multi-faceted, being driven by both the base rate and the amount of individuating information the individual believes they possess about their chance of experiencing a particular event. In fact, the variability of these subjective confounds can be seen even when an identical set of life events is used. Across our three studies, there are differences in the distributions of implied likelihood ratios (LHRs), estimation error ($|E1 - BR|$), and base rate error ($|eBR - BR|$) (see Figures A.2 to A.4 for the empirical distributions of these variables in each study; and see Figure 2.8 for a numerical simulation of the relationship between these variables). It may be these discrepant statistical characteristics that give rise to varying results across studies.

One tempting approach to side-step the variability in results across studies and analyses is to focus solely on the aggregated data from Studies 1-3 as the single "answer." This

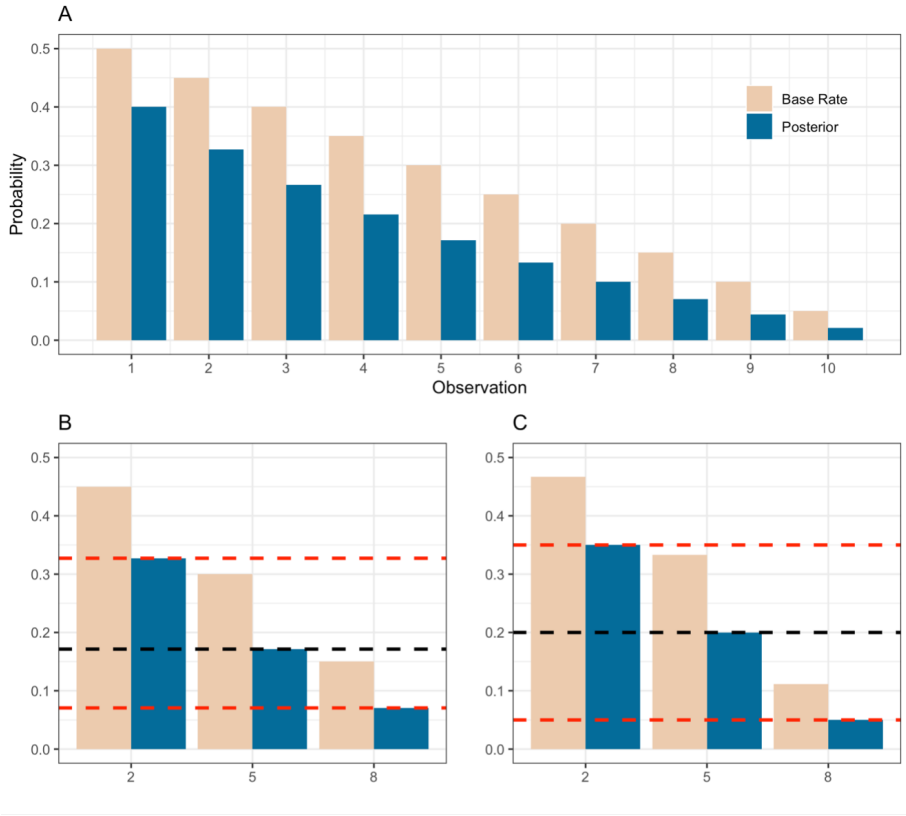

Figure 2.8: Numerical simulation of probability scale compression and the relationship between base rate (BR) error, belief change, and implied likelihood ratios (LHR). [A] Ten simulated observations (x-axis) of paired base rates and posterior probabilities (y-axis). Across all ten pairs the LHR, $\frac{E1}{(1-E1)} \div \frac{eBR}{(1-eBR)}$, that is, the degree of individuating knowledge is the same (in this plot, 0.4), and the posterior probability is derived via Bayes' theorem, by combining that individuating knowledge with the respective base rate. BR error is [B] Observations 2, 5, and 8 from Plot A. While the LHR (0.4) and absolute BR error (0.15) is held constant, the absolute belief change cannot be equal when updating in opposing directions (0.16 moving upwards on the scale from observation 5 to 2; 0.10 moving downwards on the scale from observation 5 to 8). [C] Observations 2, 5, and 8 when absolute belief change (0.15) and LHRs (0.4) are held constant, which results in unequal absolute BR errors (0.12 to move upwards on the scale from observation 5 to 2; 0.06 to move downwards on the scale from observation 5 to 8).

approach is tempting as it achieves the highest possible statistical power, which has been useful in supporting our argument. This approach is indeed logical in the sense that it achieves the highest possible statistical power, but it overlooks a key point: we are observing variability in results in already high-powered studies with 100 participants each when the update method has been widely used for neuroscientific studies that often have far fewer participants. The original work of Sharot et al. (2011) relied on a sample of just 19 participants, N. Garrett et al. (2014) had 30 participants, N. Garrett and Sharot (2014) had 32 participants, Kuzmanovic and Rigoux (2017) had 27 participants, and even the large (by neuroscience standards) sample of Moutsiana et al. (2013) was just 52 participants.

Our findings suggest that there is not a consistent, behaviourally-large, motivational asymmetry in belief updating, as would be required to inspire confidence in the neural correlates – and therefore the existence – of the optimistic bias phenomenon. While variability with "sensible" stimuli (i.e., negative or positive events) is often attributed to natural noise in the world, we observed similar variability with entirely nonsensical stimuli (i.e., neutral events). The dominant theory for a general optimism bias is that it is an adaptive, self-serving mechanism that enhances exploratory behaviour and reduces stress and anxiety as a regular feature of healthy human cognition (Sharot et al., 2011). Further, it has been argued that the ability to integrate desirable and undesirable information reflects two dissociated processes with different developmental trajectories in the human brain (Chowdhury et al., 2014; Moutsiana et al., 2013). On neither perspective would one expect the findings presented here. Were the update method a suitable tool for probing optimistic bias, it simply should not show "bias" with valence-neutral events. The fact that it can, and that results vary across samples and analyses, renders both the behavioural results of earlier studies and the underlying neurological correlates, based on small samples, uninterpretable.

Whilst our investigation included small changes in methodology across studies and covered several supplementary analyses, readers familiar with the optimistic belief updating literature might point out that we did not control for certain covariates that have been considered in past studies, such as participants' familiarity, vividness, emotional arousal, and perceived controllability of the events presented. While this could indeed be viewed as a limitation of the present work, we do not see it as such because past studies consistently report no change in results with or without these controls (e.g., Moutsiana et al., 2013; Shah et al., 2016; Sharot et al., 2011), and moreover, certain covariates like emotional arousal are, by definition, irrelevant for neutral life events. Likewise, it could be argued that our results are limited due to the fact that we only considered one framing; we only asked participants how likely they were to experience the events presented and not how likely they were to not experience the events. However, the potential of a confounding framing effect also seems unlikely given that the past studies that do control for it report there to be no consequence (e.g., N. Garrett & Sharot, 2014; N. Garrett et al., 2014; Korn et al., 2014; Sharot et al., 2011).

Other methods for optimism research need to be used, but there is no quick fix. One plausible direction is to test for bias as asymmetric deviations from Bayes' theorem, which

acknowledges that updates in different parts of the probability scale cannot be mathematically equated to one another. Yet, Shah et al. (2016) demonstrate that the fundamental problems with the scale are inherited in such analyses given that participants' responses will be noisy (see "comparisons with rational Bayesian predictions" above). Alternatively, some studies in economics have utilised highly-contrived "lottery" methods incorporating objective probabilities (Barron, 2018; Buser et al., 2018; Coutts, 2019; Eil & Rao, 2011; Ertac, 2011; Gotthard-Real, 2017; Möbius et al., 2014). These studies have produced markedly inconsistent results, with some observing an optimistic asymmetry (Eil & Rao, 2011; Möbius et al., 2014), and others finding no evidence thereof (Barron, 2018; Buser et al., 2018; Coutts, 2019; Ertac, 2011; Gotthard-Real, 2017).

Optimism research is of importance to researchers and practitioners alike. Yet, for there to indeed be a true optimism bias in belief updating, the evidence should not be able to be produced by rational, unbiased agents (e.g., Shah et al., 2016) or in cases where the variable upon which it depends — valence — is removed. The update method thus remains unfit for purpose, and assuming evidence produced by it to be solid is ill-advised. Yet, optimistic belief updating, along with optimism's other forms, may very well exist in the real-world. Our empirical results argue for an *absence of evidence*, rather than *evidence of absence*. More work would be needed to actively falsify the optimistic belief updating hypothesis. For example, is asymmetric updating explained away by the statistical artefact identified here, or does it co-exist with actual motivated updating? Nevertheless, our analyses suggest the foundational method that research is building upon continues to fail critical tests. Here, it has failed to display a consistent valence-dependence, an inherent attribute of optimistic belief updating's very definition.

2.2 Chapter conclusion

Psychological explanations for why people may form inaccurate beliefs in the context of a digital world centre around observations that individuals' cognitive processes favour desirable, identity-concordant information. This account of a psychological problem "in the mind" suggests that the ever-larger amounts of information granted online leads to increasingly skewed beliefs due to asymmetric sampling, reasoning, and updating. However, as the empirical results presented in this chapter emphasise, observing asymmetries in information processing may not necessarily constitute evidence of irrational, motivated cognition. For such asymmetries to have bearing on our overarching question of belief

accuracy in a digital world, they must be shown to be true, accuracy-cost-incurring bias by deviating from normative standards.

Needless to say, the results presented in this chapter only target the specific phenomenon of optimistic belief updating and do not address the myriad other paradigms that comprise the psychological account of inaccurate belief formation. But in a broad sense, the observation of asymmetric belief updating in response to neutral stimuli points to the methodological limits of psychology-oriented experimentation in artificial settings. For one, it is difficult to distinguish "errors" from systematic bias that incurs costs in the long-run with contrived experimental designs. And further, specifying the appropriate normative model for any given task is not only a non-trivial exercise, but one that frequently runs counter to experimenter intuitions — as is the case with Bayesian agents displaying "optimistically" asymmetric belief updating (U. Hahn & Harris, 2014; Shah et al., 2016). Elsewhere, these issues have also been shown to plague related research on politically motivated reasoning, where supposed evidence of biased reasoning could equally be produced by rational, Bayesian reasoning with differential prior beliefs (Tappin et al., 2020a, 2020b, 2020c). Taken together, these features of characteristic methodological approaches in psychology suggest that empirical evidence underlying the account of a psychological problem "in the mind" may be less conclusive than it seems. In order to supplement such evidence, it is necessary to consider the real-world environments in which the psychological processes involved in belief formation take place.

Chapter 3

A structural problem "in the world"

> The apparent complexity of our behaviour over time is largely a reflection of the complexity of the environment in which we find ourselves
>
> H. A. Simon, 1969, p. 53

Running alongside the account of a psychological problem "in the mind" is much research lamenting structural characteristics of online information environments as inherent threats to the accuracy of our beliefs. In this account of a structural problem "in the (digital) world" the emphasis is not on the cognitive capacities of human users, but on the design of new digital media and the ways in which it presents users with distorted information (e.g., Brady et al., 2017; Caplan & boyd, 2016; Ferrara et al., 2016; Lazer, 2015; Moore, 2019; Pariser, 2017; Persily & Tucker, 2020; Seifert, 2017; Shao et al., 2018; Sunstein, 2018; Vosoughi et al., 2018; Woolley, 2016). Specifically, this perspective points to three interacting features: (1) perverse incentives to propagate low-quality information driven by "attention economics," (2) the presence of manipulative bots, and (3) proprietary, content-curating algorithms that personalise and fragment individuals' information environments. As a result of these features, users' truth-seeking intentions are said to be impeded by the structure of the digital world itself.

Popularised by Goldhaber (1997) and Davenport and Beck (2001) (but also see H. A. Simon, 1971), the term "attention economy" refers to the observation that in the digital world the most scarce, valuable resource is human attention, rather than information or

any material capital. In this new economic paradigm it is argued that our attention is now effectively a currency, and the fortunes of businesses depend on their ability to capture and monetise the attention of their audience (Davenport & Beck, 2001).[1] While, in many ways, this has always been the case, the dynamics of the attention economy have seemingly been supercharged by new digital media. Collective attention spans have shortened, moving more rapidly between popular topics and events on social media (Lorenz-Spreen et al., 2019), and the digital trace data left behind by users leaves them open to microtargeted advertisements optimised for engagement (Moore, 2019). What does this mean for the accuracy of our beliefs? First and foremost, it suggests that the online information environments people increasingly rely upon to stay informed will be dominated by content that is attention-grabbing, regardless of its actual informational quality. For digital news organisations and content creators whose revenue relies on readership metrics (e.g., clicks, shares, channel subscribers) to sell advertisement space to third-parties, the long-term incentive to build a journalistic reputation has arguably been been supplanted by the myopic, profit-first incentive to "go viral." Unfortunately, the factors influencing the diffusion of information online identified in existing literature seem to be either uncorrelated or negatively correlated with conventional markers of credibility. These include, for example, emotion-laden language (Berger & Milkman, 2012; Brady et al., 2017; Ferrara & Yang, 2015; Hansen et al., 2011; Stieglitz & Dang-Xuan, 2013), negativity directed towards one's political out-group (Pew Research Center, 2017; Rathje et al., 2021), and news content that is novel, surprising, and flat out false (Vosoughi et al., 2018). Taken together, the implication of these findings is that the attention economics of the digital world reward the sharing of low-quality content in the name of "virality," and in turn inundate users with misleading information.

A second feature of online information environments that has been a cause for concern is the prevalence of so-called *bots*. In its purest form, a bot (derived from "robot") is a computer programme that functions as an autonomous agent on an online platform (Franklin & Graesser, 1996). However, the term has come to encompass a wide range of typologies including the fully automated (e.g., web crawlers and chatbots), semi-automated (e.g., social media accounts with scheduled postings), and manually controlled (e.g., sockpuppets and trolls) (Gorwa & Guilbeault, 2020). Whereas many instances for deploying bots online

[1] For example, the cost of running an advertisement on Twitter is determined by the number of user "engagements" received; instead of purchasing a premium Spotify account to stream music, you can listen for "free" by having your music interrupted by advertisements; in 2017, the co-founder of Netflix, Reed Hastings, explained that one of their main competitors is sleep (Hern, 2017).

are well-meaning or benign, such as web crawlers that automatically index massive loads of websites for search engines (Pant et al., 2004) or chatbots acting as virtual assistants (Sansonnet et al., 2006), they can have unintended effects on the sorting and spread of information. For example, search engine-supporting web crawlers typically index content based on solely on its "relevance," meaning that a user who enters an obscure, outdated, or otherwise problematic search query will not be presented with the most accurate and informative information, but with whatever content has certain keywords and high levels of viewer traffic (cf. *data voids*, Golebiewski & boyd, 2018). This feature leaves such bots vulnerable to being gamed, be it through the seemingly accepted business practice of "search engine optimisation" or by bad actors pushing disinformation around specific topics (Gorwa & Guilbeault, 2020; Woolley & Howard, 2016). At the centre of many recent studies, however, is the coordinated use of bots on social media, which can manipulate users by disrupting political discourse (Bessi & Ferrara, 2016), warping perceptions of consensus (Lerman et al., 2016), exacerbating social tensions (Stella et al., 2018; L. G. Stewart et al., 2018), and amplifying misinformation (Ferrara et al., 2016; Lazer et al., 2018; Shao et al., 2018; Vosoughi et al., 2018). While some studies evaluating the role of bots in online information environments have questioned their real-world consequence (e.g.,. Bail et al., 2020; Dunn et al., 2020; González-Bailón & De Domenico, 2021), the well-documented prevalence of malicious bots — and the ease with which anonymous and automated accounts can be set up — casts doubt over the informational integrity of the digital world.

A third feature of the digital world that has been cited for its undermining effect on the accuracy of people's beliefs is the ubiquitous use of proprietary algorithms for content curation. In concert with bots, algorithms provide an indispensable service by assisting users' navigation of content online, which would otherwise leave finding relevant, needed information an unmanageable task. However, a dark side to these algorithms emerges once their personalisation of content goes beyond the linkage structure of the web (as in Google's original PageRank system, Brin & Page, 1998) to include factors like users' browsing history and geographic location (Lazer, 2015). For instance, Le et al. (2019) observed that, based on browsing history alone, personalised Google search results tend to reinforce the presumed political preferences of the user. This finding demonstrates what Pariser (2017) dubbed the *filter bubble* effect (cf. *echo chamber* effect, Jamieson & Cappella, 2008), whereby algorithmic personalisation is said to entrench individuals in

their own, unique, self-affirming information environment online, thereby fragmenting the digital world into niches or "bubbles" with little interaction between them. Such an effect becomes particularly concerning in the context of social media and the aforementioned attention economy. Given that the business model of social media platforms relies on attracting and retaining user engagement to sell advertising, it is perhaps not surprising that the integrity of the proprietary, often opaque algorithms used to recommend content to platforms' users has been called into question (Pariser, 2017; Persily, 2017; Sunstein, 2018). For example, research has found that Twitter's default news feed algorithm unequally promotes friends' posts to users based on popularity metrics (Bartley et al., 2021), YouTube's recommendation algorithm presents increasingly radical content to users (e.g., Ribeiro et al., 2020; Tufekci, 2018; for a review see Yesilada & Lewandowsky, 2021), and Facebook's algorithmic curation conforms to users' ideology, albeit to a fairly small degree (Bakshy et al., 2015). However, in contrast to the original filter bubble hypothesis of Pariser (2017), several studies have pointed out that such filtering effects are prompted by users' choices — including whom to befriend and what content to "like" — to a much greater extent than the content-curating algorithms themselves (e.g., Bakshy et al., 2015; Chen et al., 2021; Hosseinmardi et al., 2021; Möller et al., 2018). Going further, some have downcasted the existence of the filter bubble effect altogether (e.g., Bruns, 2019; Dubois & Blank, 2018; Hannak et al., 2013).

Whereas few researchers would argue that the digital world's attention economics, prevalence of bots, and reliance on proprietary algorithms for content-curation do not pose plausible threats to users' ability to form accurate beliefs, it is widely-acknowledged that estimating their actual causal effects remains a difficult task. This is arguably due to the fact that, notwithstanding recent innovation in digital field experiments (e.g., Mosleh, Pennycook, et al., 2021) and algorithmic auditing techniques (e.g., Bartley et al., 2021; Sandvig et al., 2014), independent investigations of the digital world have largely been limited to non-experimental approaches. While, nowadays, researchers are able to collect vast quantities of digital trace data to power their studies, the nature of such data as observational, highly confounded, and gate-kept by the platforms they exist on means making the jump from data to meaningful scientific conclusions faces new hurdles (Lazer et al., 2021; Munger, 2019; Salganik, 2017; Tufekci, 2014; Wagner et al., 2021). In order to understand belief accuracy in a digital world, it thus seems crucial to remain critical of the data and methodologies that are being applied across the computational social sciences so

that they may be bettered.

In the remainder of this chapter, I present a study probing a high-profile finding that is said to evidence the effect of social media's perverse incentives to share emotional, outrage-inducing content, regardless of its informational quality (*moral contagion*, Brady et al., 2017). By testing the methodological limits of a conventional approach to the study of information diffusion on social media, I emphasise the need for deeper understanding of the analytic challenges faced if meaningful links between structural characteristics of the digital world, the integrity of online information environments, and the accuracy of people's beliefs are to be established.[2]

3.1 Reconsidering evidence of moral contagion in online social networks

In 2017, a study leveraging large-scale social media data presented evidence of a *moral contagion* effect, which seemingly corroborates mainstream concerns about attention economics online and the perverse incentives they impose (Brady et al., 2017). In the study, Brady et al. (2017) apply a dictionary-based text analysis procedure to quantify moral-emotional language in hundreds of thousands of tweets capturing the naturally-occurring communications of Twitter users. By then fitting a regression model and performing a series of robustness checks, they show that the mere presence of moral-emotional words increases messages' retweet counts by a factor of 20%, regardless of the messages' informational quality (Brady et al., 2017). The implications of this moral contagion phenomenon, where the exposure to moral emotions shapes the diffusion of information due to their attention-grabbing nature, are undoubtedly significant. Invoking morality in reasoning has previously been shown to harden existing belief structures, delegitimize authority, and, in extreme cases, dehumanize opposing perspectives (Ben-Nun Bloom & Levitan, 2011; Crockett, 2017). While injections of moral reasoning into discourse can be beneficial — providing shared identities and guiding ethical behaviour — the introduction of unnecessary moralization and its emotional underpinnings may jeopardize rational debate. It is for this reason that moral justifications carry weight in some domains but not others. For example, loading an argument with moral-emotional language might be an effective strategy in a debate over social policy and human rights, yet that same strategy is likely to be

[2]The empirical work presented in this chapter is based on a collaboration between myself, Nicole Cruz, and Ulrike Hahn, which has been published in *Nature Human Behaviour* (Burton, Cruz, et al., 2021). All relevant data and code has been made available on an OSF project page.

penalized in an argument over mathematics. However, if moral contagion is as widespread and domain-general as Brady et al. (2017) suggest, then it seems plausible that sentiments about where moralization is appropriate are changing as a result of the attention economics of the digital world. This also suggests that we are susceptible to new forms of political persuasion online. As Brady et al. (2017) conclude, "it seems likely that politicians, community leaders, and organizers of social movements express moral emotions...in an effort to increase message exposure and to influence perceived norms within social networks" (p. 7316). Beyond this substantive contribution, the authors also recognize the methodological implications of their study, because "in comparison with laboratory-based studies, the social network approach offers much greater ecological validity" (Brady et al., 2017, p. 7317).

Brady et al. (2017) is one example of what is an ongoing methodological shift across the social sciences (also see, e.g., De Choudhury et al., 2013; Garcia & Rimé, 2019; Rathje et al., 2021; Tumasjan et al., 2010), whereby statistical analyses of large-scale digital data traces — namely, social media data — form the basis for studies of human behaviour in the context of the digital world. But digital data traces produced by social media users are inherently noisy and high-dimensional. In contrast to the "custom-made" data generated via controlled experimentation, material harvested from online platforms is usually not created with research in mind (Salganik, 2017). Social media data can be ambiguous, confounded by proprietary algorithms and restricted access, and unrepresentative of wider populations, which may limit the generalizability of findings between platforms and between online and offline populations (Ruths & Pfeffer, 2014; Salganik, 2017; Tufekci, 2014). These documented observations may be less problematic if one's research objective concerns itself only with understanding behaviour on a given platform itself; however, in the absence of agreed upon methodological standards for handling social media data, the space for "researcher degrees of freedom" (Simmons et al., 2011) is particularly vast. This means that conclusions from analyses of observational social media data alone may face deeper issues, insofar as they are intended to teach us something about real human behaviour or meaningful effects of the digital world's design.

In this study, we probe the finding of moral contagion, illustrating possible methodological pitfalls that might be encountered when standard practices of null hypothesis significance testing are applied to large-scale social media datasets. How robust is correlational evidence from large-scale observational data? What inferences and generalizations

can be made from such evidence? Answering these questions seems crucial to make sense of the existing literature concerning the role of the digital world in shaping human behaviour.

3.1.1 Method

The diffusion of information in social networks has been likened to a biological pathogen, spreading from person to person through direct contact. For a behaviour, psychological state, or other condition to qualify as a simple social contagion, the probability of the condition being adopted by an individual should increase monotonically with the number of times that individual is exposed to said condition (Hodas & Lerman, 2014). In the case of moral contagion, moral-emotional words (e.g., *kill*, *protest*, *compassion*) are considered to be the "contagious" cue because their presence is presumed to be a central factor in an individual's decision to retweet (or diffuse) the message in which it is included. Based on this logic, moral contagion should be present in other corpora of tweets pertaining to contentious, politicised topics. To test this proposal, we recreated Brady et al.'s (2017) methodology and applied it to other Twitter corpora spanning a variety of socio-political issues and events.

Measuring language

As in Brady et al. (2017), we used a dictionary-based text analysis to quantify distinctly emotional ($N_{words} = 819$; e.g., panic, fear, heartwarming), distinctly moral ($N_{words} = 316$; e.g., fair, racism, solidarity), and moral-emotional language ($N_{words} = 72$; e.g., shame, victimize, disgust). Importantly, there is no overlap in the dictionaries, meaning that each tweet could be allocated three discrete scores forming three independent predictor variables. To ensure our scripts were accurately counting words and word stems, we performed a check in which we re-ran the scripts with a random sample of 10 word stems and 10 words and manually checked that the correct counts were displayed on a random sample of 20 tweets from each corpus that had at least one word/stem counted. By selecting a manageable number of tweets, words, and word stems, we were able to check for both false positives and false negatives and then simply scale up our scripts. We found that our scripts were accurately counting words and word stems, and the tweets included in each corpus were relevant to their respective topics.

Measuring diffusion

The key dependent variable in this study is message diffusion. Diffusion was calculated as the sum of a message's retweet count as captured in the metadata and the number of times that message's text appeared in a corpus. Identical messages were then collapsed into a single observation with other relevant metadata from the earliest posting (e.g., the number of followers a message poster has; whether the post included URLs, an image, or video media). This approach avoids penalising retweet chains, which are important indicators of diffusion on Twitter, while also accounting for unconventional retweets where a user copies and pastes someone's message rather than clicking the retweet button. With diffusion as our dependent variable and the three language measures as predictors, we then followed Brady et al. (2017) in fitting a negative binomial regression model with maximum likelihood estimation — to best handle the overdispersed count data being analysed (Hilbe, 2011) — to each dataset (henceforth referred to as the "main moral contagion model"). The presence of contagion was determined by exponentiating the regression coefficients of each predictor (i.e., distinctly emotional, distinctly moral, and moral-emotional language) to generate incidence rate ratios (IRR) — the most central measure being moral-emotional language's IRR. Note that as a ratio measure, IRRs greater than 1.00 signify a positive contagion effect (e.g., $IRR = 1.10$ suggests a 10% increase in diffusion), and vice versa.

Datasets

With the above measures we tested the influence of language use on message diffusion across six corpora of tweets that capture the naturally-occurring communications among users (see Table 3.1 for full descriptive statistics of each corpus. While no specific corpus or topic was initially targeted, certain criteria were employed. To be considered for this study, corpora had to contain Twitter data (i.e., tweet messages and retweet counts), contain messages written in English, and relate to a polarising or morally-charged real-world issue, event, or social movement.

Once retrieved, corpora were further narrowed by collapsing repeated messages into a single observation (as described in "measuring diffusion") and removing non-English messages. Since the pre-existing corpora did not include language identifying metadata, the `textcat` package (Hornik et al., 2013) was employed to extract English tweets in these instances. Additional preprocessing was done with the `tm` (Feinerer et al., 2015) and `tidyverse` packages (Wickham et al., 2019) prior to applying the dictionary-based

text analysis. This included converting all text to ASCII characters and removing retweet prefixes (i.e. "RT"), usernames, punctuation, and URLs. Observations in which no text remained after the preprocessing were removed from the analysis. All preprocessing and analysis was done in `R` and scripts are available on the public OSF project page. The following paragraphs describe the six corpora covered in this study.

COVID-19. For this corpus we collected tweets pertaining to the (ongoing at the time of writing) COVID-19 pandemic. Using the `rtweet` package (Kearney, 2019), we specified a search for English tweets including at least one of the following terms: #COVID-19, COVID-19, COVID19, covid19, COVID, covid, or coronavirus. Collected tweets were posted on 23-24 March 2020, a period in which nation-wide lockdowns were being put into effect across the globe. While the topic of infectious disease does not necessarily evoke feelings of morality or polarisation a priori, the COVID-19 pandemic has elicited highly contentious debate in political, scientific, and public spheres. For example, Reuters reported results of a poll showing that Democrats are about twice as likely as Republicans to say COVID-19 poses an imminent threat to the US (Heath, 2020), and researchers identified political polarisation as an important part of the social context that should be addressed in responses to COVID-19 (Van Bavel et al., 2020).

#MeToo. Our second corpus comprised of Twitter messages containing the #metoo hashtag was obtained from the data.world repository. The tweets were collected from the Twitter API between 29 November and 25 December 2017, little more than a month after the #metoo hashtag first appeared online in coordination with the "Me Too movement" (Turner, 2018). The "Me Too movement" is a movement against sexual harassment and assault. It was ignited by Hollywood sexual abuse allegations and has since become an international phenomenon garnering widespread media attention, support, and critique.

#MuellerReport. A third corpus was collected by using the #muellerreport hashtag to retrieve tweets from the Twitter API created between 23 and 25 March 2019 — the weekend during which US Attorney General William Barr released his summary of Special Counsel Robert Mueller's investigation into Donald Trump's 2016 presidential campaign. This corpus was of special interest because the Mueller Report has been a major source of controversy. While originally a non-polarised issue, the public opinion divided over time (Thomson-DeVeaux, 2019) meaning that moral-emotion could have plausibly played

a part in moralising conversations on Twitter.

2016 US Presidential Election. Our fourth corpus containing viral tweets (those with 1,000+ retweets) from the 2016 US Presidential Election was obtained from the Zenodo repository. The set of tweets was collected with the Twitter API and extracted messages that contained specific hashtags (#MyVote2016, #ElectionDay, and #electionnight) and/or user handles (@realDonaldTrump and @HillaryClinton) (Amador et al., 2017). This corpus was of special interest as it contained many "fake news" messages as coded by the curators, which one might expect to use especially morally- and emotionally-charged language to garner extra attention given the conclusions of Brady et al. (2017).

Post-Brexit. A fourth corpus containing unfiltered tweets and metadata from the morning that Brexit was announced was obtained from the Mendeley Data repository. These tweets were collected with NCapture from QSR and employed a tight temporal parameter so as to capture the public's reaction to the political event (Parker, 2017). Brexit refers to the result of the 2016 EU Referendum in the United Kingdom, and this dataset includes Twitter responses from across the globe.

#WomensMarch. Our sixth and final corpus with tweets containing the #womensmarch hashtag was obtained from the data.world repository. Using the Twitter API, 15,000 messages were collected that referenced the pro-women's rights, and effectively anti-Trump, protest that took place in the wake of the presidential inauguration on 21 January 2017 (Adhokshaja, 2017). The Women's March has since become a worldwide movement with annual marches in late January to non-violently protest for women's reproductive rights, LGBTQ rights, immigration and healthcare reform, as well as racial, gender, and religious equality.

3.1.2 Results

Out-of-sample prediction

Prior to analysing our corpora, we checked our model specifications by reanalysing Brady et al.'s (2017) cleaned data, which they have made available online. Their data focused on topical political issues in the United States: gun control ($n = 48,394$), same-sex marriage ($n = 29,060$), and climate change ($n = 235,548$). Using the Twitter API and sets of topic-related filter words (e.g., guns, gun control, and NRA for the gun control topic),

	COVID-19	*#MeToo*	*#Mueller Report*	*2016 US Election*	*Post-Brexit*	*#Womens March*
N	701,925	393,135	229,046	9,001	17,998	15,000
n	172,697	151,035	39,068	8,233	5,660	3,778
Diff. min.	0	0	0	1,001	0	0
Diff. max.	368,611	56,750	25,842	100,000	31,901	170,518
M Diff.	266.78 (4,152.64)	8.93 (222.63)	15.54 (312.87)	3,372.65 (5,222.76)	119.60 (923.57)	705.80 (4,983.11)
M moral-emo. words	0.23 (0.52)	0.20 (0.47)	0.18 (0.47)	0.16 (0.43)	0.09 (0.32)	0.16 (0.42)
M Moral words	0.49 (0.79)	0.26 (0.53)	0.47 (0.77)	0.33 (0.61)	0.22 (0.48)	0.28 (0.57)
M Emo. words	1.09 (1.24)	0.89 (0.97)	0.99 (1.17)	0.85 (1.02)	0.69 (0.90)	0.67 (0.83)
M XYZ count	2.61 (2.18)	1.84 (1.44)	2.43 (2.08)	1.68 (1.41)	2.28 (1.37)	1.53 (1.32)

Table 3.1: Descriptive statistics of each analysed corpus, including the minimum, maximum, and mean (M) diffusion ($Diff.$), and the mean count of moral-emotional words, distinctly moral words, distinctly emotional words, and Xs, Ys, and Zs ($XYZcount$). N refers to the total number of raw tweets included in the corpus (including duplicates, non-English tweets, and tweets with no text), and n refers to the number of clean, unique tweets analysed in the paper. For all means, standard deviations (SD) are reported in parentheses.

tweets and metadata were extracted between 30 October and 15 December 2015. Across the three corpora comprising 313,002 analysable tweets spanning three topics, our analysis reproduced their findings. Moral-emotional language was significantly associated with an increase in retweets in each corpus when covariates were controlled for (same-sex marriage, $IRR = 1.17$, $p < 0.001$, 95% $CI = 1.09, 1.27$; gun control, $IRR = 1.19$, $p < 0.001$, 95% $CI = 1.14, 1.23$; climate change, $IRR = 1.24$, $p < 0.001$, 95% $CI = 1.22, 1.27$), and in two out of three corpora when covariates were not controlled for (same-sex marriage, $IRR = 1.08$, $p = 0.059$, 95% $CI = 0.99, 1.18$; gun control, $IRR = 1.36$, $p < 0.001$, 95% $CI = 1.30, 1.42$; climate change, $IRR = 1.15$, $p < 0.001$, 95% $CI = 1.12, 1.17$). However, these results did not consistently generalize across the six corpora we analysed.

Taking Brady et al.'s main moral contagion model, as well as the nested single-variable model in which only moral-emotional language is used as a predictor, we found moral contagion to be present in only two of six corpora before controlling for covariates: COVID-19 tweets ($IRR = 1.15$, $p < 0.001$, 95% $CI = 1.11, 1.18$) and #MuellerReport tweets ($IRR = 1.28$, $p < 0.001$, 95% $CI = 1.16, 1.42$). In the four pre-existing corpora, moral-emotional language either had no significant relationship with message diffusion or had a negative effect where moral-emotional language predicted a decrease in diffusion (Ta-

ble 3.2). While we could not control for the same covariates as Brady et al. (2017) and were therefore unable to provide direct replications in the four pre-existing corpora due to missing metadata, we did so in the COVID-19 and #MuellerReport corpora (we do this to aid comparison with Brady et al.'s original results; however, we strongly caution against basing one's interpretation of these results on covariates – see section on "covariates, outliers, and the analytical multiverse"). Once Brady et al.'s (2017) chosen covariates were controlled for in the regression model to provide a direct replication of the original analysis, the significant association between moral-emotional words and message diffusion remained in the #MuellerReport tweets ($IRR = 1.27$, $p < 0.001$, 95% $CI = 1.16, 1.40$), but no statistically significant relationship was observed in the COVID-19 tweets ($IRR = 1.01$, $p = 0.320$, 95% $CI = 0.99, 1.04$).

The limits of correlational data

The inconsistent results of out-of-sample prediction tests point toward the limitations of purely correlational data. The inherent difficulty of distinguishing true causal contagion from confounding network homophily has been noted in detail elsewhere (e.g., Aral et al., 2009; Shalizi and Thomas, 2011). But large sets of observational data carry even more fundamental risks of spurious correlation and endogeneity. To demonstrate this, we conducted a follow-up analysis in the spirit of Hilbig (2010).

In his study, Hilbig (2010) re-evaluated conclusions made by Gigerenzer and colleagues (e.g., Gigerenzer, 2008; Goldstein & Gigerenzer, 2002; Marewski et al., 2010) that people employ heuristics in judgement and decision making tasks on the basis of correlational evidence alone. Specifically, Hilbig (2010) examined the recognition heuristic, which Goldstein and Gigerenzer (2002) define as: "if one of two objects is recognised and the other is not, then infer that the recognised object has the higher value with respect to the criterion" (p. 76). Seemingly straightforward evidence of the recognition heuristic in action is provided by asking different sets of individuals to identify which of two cities has a larger population. For example, Goldstein and Gigerenzer (2002) report that upon asking whether San Diego or San Antonio has a larger population, approximately two thirds of the Americans asked correctly identified San Diego as having the larger population, yet 100% of the Germans asked correctly identified San Diego. The reason for this, they argue, is that the Germans — who are presumably less knowledgeable on the topic of American cities — relied on the recognition heuristic. Since the Germans were more unfamiliar with

San Antonio, they successfully associated the recognisability of San Diego with a larger population size (Goldstein & Gigerenzer, 2002). However, Hilbig (2010) argued that this conclusion may be unfounded since recognisability is not the only available cue that might be used to make the judgement of population size. To do so, he introduced a humorously implausible *alphabet heuristic*, which stipulates that people can infer cities' population size by ranking the city names in alphabetical order and selecting the latter option (e.g., San **A**ntonio comes before San **D**iego). Using the same judgement task as Goldstein and Gigerenzer (2002), Hilbig (2010) shows that the alphabet heuristic is at least as useful as the recognition heuristic, and that the same data previously taken as evidence of the recognition heuristic in action could equally be taken as evidence of the alphabet heuristic in action. Of course, the purpose of this analysis is not to propose that people actually use the alphabet heuristic, but rather to critique the methods employed and conclusions made by Gigerenzer and colleagues.

Returning to the issue of moral contagion with the logic of Hilbig (2010), we created an absurd factor for illustrative purposes, what we call *XYZ contagion*, and tested whether the number of X's, Y's, and Z's included in messages' text predicted diffusion (note that we were unable to test for XYZ contagion in Brady et al.'s original data because their raw data did not include metadata retweet counts, which meant that our analysis scripts could not be properly applied). Our analysis found XYZ contagion to be present in four of our six corpora such that the presence of the letters X, Y, and Z predicted an increase in message diffusion: COVID-19 tweets ($IRR = 1.08$, $p < 0.001$, 95% $CI = 1.07, 1.08$), #MeToo tweets ($IRR = 1.13$, $p < 0.001$, 95% $CI = 1.12, 1.15$), #MuellerReport tweets ($IRR = 1.12$, $p < 0.001$, 95% $CI = 1.10, 1.14$), and the 2016 US Election tweets ($IRR = 1.01$, $p = 0.030$, 95% $CI = 1.00, 1.03$). While there was no positive relationship between the presence of X, Y, and Z and message diffusion in the #WomensMarch and Post-Brexit tweets, the finding that XYZ contagion performs well in a key test of robustness, out-of-sample prediction, demonstrates the potential of large-scale social media datasets to contain spurious correlations (Table 3.2; also see Appendix B.1.2 for a bootstrap resampling analysis).

In addition, we calculated Akaike Information Criteria (AIC) as measures of model adequacy and found that our model of XYZ contagion actually outperforms the main, multi-variable moral contagion model in two of the six corpora (Table 3.2). We further tested the XYZ contagion model against the single variable moral contagion model such

that the predictive value of the count of letters X, Y, and Z was compared to the count of moral-emotional words in isolation. This analysis revealed that the count of letters X, Y, and Z was in fact a better predictor of message diffusion than moral-emotional words in five out of six corpora, despite being nonsensical (Table 3.2).

	COVID-19	*#MeToo*	*#Mueller Report*	*2016 US Election*	*Post-Brexit*	*#Womens March*
N	172,697	151,035	39,068	8,233	5,660	3,778
Main Multi-Variable Moral Contagion Model						
IRR	1.15 [1.11, 1.18]	0.91 [0.88, 0.95]	1.28 [1.16, 1.42]	1.02 [0.98, 1.06]	0.89 [0.72, 1.13]	1.01 [0.77, 1.38]
p	<0.001	<0.001	<0.001	0.465	0.370	0.925
$\Delta_i(AIC)$	0.00	138.88	20.66	0.00	0.00	0.00
$w_i(AIC)$	>.9999	<.0001	<.0001	>.9999	.9995	0.9629
Single-Variable Moral Contagion Model						
IRR	1.19 [1.15, 1.23]	0.92 [0.89, 0.96]	1.40 [1.28, 1.55]	1.02 [0.98, 1.06]	0.81 [0.66, 1.02]	0.90 [0.68, 1.24]
p	<0.001	<0.001	<0.001	0.337	0.101	0.494
$\Delta_i(AIC)$	690.53	334.28	49.41	36.39	15.39	14.28
$w_i(AIC)$	<.0001	<.0001	<.0001	<.0001	.0005	0.0009
XYZ Contagion Model						
IRR	1.08 [1.07, 1.08]	1.13 [1.12, 1.15]	1.12 [1.10, 1.14]	1.01 [1.00, 1.03]	1.00 [0.95, 1.06]	0.89 [0.82, 0.96]
p	<0.001	<0.001	<0.001	0.030	0.998	0.011
$\Delta_i(AIC)$	386.72	0.00	0.00	32.67	18.56	6.87
$w_i(AIC)$	<.0001	>.9999	>.9999	<.0001	.0001	0.0362

Table 3.2: Negative binomial regression model results and comparisons. Incidence rate ratios (IRR) indicate the size of contagion effects in each dataset (for the main moral contagion model only the effect of moral-emotional language is reported), with 95% confidence intervals in brackets and corresponding p-values in the row below. For each model, the differences in AIC with respect to the best candidate is calculated, $\Delta_i(AIC)$, meaning that an $\Delta_i(AIC)$ equal to zero signals that the corresponding model is the best fit for the given dataset. AIC values are further transformed into Akaike weights, $w_i(AIC)$, which are the conditional probabilities that the model in question, i, is the best model given the data and the set of candidate models

Covariates, outliers, and the analytical multiverse

Out-of-sample prediction tests and model comparisons demonstrate how social media datasets may be susceptible to unfounded correlations. However, we need to consider the influence of outliers and covariates in more detail, which are indeed sensible and widely-recognised checks that can and have been put in place to guard against spurious results. But as we show next, in the context of social media data, neither of these are sufficient to solve the problems identified here, facing both methodological and conceptual

limitations.

Regarding outliers, the problem is that social media data are a typical case of fat-tailed distribution, and it is unclear how "outlier" should be defined. The prevalence of extreme values (e.g., a tweet garnering 100,000 retweets when the median is 0) is likely a constitutive feature of the dataset, rather than a bug or error to be neglected. Consequently, decisions on outliers are seemingly arbitrary. For example, consider a traditional psychology experiment measuring reaction times in the lab. Outliers in this case are readily identifiable: A reaction time that is ten times the mean indicates that a participant was not paying attention, had not read the instructions, or the data was entered incorrectly. Yet, in the domain of social media, there is no such judgement that can be made. That a message may be retweeted zero, one, or 100,000 times is in fact an intrinsic part of the paradigm. What does it mean if, in a study of message sharing, the top ten or one hundred most shared messages determine what statistical results are retrieved from a corpus of hundreds of thousands of messages? Are these observations to be excluded, or are they meaningful indicators of a recipe for going viral?

Covariates might be considered even more important. Indeed, there is a wide range of potential covariates that plague social media data, relating to both the content of messages and the accounts of message posters. Specifically relating to Twitter, it has previously been shown that the presence of hashtags and URLs in a message, the number of followers and followees a message poster has, and the age of the message poster's account all influence retweet rates (Suh et al., 2010). There are also questions around the potential need to account for the influence of automated and semi-automated bots (Kollanyi et al., 2016; Lazer et al., 2018; Ruths & Pfeffer, 2014). Despite existing literature highlighting these covariates, the controls that researchers put in place are often inconsistent, even when the hypotheses in question are relatively similar. For example, consider three studies investigating the role of emotion in message sharing on Twitter: Stieglitz and Dang-Xuan (2013) control for the number of hashtags a tweet contains, the presence of URLs, the number of followers a message poster has, and the number of tweets a user has posted during the sampling period; Ferrara and Yang (2015) excluded tweets containing URLs or media (i.e., a photo or video); and Brady et al. (2017) control for the number of followers the message poster has, whether media or URLs are present in a tweet, and whether the message poster is "verified" (a status indicating that the user is a celebrity or public figure). Not only do these studies identify different covariates, but they also

control for them in different ways. For instance, where Ferrara and Yang (2015) excluded tweets containing URLs and media, Brady et al. (2017) input these covariates as binary variables in a regression. While each study's controls are certainly defensible, this points to another problem: any given set of controls will not be exhaustive and there is no agreed upon standard for what controls must be made to separate a publishable finding from a coincidental statistic; and even more fundamentally, against what ground truth could these methodological practices be evaluated?

Taken together, the ambiguity surrounding outliers and covariates highlights the increased "researcher degrees of freedom" (Simmons et al., 2011) in analyses of social media data. That is, researchers must make many arbitrary analytical decisions when collecting, processing, and analysing the data. While this is not unique to social media data or any type of digital data traces, it may be especially consequential in this context. To investigate how decisions on covariates and outliers influence the moral contagion and XYZ contagion results, we conducted specification curve analyses (SCA) (Simonsohn et al., 2020) on our three largest corpora (COVID-19, #MeToo, and #MuellerReport). In short, SCA is a way to make analytic flexibility transparent by running all justifiable model specifications (e.g., what covariates to control for, what data subsets to analyse, what independent variable to assess, etc.), and then making joint inferences across the results of all these specifications (Simonsohn et al., 2020). SCA is closely related to the concept of a "garden of forking paths" (Gelman & Loken, 2014) and "multiverse analysis" (Steegen et al., 2016), and serves to clarify the fragility or robustness of statistical findings by identifying which analytical choices they hinge on.

For our SCA, we consider the results of negative binomial regression specifications with either the number X's, Y's, and Z's or the number of moral-emotional words in a tweet predicting diffusion, with or without controlling for covariates, and with or without the removal of (arbitrary) increments of outliers (the tweets with the top 10, 100, and 1,000 diffusion counts). The covariates we consider are the number of distinctly moral words, the number of distinctly emotional words, and the number of characters in a tweet, the number of followers a message poster has, whether the message poster's account is verified, and whether media, URLs, and hashtags are present (binary). Because the #MeToo corpus is a preexisting dataset that was not collected by the authors of the present study, not all of the relevant metadata is included and only some of the covariates could be considered. Figure 3.1 displays the outcome (unstandardised regression coefficient) of

each model specification (x-axis) when fitted to each corpus as three, vertically-aligned points corresponding to the independent variable, covariates, and outliers accounted for (y-axis). We then plot these outcomes as specification curves in Figure 3.2, visualising how negative, positive, and nonsignificant moral contagion effects can be retrieved, depending on the chosen corpus and model specification (also see Appendix B.1.3 for SCA applied to Brady et al.'s original corpora). The specification curves also allow for comparative evaluations between moral contagion and XYZ contagion. Namely, we observe that while the median regression coefficient across model specifications with moral-emotional words as the independent variable is positive in the COVID-19 ($n = 40$, median $B = 0.18$, $SD = 0.08$) and #MuellerReport corpora ($n = 39$, median $B = 0.10$, $SD = 0.13$), it is negative in the #MeToo corpus ($n = 28$, median $B = -0.02$, $SD = 0.08$). Meanwhile, the median regression coefficient across model specifications with the number of X's, Y's, and Z's as the independent variable is positive in all three corpora (COVID-19, $n = 39$, median $B = 0.07$, $SD = 0.05$; #MeToo, $n = 28$, median $B = 0.04$, $SD = 0.06$; #MuellerReport, $n = 39$, median $B = 0.05$, $SD = 0.05$). This could be taken to suggest that the XYZ contagion effect is, if anything, more stable than the moral contagion effect across theoretically-justifiable model specifications in the three corpora addressed here. Of course, we strongly doubt that the letters X, Y, and Z play a central role in shaping the diffusion of information on Twitter. What our analyses show, however, is that the evidence of moral contagion provided by Brady et al. (2017) seems to be virtually indistinguishable from our atheoretical XYZ contagion effect, regardless of whether it is framed as a causal or correlational effect.

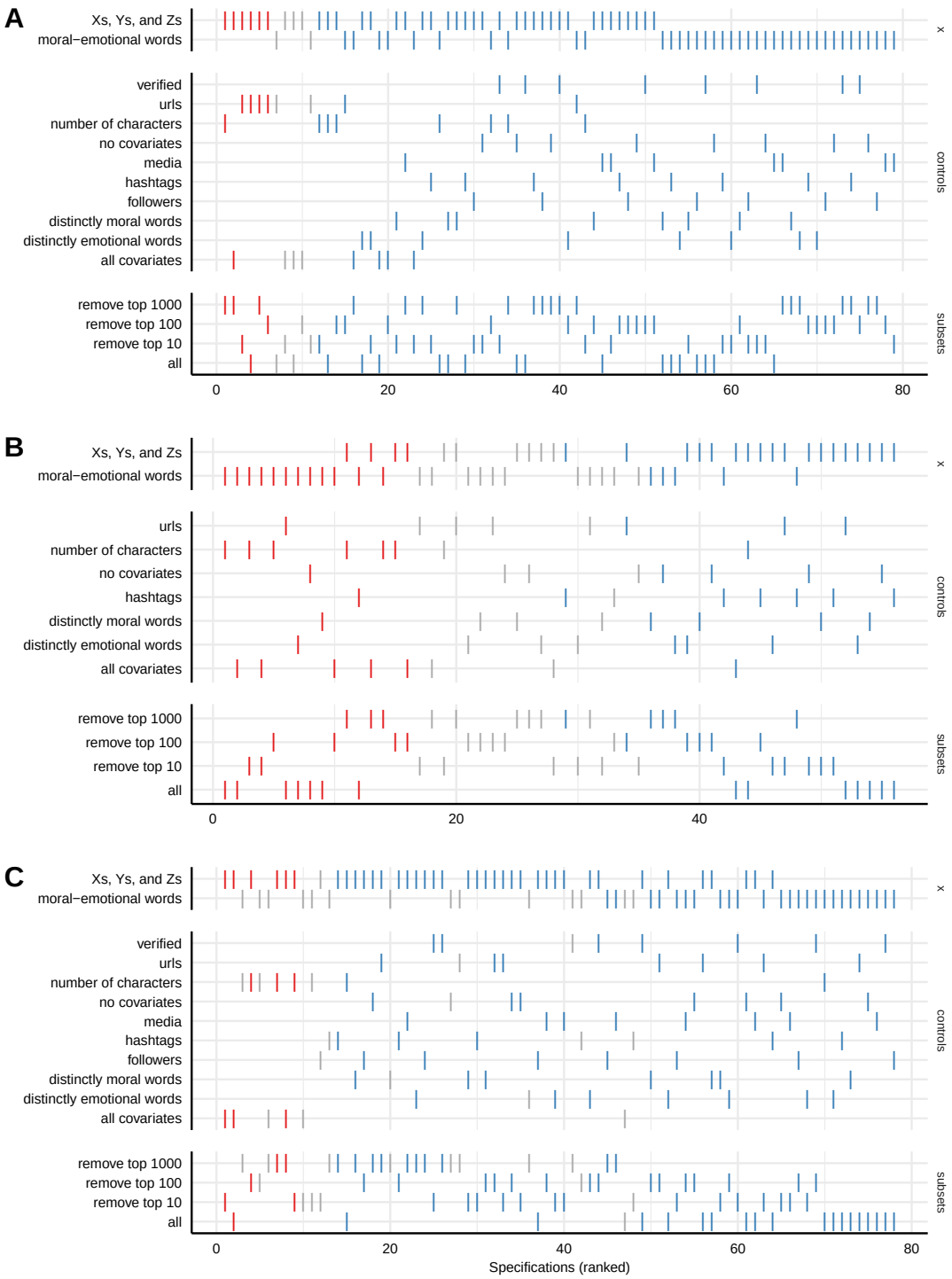

Figure 3.1: Qualitative results of specification curve analyses (SCA). (A) COVID-19 corpus. (B) #MeToo corpus. (C) #MuellerReport corpus. Each possible model specification (x-axis) is represented by three vertically-aligned points corresponding to the outliers removed and covariates and independent variable included in the negative binomial regression equation (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. There are fewer specifications (56) in the #MeToo SCA (B) because metadata on some covariates was absent. Of the 80 possible specifications for the COVID-19 and #MuellerReport data, one specification was excluded from the COVID-19 SCA and two specifications were excluded from the #MuellerReport SCA because these algorithms did not converge.

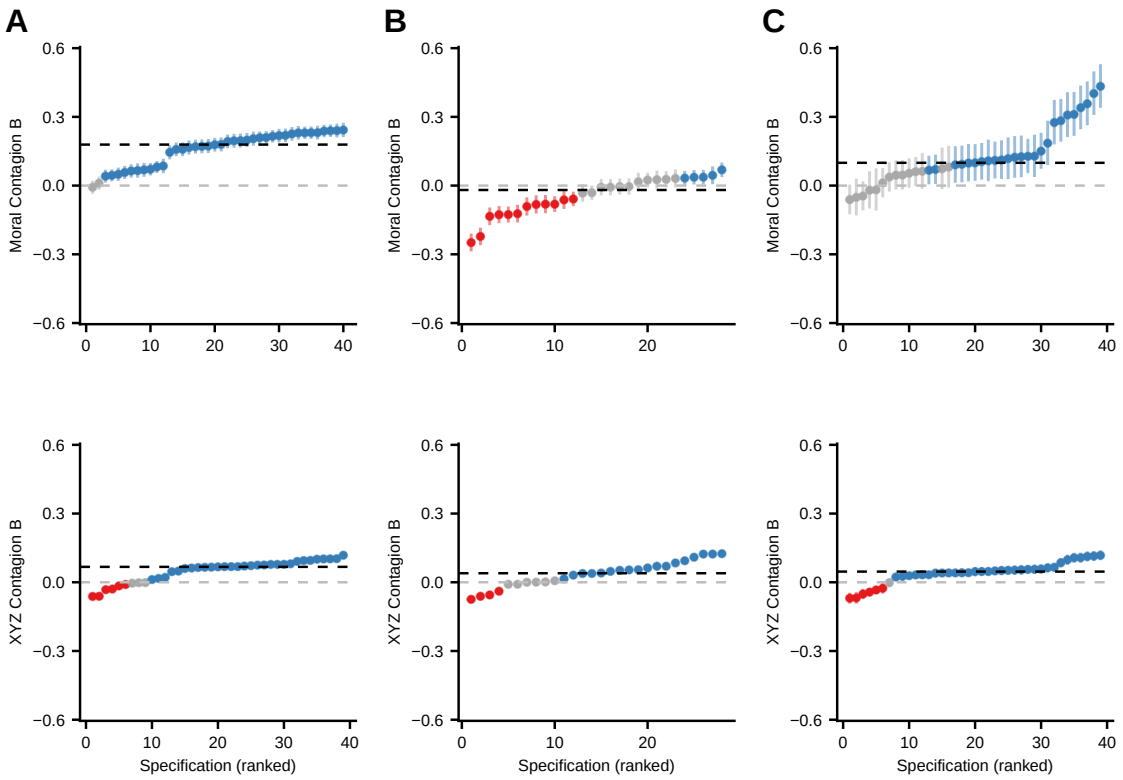

Figure 3.2: Specification curves for moral contagion (top plots) and XYZ contagion (bottom plots) effects. (A) COVID-19 corpus (moral contagion, $n = 40$, median $B = 0.18$, $SD = 0.08$; XYZ contagion, $n = 39$, median $B = 0.07$, $SD = 0.05$). (B) #MeToo corpus (moral contagion, $n = 28$, median $B = -0.02$, $SD = 0.08$; XYZ contagion, $n = 28$, median $B = 0.04$, $SD = 0.06$). (C) #MuellerReport corpus (moral contagion, $n = 39$, median $B = 0.10$, $SD = 0.13$; XYZ contagion, $n = 39$, median $B = 0.05$, $SD = 0.05$). Each model specification (x-axis) is represented by a single point indicating the resulting unstandardised regression coefficient and vertical bars indicating 95% confidence intervals (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. There are fewer specifications in the #MeToo corpus (B) because metadata on some covariates was not recorded. Two specifications in the #MuellerReport corpus (C) and one specification in the COVID-19 corpus (A) are excluded because the algorithm did not converge.

3.1.3 Discussion

Out-of-sample prediction, model comparisons, and SCA question the evidence that a meaningful moral contagion effect has been identified on Twitter. To be clear, moral contagion may very well exist, as lab-based work seems to support (Brady et al., 2020), but our results caution against basing such a conclusion on large-scale, observational data alone. They also caution against the idea that such data provide stronger evidence than lab-based studies due to greater ecological validity. Not only does our analysis challenge the moral contagion hypothesis, but, perhaps most worryingly, it shows that current methodological standards can support patently absurd models, such as the XYZ contagion. One limitation of our analysis is that it is indeed possible to hypothesise why the

XYZ contagion might exist after seeing our results (e.g., perhaps X's, Y's, and Z's are attention-grabbing because they are infrequently used). However, there is no reason to believe that the presence of these letters is causally relevant a priori and there is currently no evidence to suggest such a theory. While one might expect such causally irrelevant factors to be randomly distributed, there is no guarantee that they do not exhibit some artefactual, spurious correlation with the target phenomenon of interest. Yet crucially, the analyst has no way of telling in advance what state of affairs they will face. For analyses of digital data traces collected from social media platforms to effectively inform our understanding of human behaviour in the context of the digital world, we make two suggestions for future research utilising such data: (1) do not settle for correlational evidence alone, and (2) make the consequences of analytic flexibility transparent.

Both the fragility of moral contagion and the seeming "success" of XYZ contagion in our data highlight how the conclusions afforded by standard statistical procedures, like linear regression models and significance testing, are limited when applied to large-scale social media datasets. While correlational evidence can be informative (e.g., for predictive purposes), this overlooks the crucial point of why findings such as the moral contagion phenomenon are typically interesting. Arguably, the correlational findings of moral contagion are interesting precisely where they seem to be indicative of a meaningful causal relationship (Rohrer, 2018). This is why it would be highly unlikely that any academic journal would publish a paper on XYZ contagion. It thus seems necessary for researchers interested in understanding human behavior to either triangulate correlational findings with data from controlled experimentation (e.g., Dehghani et al., 2016; Mooijman et al., 2018); apply alternative statistical techniques, such as structural equation modelling (SEM) (e.g., Westfall & Yarkoni, 2016) or directed acyclic graphs (DAGs) (e.g., Rohrer, 2018); or use other design methods for causal inference with observational data, if large-scale observational data is to be relied upon.

Our analysis also highlights the need to address analytic flexibility when utilising social media data. The SCA results presented show how justifiable decisions on covariates and outliers are empirically consequential, capable of giving rise to directly conflicting results on the same predictive relationship in the same dataset. Yet our demonstration only scratches the surface of the analytical "multiverse" that researchers must navigate when handling social media data. For instance, text-as-data research such as that examined in the present work requires heavy data preprocessing, for which there is no agreed upon

standard. In analysing tweets, one may or may not decide to employ stemming, lemmatization, remove "stop words," remove usernames and hashtags, disambiguate homographs with part-of-speech tagging (e.g., "be *kind* to your dog" vs. "what *kind* of dog is that?"), and so on. While seemingly mundane, these preprocessing decisions can lead substantively different interpretations of the data to emerge (Denny & Spirling, 2018). The same can also be said of feature engineering. For example, the decision to use a dictionary (or bag-of-words) approach versus a machine learning strategy for text classification can lead to different measurements of moral expressions within the same corpus, and classification performance can vary across contexts (Hoover et al., 2019). While a logistic regression model fitted with Brady et al.'s (2017) dictionaries seems to be a good predictor of human judgements of moral expression in tweets related to #MeToo ($AUC = 83.2\%$), it is essentially as good as random when applied to a corpus containing hate speech messages ($AUC = 51.7\%$) (see Appendix B.1.1 for more analysis of the Moral Foundations Twitter Corpus; Hoover et al., 2019). At present, the focal strategy for managing analytic flexibility is pre-registration, but this seems ineffective for the issues raised here. Pre-registering an analysis plan might ensure researchers commit to a chosen analytical pathway and guard against "p-hacking," but given the underlying multiverse of divergent but theoretically-defensible results, this is not enough to guarantee that the specific results retrieved are ultimately informative. While there is indeed a longstanding tradition in the social sciences to consider alternative model specifications as a check of robustness, methods like SCA (Simonsohn et al., 2020) should be encouraged so as to make this tradition more transparent and exhaustive, and to better display exactly which analytical decisions are responsible for potentially conflicting results.

While the use of observational "big data" is relatively new to many social science domains, the obstacles outlined here are not particularly novel in other fields. For instance, large longitudinal datasets have been integral to the study of public health and epidemiology, where it has previously been shown that the standard use of regression models can produce implausible findings, such as statistics that suggest acne, headaches, and height are "contagious" (Cohen-Cole & Fletcher, 2008). If analyses of social media and other digital data traces are to contribute to understandings of human behaviour, it seems unlikely that the standard practices of null hypothesis significance testing and robustness checks will suffice. As demonstrated here, the inferences and generalizations that can be made from purely correlational findings in observational social media data can sometimes

be remarkably fragile.

3.2 Chapter conclusion

Explanations of why people may form inaccurate beliefs due to structural features of the digital world emphasise the interplay of attention economics and proprietary tools for content-curation and manipulation — namely, bots and algorithms. This account of a structural problem "in the (digital) world" argues that despite the increased accessibility of information, the design of online information environments leads to users being presented with misleading signals. However, as the investigation presented in this chapter demonstrates, conventional methodological tools propping up this account may not be providing the meaningful insights that many believe them to be. In contrast to the contrived experimental methods that dominate psychological accounts of a problem "in the mind," researchers focused on this account have looked to exploit users' naturally-occurring digital trace data. While utilising such "ready-made" data seemingly side-steps the issues of ecological validity and statistical power that have plagued classic psychology studies, our results highlight how this computational social science approach faces its own, poorly understood challenges. As such, further methodological innovation seems required for structural features of the digital world to be confidently identified as having meaningful, causal effects on the accuracy of people's beliefs.

Chapter 4

Engineering digital tools to support belief accuracy online

> Human rational behaviour is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor
>
> H. A. Simon, 1990, p. 7

In spite of the methodological challenges highlighted in Chapters 1 and 2, a natural question that arises when reading the literature around belief accuracy in a digital world is whether past findings can inform new ways of supporting people's information processing with technological tools. Given that online information environments are human-made, is it possible that they could be re-imagined in a way that mitigates people's cognitive limitations? Or, could those same structural features cited for their information-distorting effects be re-designed as tools with the specific purpose of enhancing people's truth-seeking ability? Approaching questions such as these offers a route towards not only reducing potential harms of the digital world, but towards realising its early promises for a more informed, more engaged public. Indeed, efforts to design and engineer digital tools to support belief accuracy online are already underway. These efforts duly recognise the metaphor introduced by H. A. Simon (1990) in the quote above and have sought out ways to re-align structural features "in the (digital) world" with our understanding of psychological mechanisms "in the mind." By and large, these efforts take their shape as either (1) interventions to assist people's information processing in existing online infrastructures, or (2) purpose-built civic technologies that provide novel online spaces for deliberation and

decision making.

Digital tools that function as interventions aim to provide simple prompts to users so that their information processing might be made easier, either by making small changes in the user interface (cf. *choice architecture*, Kozyreva et al., 2020; Thaler & Sunstein, 2008) or by promoting basic digital competencies. The most straightforward instantiation of this is fact-checking, whereby a third-party (e.g., Snopes, Full Fact) verifies the veracity of claims being reported. Whereas fact-checking has long been applied to mainstream news outlets, new digital media poses extra challenges. For one, fact checks will only be valuable to users if the users readily encounter them, which is not a given in the decentralised information environments online. Even if an online platform accepts and integrates official fact-checking labels [as is the approach taken by Facebook (n.d.)], the scale and structure of the digital world means the task of verifying every disputed claim is impossible for any traditional organisation, even if supported by automated, AI-driven tools (for an overview of automated fact-checking at Full Fact, see Corney, 2021). With this in mind, it has been proposed that online platforms enable "crowdsourced" fact-checking — where users themselves provide ratings of information quality that are then aggregated — as a cost-efficient, scalable solution that leverages the participatory nature of new digital media (Allen et al., 2021; Pennycook & Rand, 2019). Yet, as major social media platforms begin pilot testing such measures (e.g., *Birdwatch*, Twitter, 2021), other researchers argue that community-based approaches to fact-checking are vulnerable to new forms of manipulation (e.g., Coscia & Rossi, 2020; Pröllochs, 2021; for counter evidence see Epstein et al., 2020). Relatedly, there is also existing literature documenting how interventions like fact-checking that directly correct or "debunk" fallacious claims may not effectively change users beliefs (e.g., the *continued influence effect*, Desai et al., 2020; Johnson & Seifert, 1994; Lewandowsky et al., 2012; Wilkes & Leatherbarrow, 1988; the *backfire effect*, Nyhan & Reifler, 2010; Peter & Koch, 2016; for counter evidence see Ecker et al., 2020; Lee, 2021; Swire-Thompson et al., 2020; T. Wood & Porter, 2019), and may, in some cases, actively increase the sharing of low-quality information (Mosleh, Martel, et al., 2021).

These technological and psychological challenges for fact-checking in the digital world have encouraged the development of more domain-general interventions for assisting users' information processing online (for overviews, see Kozyreva et al., 2020; Lorenz-Spreen et

al., 2020)[1]. Perhaps most popular is the application of *nudges*, which involve influencing people's behaviour by making small alterations to the choice architecture (Mirsch et al., 2018; Thaler & Sunstein, 2008; Weinmann et al., 2016). For example, Pennycook et al. (2021) show how simply shifting individuals' attention to accuracy by sending them an unsolicited message asking them to rate the veracity of a headline can counteract the perverse incentives to spread low-quality information on social media (also see Pennycook et al., 2020; Roozenbeek et al., 2021). Similarly, Lewandowsky et al. (2017) advocate for so-called *technocognition*, or the development of cognitively-inspired features for online information environments. For instance, Fazio (2020) shows how incorporating "friction" into social media platforms' sharing functions (i.e., forcing users to pause, reflect, or make extra clicks) can reduce the spread of misinformation, which both WhatsApp (2019) and Twitter (2020) have already begun implementing. Finally, in a less paternalistic approach, Hertwig and Grüne-Yanoff (2017) suggest *boosting*, which aims to foster people's competencies to navigate the digital world on their own. Like nudges, boosts can take many forms such as pop-ups with fast-and-frugal trees or visualisations of information sharing cascades to aid users' reasoning (Lorenz-Spreen et al., 2020). The most popular instantiation of boosting to date involves the gamified "inoculation" of users, whereby users are encouraged to adopt the perspective of a disinformation campaigner so as familiarise themselves with common manipulation techniques used online (Basol et al., 2020; Maertens et al., 2020; Roozenbeek & van der Linden, 2019). Altogether, digital implementations of each of these tools — fact-checking, nudging, technocognition, and boosting — has potential to improve the experience of information online for users, and in turn support their ability to form accurate beliefs. Still, it could be argued that these intervention-style tools simply address symptoms of the "post-truth malaise" rather than root causes, because they accept the existing infrastructures and economics of the digital world. This in turn leaves one to wonder what an alternative, socially responsible digital world might look like.

In recent years, there has been increasing interest in the development of purpose-built civic technologies that leverage structural features of the digital world for social good. Such

[1]Since the focus of this chapter is on digital tools, I have excluded interventions such as digital literacy education (e.g., A. M. Guess et al., 2020; Weinmann et al., 2016; Wineburg & McGrew, 2019), "self-nudging" practices (Center for Humane Technology, n.d.; Reijula & Hertwig, 2020), and policy-level regulation [e.g., the General Data Protection Regulation (GDPR), which nudges individuals to protect themselves against algorithmic personalisation by mandating that online platforms use opt-in defaults for obtaining users' consent to process personal data (European Parliament, 2016)]. While each of these has an important role to play in supporting belief accuracy online, they are out of scope for the present research.

technologies have been built for wide-ranging functions, from participatory budgeting to data-driven urban planning to crowdfunding platforms (Jungherr et al., 2020; Saldivar et al., 2019). However, within our focus of supporting belief accuracy online, there are two particularly relevant examples. The first example is *online prediction markets*, which are online platforms built to collect forecasts of real-world events from their userbase. By rewarding users for accurate forecasts, prediction markets are able to harness the "wisdom of the crowd" by aggregating large quantities of individual forecasts to generate highly-accurate collective predictions for consequential events like political elections and scientific breakthroughs (Arrow et al., 2008; Surowiecki, 2005; Wolfers & Zitzewitz, 2004). While prediction markets need not be hosted online, the digital world opens up new opportunities for their design and application. For instance, whereas prediction markets traditionally involved the exchange of contracts or "futures" that would yield monetary payments for correctly forecasting real-world outcomes, contemporary online prediction markets (e.g., Metaculus) have implemented reputation-based incentives similar to what is seen on social media. In addition, the digitalisation of prediction markets enables both the rapid generation of data visualisations and more effective outreach to a global userbase, which may ultimately serve to engage and inform more users and observers alike.

The second example of how belief accuracy can be supported with civic technology is *online deliberation tools*, which aim to provide virtual spaces for structured, equitable information exchange. In their most basic form, online deliberation tools are built as collaborative interfaces where individual users share arguments with one another on a pre-specified topic for some collective objective (e.g., drafting policy priorities or reaching a consensus). However, recent innovations have seen the introduction of data-driven back ends where, for instance, algorithms operate on users' inputs to enhance the quality of desired outcomes. The most well-known example of this type of tool was mobilised in Taiwan, where the Polis platform was used to bring citizens and policymakers together online for machine-mediated debates in a process labelled "virtual Taiwan" (vTaiwan). Tasks at hand included identifying policy concerns around how to regulate the ride-sharing company, Uber, and finding consensus on the politically-contentious decision of whether Taiwan should share a timezone with mainland China (Miller, 2020). In short, what Polis provided was an online platform where thousands of individuals could submit statements and vote in (dis)agreement with each other's positions, whilst, in the back end, a clustering algorithm organised and visualised users' responses in a way that discounted divisive

statements and highlighted common ground between opinion groups. As a result of this digitally-supported deliberation, Taiwanese citizens were able to transparently participate in government decision making, thereby ensuring that all stakeholders were informed and engaged, and that the most accurate (i.e., true to public opinion) collective outcomes were achieved. Despite successes like vTaiwan[2] demonstrating how the digital world can be reclaimed to support belief accuracy and decision making, experimental studies identifying the causal effects of such tools on accuracy-related outcomes, and how to improve them, are noticeably absent from existing literature (for a review of online deliberation research, see Friess & Eilders, 2015; Strandberg & Grönlund, 2018). Nevertheless, the development of online prediction markets and deliberation tools point towards promising opportunities for supporting belief accuracy by re-appraising and re-designing online information environments.

In the remainder of this chapter, I present an exploratory study introducing a novel digital tool for supporting belief accuracy and decision making in online social networks.[3]

4.1 Rewiring online social networks to enhance collective decision making

The increasing digitalisation of society has renewed interest in "wisdom of the crowd" effects, where the collective judgement of a group is more accurate than the judgements of individual experts or the individual group members themselves (Condorcet, 1785; Galton, 1907; Grofman et al., 1983; Surowiecki, 2005). Not only do the new means for information exchange and aggregation provided by the digital world promise more informed individuals, but also ready access to larger, wiser crowds, as is demonstrated by modern applications like online prediction markets (Arrow et al., 2008; Wolfers & Zitzewitz, 2004), crowdsourcing (Howe, 2006), and digital democracy tools (Morgan, 2014; J. Simon et al., 2017). In the present work, we draw from existing literature on wisdom of the crowd effects and judgement aggregation to design, deploy, and evaluate a new digital tool for supporting collective belief accuracy and decision making: *rewiring algorithms*.

[2]For other examples see Swae, MIT's Deliberatorium, and Stanford's Online Deliberation Platform.

[3]The following work is based on a project funded by Nesta's Centre for Collective Intelligence and subsequent collaborations between myself, Abdullah Almaatouq, M. Amin Rahimian, and Ulrike Hahn, which has been presented at the 9th ACM Collective Intelligence Conference (Burton, Hahn, et al., 2021) and the 43rd Annual Meeting of the Cognitive Science Society (Burton, Almaatouq, et al., 2021). All data and code has been made available on a GitHub repository.

Social influence, network structure, and collective estimation

The earliest results on wisdom of the crowd effects in collective estimation tasks assumed that individuals' judgements are made independently, meaning that their errors are uncorrelated and cancel out in aggregate (Condorcet, 1785). However, this independence assumption often goes unmet in the real world because people communicate with or otherwise influence one another. Past research on the effects of social influence in collective estimation tasks has produced seemingly contradictory findings. On one hand, there is evidence that social influence indeed undermines crowd wisdom by causing individuals' judgements to become correlated (U. Hahn et al., 2019; Lorenz et al., 2011; Muchnik et al., 2013); while on the other, there are studies that report an increase in collective accuracy following social influence (Almaatouq et al., 2020; Becker et al., 2017; Becker et al., 2019; Gürçay et al., 2015).

Formal results that incorporate the possibility of non-independence provide a potential explanation of these seeming contradictions (e.g., Ladha, 1992; Page, 2008). Such results show that social influence is neither inherently beneficial nor inherently detrimental to crowd wisdom; instead its effects depend on whether the benefits of communication to individual accuracy outweigh the detrimental effects of non-independence on collective accuracy. The logic of this is made clear in the Diversity Prediction Theorem, which states that collective error squared is the difference between the average individual error squared and the diversity of the individuals' judgements (Page, 2008). While providing a mathematical guarantee that the collective estimate will always be more accurate, in terms of error squared, than the average individual's as long as there exists some diversity in the group, this theorem formalises how social influence can be both good for collective accuracy (if it leads to an increase in average individual accuracy) and bad (if it leads to too much of a decrease in diversity). Whether social influence will increase or decrease collective accuracy for any given group thus depends on which one of these duelling effects is greater.

To provide predictions for when groups will benefit from social influence, recent research has turned towards studying how different social network structures affect collective accuracy (e.g., Almaatouq et al., 2020; Becker et al., 2017; U. Hahn, Hansen, et al., 2018; U. Hahn et al., 2019; Jönsson et al., 2015). Because social network structures delineate the paths through which social influence can be exerted in a group, it follows that different structural characteristics will feature in determining whether the net effect of social

influence will be beneficial for collective accuracy. For example, high levels of connectivity and free-flowing information can lead to "excess correlation" (i.e., correlation between individuals that is not accuracy inducing, Jönsson et al., 2015), high levels of centralisation can lead to certain individuals wielding excessive influence over the network (Becker et al., 2017), and a lack of structural plasticity can prevent networks from effectively responding to feedback about individuals' performance (Almaatouq et al., 2020).

Rewiring algorithms for collective accuracy

A reading of the literature linking network structure and collective accuracy begs the question: can we build optimal social network structures for eliciting the wisdom of the crowd? Despite the abundance of knowledge on the relationship between network structure and collective accuracy, strategies for exploiting network structure to increase collective accuracy remain under-explored. While there may be considerable difficulties in manipulating the structure of social networks in the analog world, the digital world provides new opportunities. Just as algorithms have already been used to mediate the information presented to online social networks (Lazer, 2015) and to identify influential nodes in social networks (Wei et al., 2018), it seems plausible that algorithms could be used to rewire the structure of online social networks to boost the wisdom of crowds.

In this work, we explore the viability of *rewiring algorithms* — programmable rules for manipulating who communicates with whom — as a tool for supporting collective belief accuracy and decision making online. Specifically, we develop and test three candidate algorithms and evaluate their effects on the collective accuracy of estimates made by communicating social networks.

4.1.1 Modelling and simulations

We first employ agent-based modelling and simulation to efficiently operationalise the parameter space and prototype different algorithm designs. Our modelling framework uses networks of 16 simulated agents who are tasked with judging a single binary hypothesis (i.e., each agent can favour either 0 or 1, with exact beliefs falling between these points). Such judgements readily map on to a broad range of real-world scenarios: assessing the truth or falsity of proposition, deciding whether or not to vote for a political candidate, or predicting whether or not a future outcome will occur.

Our model is initiated by first sending vectors of binary evidence to each agent, which

they integrate with a starting prior of 0.5 via Bayes' theorem. This procedure serves to simulate how individuals would have accrued their own independent knowledge on a given topic, rather than entering a discussion with a purely indifferent prior of 0.5. To represent a population of individuals with varying knowledge about the hypothesis at hand, we vary the amount of evidence each agent receives such that some individuals may be more familiar with or knowledgeable on a given hypothesis. We additionally vary the quality of the evidence sent to the agents by introducing two parameters: *sensitivity*, the probability of receiving positive evidence when the hypothesis is true (i.e., the so-called "hit rate" familiar from signal-detection theory), and *specificity*, the probability of receiving negative evidence when the hypothesis is false (i.e., the so-called "correct rejection rate"). These parameters allow us to model "kind" environments where true positive and true negative evidence is prevalent and a majority of the population is already nearly certain of the truth, as well as less favourable environments where true evidence is rare, and the beliefs possessed by the population are more widely distributed.

Once their initial estimates are assigned, the agents communicate with one another across a randomly generated network structure over the course of four discrete time points, t=1,2,3,4. At each time point, each agent i revises their estimate in light of those communicated by their network neighbours according to a DeGroot belief updating rule (Becker et al., 2017):

$$R_{t+1,i} = \alpha_i \times R_{t,i} + (1 - \alpha_i) \times \bar{R}_{t,j \in N_i}, \tag{4.1}$$

where $R_{t+1,i}$ is the agent's revised estimate following communication; $R_{t,i}$ is the agent's current estimate; $\bar{R}_{t,j \in N_i}$ is the average current estimate of the agent's network neighbours; and α_i and its complement $(1 - \alpha_i)$ represent the weight that the agent places on its own estimate versus those of its peers, respectively. Following the empirical analysis of belief revision in Becker et al. (2017), each agents' α at any given time point is determined by the following regression equation:

$$\alpha_i = 0.75 - 0.05\epsilon_i + \mathcal{N}, \tag{4.2}$$

where ϵ_i is the agent's absolute error, and $\mathcal{N}$ is Gaussian noise with $\mu = 0$ and $\sigma = 0.06$. This stochastic process means that there is a modest association ($r \approx 0.21$) between accuracy and resistance to social influence among our agents (Becker et al., 2017).

Network conditions

Of particular interest to the present work is how different network conditions perform in the general modelling framework outlined above. Here we consider collective accuracy in four conditions: static networks (i.e., unchanging network structure), and networks to which we apply one of three candidate rewiring algorithms. For *static* networks (our control condition), the initial, randomly generated network structure does not change and each agent communicates with the same other agents at each time point. However, in our three experimental conditions, we introduce rewiring algorithms that add and/or remove connections between agents at each time point so that certain agents are exposed to the beliefs (estimates) held by certain other agents. We specifically consider three such algorithms: a *mean-extreme* algorithm, a *polarise* algorithm, and a *scheduling* algorithm. See Appendix C.1 for visual schematics of each network condition.

The mean-extreme algorithm aims to increase the average accuracy of individuals in a network by directing social influence towards individuals with potentially erroneous, outlying estimates. The algorithm first calculates the mean estimate in a network at a given time point and identifies which side of the scale midpoint (0.5 on a 0-1 probability scale) the network's mean estimate lies. If the network's mean estimate is less than the midpoint, the algorithm identifies the agent with the lowest estimate and adds directed, outgoing ties to the three agents with the highest estimates. If the network's mean estimate is greater than the midpoint, the algorithm identifies the agent with the highest estimate and adds directed, outgoing ties to the three agents with the lowest estimates. This procedure effectively brings the estimates of the outliers closer to the mean.

The polarise algorithm aims to maintain the diversity of estimates in a network and prevent a potentially biasing homogenisation. It first identifies the two most extreme agents on either side of the current distribution of estimates (i.e., the agent with the highest estimate and the agent with the lowest estimate) and cuts all incoming ties to these agents so as to preserve their beliefs from social influence. Then, the influence of these extreme agents is increased by granting each of them two directed, outgoing ties to "core" agents. These core agents are the four individuals with the median estimates in the network (e.g., in a 16-agent network, the agent with the lowest estimate receives an outgoing tie to the agents with the 7th and 8th lowest estimates, and the agent with the highest estimate receives two outgoing ties to the agents with the 9th and 10th lowest estimates). The net effect of this procedure is that the diversity of beliefs (measured as

variance) is increased by ensured both extreme, "polar" sides of the belief spectrum are heard.

The scheduling algorithm differs from the mean-extreme and polarise algorithms in that it prescribes (or "schedules") a network structure of intermixing dyads, irrespective of individuals' estimates. Specifically, the algorithm pairs agents at each time point such that no agent speaks to the same agent twice, but each individual will have the opportunity to be in possession of all the available information in the network by the end of four rounds of communication. In this way, scheduled networks will have achieved a maximum diversity in interactions — each dyad at each time point will consist of two individuals sharing information received from individuals in the network that the other has not interacted with; the algorithm prevents any redundant interactions from taking place. However, for this algorithm to function it assumes that each individual effectively fully communicates all information they possess and fully integrates all information communicated to them by their peer at each time point. This algorithmic approach offers an alternative for situations where access to individuals' current estimates at each time point is not available.

Simulation results

Following 500 iterations in nine different information environments (i.e., factorially combining sensitivity = {0.2, 0.4, 0.9} and specificity = {0.2, 0.4, 0.9}) in which four matched networks are simulated (i.e., one of each network condition starting from an identical initial network), we assess collective accuracy by calculating the squared error of the mean estimate post-communication, henceforth referred to as collective error squared (CES). In addition to *CES*, we also calculate the average individual error squared ($AIES$) and diversity, measured as variance (VAR), present in each network as a way of better understanding each algorithm's effects in the context of the Diversity Prediction Theorem (Page, 2008).

Figure 4.1 displays the results of these simulations by showing the difference between matched static and experimental networks on each measure in each possible information environment. This visualisation shows that the algorithms' effects vary across information environments. For example, consider the panels containing the results where sensitivity and specificity are symmetrically high ($sensi = 0.9$, $speci = 0.9$). In such information environments no algorithm is able to substantially influence collective accuracy because agents in the network are able to form accurate beliefs based on their independently ac-

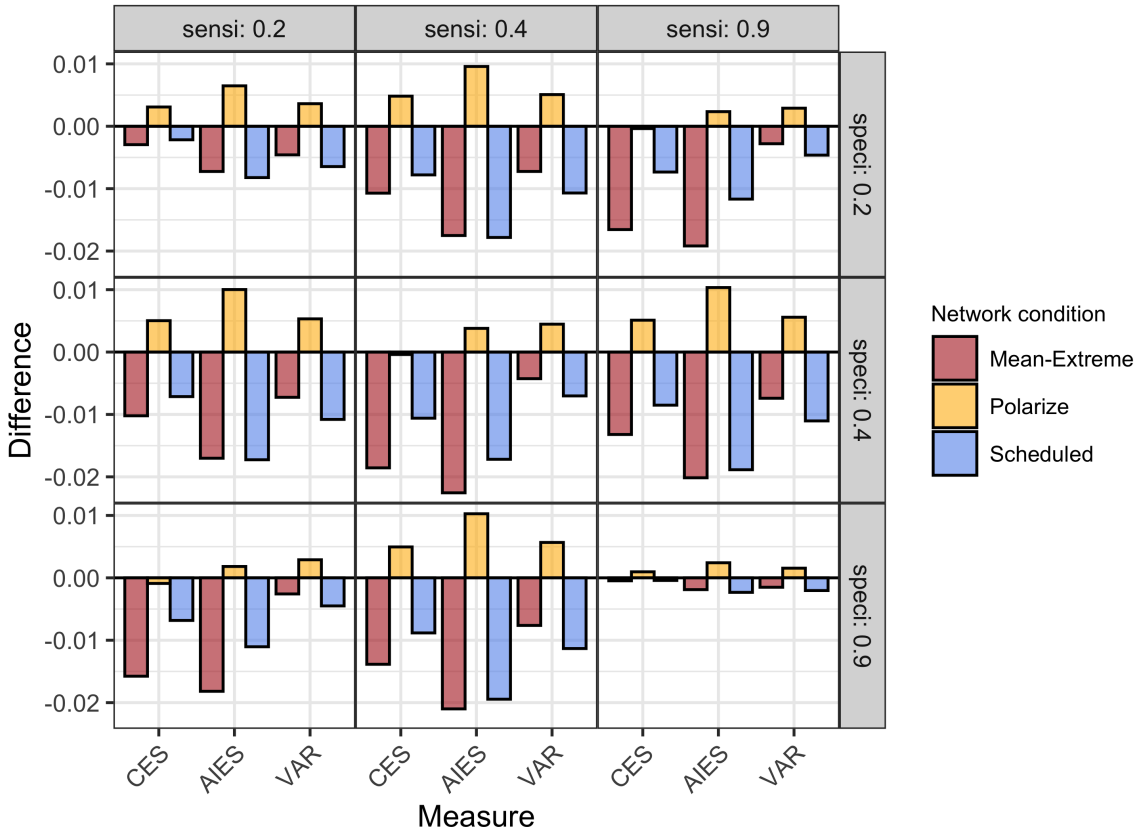

Figure 4.1: The simulated effects of each algorithm (network condition) on collective error squared *CES*, average individual error squared *AIES*, and belief variance *VAR* averaged across 500 iterations per panel. Y-axis values indicate the mean difference on a given measure as compared to a matched static network. A mean falling below zero indicates that the intervention resulted in a decrease of a given measure and vice versa.

quired knowledge, leaving little room for communication to improve the collective estimate. However, in each of the other information environments the mean-extreme and scheduling algorithm improve collective accuracy (displayed here as decreased CES), with varying degrees of magnitude. When viewed in conjunction with the impact of the intervention on $AIES$, it can be deduced that these two algorithms succeed by improving the average individual accuracy at the cost of diversity (displayed here as decreased VAR). In contrast, the polarise algorithm aims to improve collective accuracy by increasing the variance of beliefs at the cost of individuals' accuracy. However, this algorithm displays adverse effects on collective accuracy in these simulations. The failure of the polarise algorithm here seems attributable to two aspects in our modelling: the use of unbiased, optimal agents and the failure to sufficiently balance the increase in individual error with an increase in variance. The unbiased, optimal agents simulated have the ability to distinguish "anti-reliable" evidence (U. Hahn, Merdes, et al., 2018), meaning that before any communication takes place, the mean belief in the network is favourable and the distribution

of beliefs is skewed towards the truth regardless of the information environment imposed. Thus, broadcasting the extreme estimates to the median agents, who would otherwise converge towards the favourable mean estimate, will necessarily steer those receiving the erroneous extreme away from the truth. However, real human groups may possess biases that our simulated agents do not reflect, in which case the effects observed here may differ. Indeed, instilling a pre-existing bias in our model by assigning each agent a starting prior to 0.1 when the truth is 1, changes the results such that the mean-extreme algorithm often decreases collective accuracy and the polarise algorithm more frequently increases accuracy, albeit only slightly (Figure 4.2).

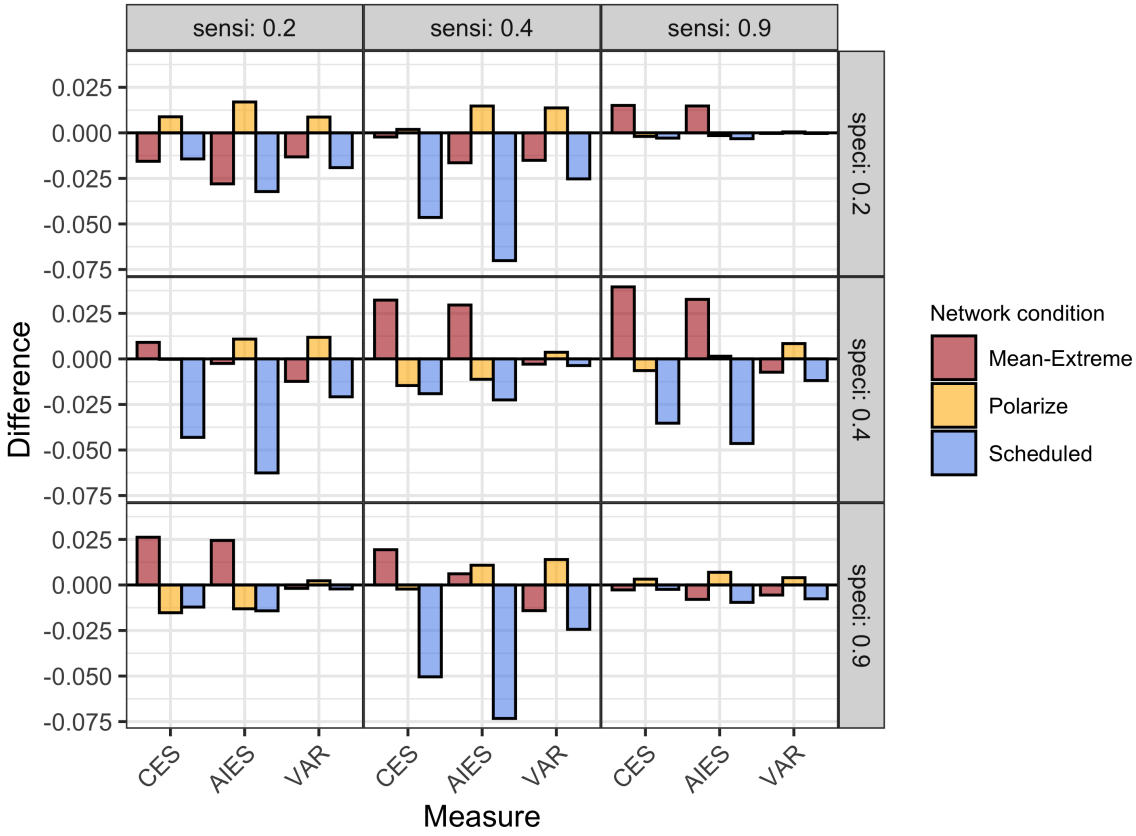

Figure 4.2: The simulated effects of each algorithm (network condition) on collective error squared *CES*, average individual error squared *AIES*, and belief variance *VAR* averaged across 500 iterations per panel when agents start with priors of 0.1 and the truth is 1. Y-axis values indicate the mean difference on a given measure as compared to a matched static network. A mean falling below zero indicates that the intervention resulted in a decrease of a given measure and vice versa.

Beyond the use of optimal agents, there are also other features of our modelling that should be taken into account when considering the robustness of these simulation results. For instance, by using the DeGroot belief updating rule specified in Equation (4.1) and Equation (4.2) we have assumed that agents' are equally receptive to influence from all

other agents, regardless of their beliefs. Yet, experimental work has shown that human belief updating may be egocentric when receiving advice from others (e.g., Volzhanin et al., 2015; Yaniv, 2004; Yaniv & Milyavsky, 2007). That is, people might discount information communicated to them by their peers if that information is too far away from their initial belief. If this is indeed the case, then the effects of the rewiring algorithms presented here may be limited since the algorithms deliberately connect agents in a network whose beliefs may be extremely opposing (as in the mean-extreme algorithm) or whose beliefs are on the edges of the network's belief spectrum (as in the polarise algorithm). Relatedly, as was explored to some degree in Figure 4.2, different distributions of initial beliefs in a network can influence the effects of each rewiring algorithm due to their underlying mechanics. Figure 4.3 shows how the simulation results change depending on the shape of the initial belief distribution, be it normally distributed around 0.5 (i.e., a network of generally uninformed or uncertain agents), log-normally distributed with a skew towards the truth (i.e., a network where most agents possess accurate information), or log-normally distributed with a skew away from the truth (i.e., a network where most agents possess inaccurate information or a pre-existing, erroneous bias, as in Figure 4.2). In conjunction with Figure 4.2, Figure 4.3 suggests that the effects of each rewiring algorithm may be context-dependent, which is explored in greater detail later in this chapter.

Next, we proceeded to test each of the rewiring algorithms with actual human social networks in an online multiplayer experiment where participants were tasked with predicting the probability that various near future events would occur.

4.1.2 Online multiplayer experiment

For our empirical study, we built an online multiplayer experiment with the Empirica software (Almaatouq, Becker, et al., 2021). This type of "virtual lab" approach allows for flexibility in the design of both a front end user interface and an experimental back end, where we could implement our rewiring algorithms. An anonymised preregistration for this study can be accessed here: https://aspredicted.org/BTJ_DKH.

Method

We recruited participants ($N = 704$) aged 18–69 ($M = 34.28$, $SD = 9.87$) via Amazon's Mechanical Turk crowdsourcing platform. Our sample size was determined by how much research funding was available for this study, and ended up being smaller than the esti-

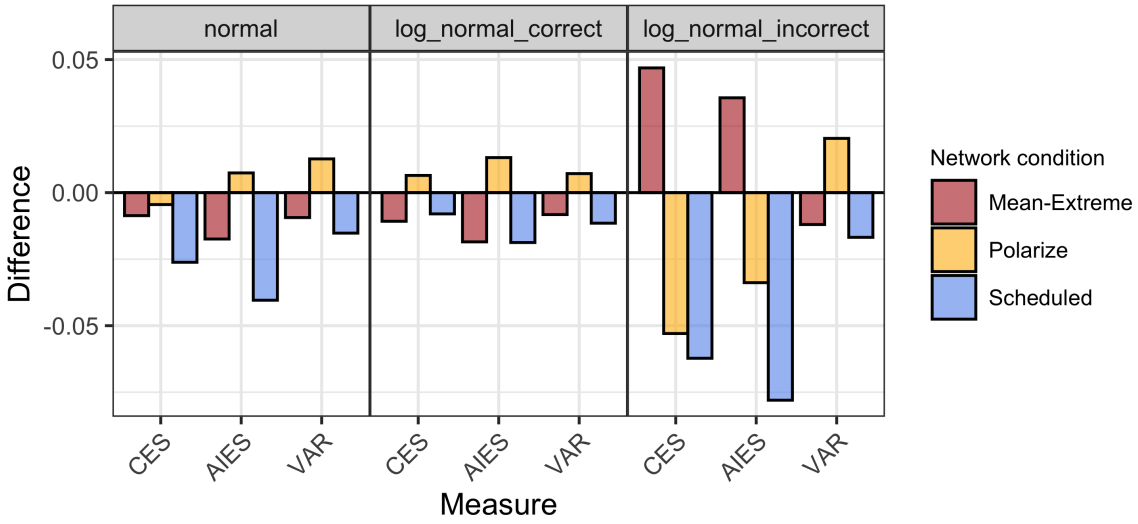

Figure 4.3: The simulated effects of each algorithm (network condition) on collective error squared *CES*, average individual error squared *AIES*, and belief variance *VAR* averaged across 500 iterations per panel when agents' initial beliefs are normally distributed around 0.5 (normal), log-normally distributed with a skew towards the true alternative (log_normal_correct), and log-normally distributed with a skew towards the false alternative (log_normal_incorrect). Y-axis values indicate the mean difference on a given measure as compared to a matched static network. A mean falling below zero indicates that the intervention resulted in a decrease of a given measure and vice versa.

mated sample size of 1,280 participants because we had not foreseen the need to use a considerable amount of funds on pilot testing our experiment. Participants were assigned into 16-person networks in one of the four network conditions (static, mean-extreme, polarise, or scheduled) and tasked with a "Collaborative Prediction Game" that consisted of ten rounds with five stages each. Each round of the game involved predicting the probability of one near future event occurring in reality (see Table 4.1 for the list of events and outcomes), with participants first providing a probabilistic prediction and short rationale for their prediction independently, and then proceeding through four stages of social exchange (or deliberation) where each participant would view the responses of their network neighbour(s) and revise their own prediction and rationale (see Figure C.5 for screenshots of the user interface). Each stage was limited to 60 seconds to prevent idle individuals from stalling the group and the entire study took approximately 50–60 minutes. Participants were given a base payment of $7.25 and given financial incentives for collective accuracy: 2x pay for the top three most accurate networks, 1.67x pay for the fourth through sixth most accurate networks, and 1.33x pay for the seventh through ninth most accurate networks. A total of 44 networks completed the study, 11 per treatment.

The four network treatments in the empirical study were identical to those simulated with our agent-based model, described in the previous section. Participants in *static*

ID	*Event*	*Outcome*
uk_covid	In the UK, the rolling seven-day average of COVID-19 deaths per day will go above 900 between 1-14 February 2021.	0
youtube_subs	There will be at least ten YouTube channels with more than 63.1 million subscribers on 8 February 2021.	1
biden_approval	Joe Biden's approval rating with be higher than 55% after three weeks as US President.	0
us_uk_vax	On 1 February 2021, the US will have administered more COVID-19 vaccination doses per 100 people than the UK.	0
bitcoin	Bitcoin will be valued at less than $30,000 on 8 February 2021.	0
super_bowl	Both teams in this year's Super Bowl will score more than 20 points.	0
us_climate	The US will rejoin the Paris Climate Agreement by 8 February 2021.	1
sp500	The S&P 500 will close higher on 8 February 2021 than it did on 31 December 2020.	1
epl	Liverpool FC will be leading the English Premier League on 7 February 2021.	0
americas_covid	The WHO will report more than 1 million COVID-19 deaths in the Americas by 8 February 2021.	1

Table 4.1: Events predicted by participants in the "Collaborative Prediction Game" experiment. An outcome of 1 indicates the event occurred in reality, and an outcome of 0 indicates the event did not occur in reality.

networks (the control condition) were placed in a randomly generated small-world network structure for each round and this network structure remained unchanged over each stage of deliberation. Participants in the *mean-extreme*, *polarise*, and *scheduled* treatments followed an identical procedure, but their network neighbours were subject to change between stages of deliberation, as determined by the given rewiring algorithm.

Experimental results

Our analyses of the empirical data focus on the accuracy of the collective, mean responses of each network pre- and post-communication. In particular, we asked the following three questions: (1) How did networks' average collective error squared (CES) differ between treatments post-communication? (2) How did communication affect CES within each network, between treatment? (3) How did the different rewiring algorithms influence networks' collective confidence calibration?

To address the first question we followed the procedure we preregistered as the main analysis, which involved a linear mixed effect model with each groups' average collective error squared (CES) across all events predicted as the dependent variable, the network treatment as a fixed effect, and random intercepts by group (Figure 4.4A). This analysis suggests that there is no significant effect of the rewiring algorithms on collective accuracy

($F(3, 436) = 0.78$, $p = 0.503$), meaning that, on average, networks to which a rewiring algorithm was applied did not achieve lower CES post-communication as compared to static networks on average[4]. However, this analysis does not account for certain key confounding variables — namely, the initial network structure and initial predictions in each network. While we could explicitly control for these in our modelling and simulation work by starting each iteration with perfectly identical networks, it was not possible to match these variables across treatments in the empirical study because each participant only completed the study one time, in one particular network, and in one particular treatment.

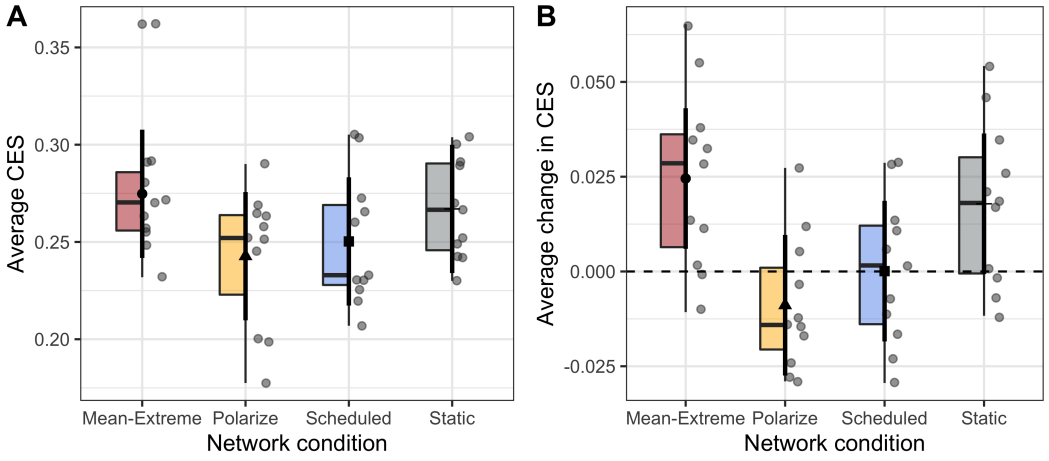

Figure 4.4: Results of linear mixed effect models. Boxplots and small points in the background display the spread of the raw data, and large shaped points indicate the model prediction with 95% confidence intervals represented by thick vertical bars. (A) Model with each groups' average collective error squared (CES) as the dependent variable, network treatment as a fixed effect, and random intercepts by group. (B) Model with each groups' average change in CES as the dependent variable (i.e., the difference between post-communication CES and pre-communication CES), network treatment as a fixed effect, and random intercepts by group.

In addressing the second question we conducted an unregistered analysis to side-step the potential confounding effects of initial network structure and initial predictions by evaluating the effect of communication *within* each network. That is, instead of directly comparing the accuracy of networks' collective predictions post-communication between treatments, we compare the *change in accuracy* between each network's prediction pre- and post-communication. Upon re-fitting our linear mixed effect model with networks' change in CES as the dependent variable, we find a significant treatment effect ($F(3, 436) = 2.72$,

[4]With this same linear mixed effect model specification, we also observed no statistically significant treatment effect when absolute error or square root error are used to measure collective accuracy. There was also no significant treatment effect on average individual accuracy when measured as squared error, square root error, or absolute error. However, there was a statistically significant treatment effect on variance, such that individuals in networks mediated by the polarise algorithm produced more diverse predictions post-communication than any other network condition ($F(3, 426) = 5.31$, $p = 0.001$).

$p = 0.044$) that suggests networks mediated by our polarise algorithm were more likely to benefit from communication, whereas communication was detrimental to mean-extreme and static networks, and neither beneficial nor detrimental to scheduled networks (Figure 4.4B). This result is encouraging because it suggests that the polarise algorithm not only prevented deliberation from leading groups astray through deleterious social influence, but the algorithmic mediation actually led groups towards more accurate predictions than those that would have been produced by aggregating the individuals' pre-communication predictions. However, this result is not robust to other loss functions for measuring collective accuracy: we observed statistically insignificant results when applying this model specification with change in collective *square root error* ($F(3, 436) = 1.02$, $p = 0.385$) and collective *absolute error* ($F(3, 436) = 1.54$, $p = 0.20$) entered as the dependent variable.

Finally, we followed our preregistration and conducted an exploratory analysis to examine the confidence calibration of networks' collective predictions pre- and post-communication. In the context of binary predictions, such as predicting whether future events will or will not occur, calibration refers to the ability to assign an appropriate degree of confidence or certainty to one's prediction (Fischhoff et al., 1977). For example, if an event's true, objective probability of occurring is 75%, then a group whose collective prediction is 90% would be considered overconfident in their judgement, whereas a group with a collective prediction of 60% would be considered underconfident, regardless of whether the event ultimately occurs in reality. While our previous analyses evaluated collective error based on the binary observable outcome of each predicted event, accurate calibration of collective predictions might be desirable in some circumstances. For instance, if a group of intelligence analysts were predicting whether our not an individual has plotted an imminent attack, overconfidence in a judgement that they did not pose a threat (e.g., a prediction of 10% when there is an actual probability of 30%) could lead the analysts to allocate insufficient resources towards monitoring the individual. To assess how our different rewiring algorithms might affect the calibration of collective predictions, we "binned" the networks' predictions by rounding them down to the nearest tenth decimal place (e.g., 0.12 and 0.19 both become 0.1) and calculated the proportion of events in each bin that occurred. If a network's predictions are perfectly calibrated, we would expect the proportion of events that occurred in each bin to match the the bin value (e.g., did 10% of the events that a network predicted to have a 10% probability of occurring actually occur in reality?). Figure 4.5 displays the results of this analysis by plotting the collective calibration of net-

works in each treatment pre-communication and post-communication side-by-side. What we find is that, despite all networks being well-calibrated pre-communication, the process of communication seems to have affected calibration differently depending on the network condition. Based on the calibration curves in Figure 4.5, it appears that communication in static network structures lead to overconfidence in collective predictions, whilst the polarise and scheduling algorithms mitigated this effect, and the mean-extreme algorithm exacerbated it. However, a closer examination of where the individual points fall suggests this conclusion is not straightforward. For instance, complete overconfidence would be reflected on the plot as a step function where every objective probability less than 0.5 would map onto a subjective probability of 0, and every objective probability greater than 0.5 would map onto a subjective probability of 1. In other words, overconfidence is indicated by undue extremity in predictions on the probability scale: predictions that should be more uncertain (closer to 0.5) tend to be closer to either 0 or 1. Yet the change in calibration curves shown in Figure 4.5 seem driven by random variations in the points plotted, rather than these points indicating unduly extreme predictions per se.

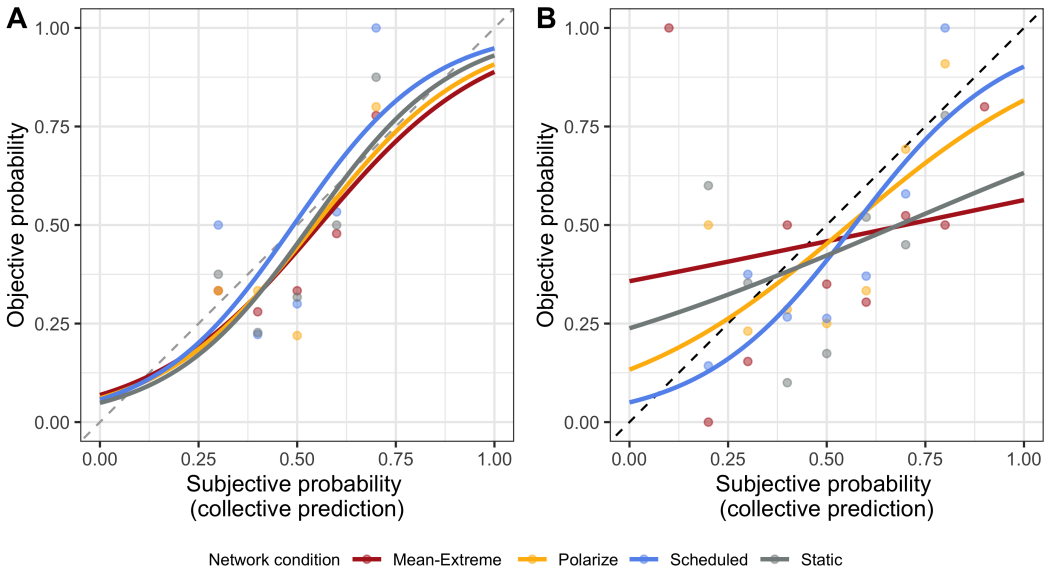

Figure 4.5: Calibration of collective predictions. Predictions (i.e., subjective probabilities) are "binned" by rounding down to the nearest tenth decimal place, and the objective probability for each bin is calculated as the proportion of events in that bin that occurred in reality. Perfectly calibrated predictions would fall on the dashed diagonal line. In cases where only one prediction fell in the corresponding bin, the data has been excluded because these data are exceptionally noisy and could only fall on either 0 or 1. (A) Pre-communication calibration. One prediction from the static condition, one prediction from the scheduled condition, and two predictions from the mean-extreme condition have been excluded due to being alone in its corresponding bin. (B) Post-communication calibration. One prediction from the static condition, one prediction from the scheduled condition, and one prediction from the mean-extreme condition has been excluded due to being alone in its corresponding bin.

To supplement Figure 4.5, an additional analysis related to overconfidence was performed where we tallied up the proportion of collective predictions in each treatment that became more extreme (or confident) in their initial position. Put simply, how often did networks' predictions become closer to 100% if their initial prediction was greater than 50%, or closer to 0% if their initial prediction was less than 50%? Across the 110 collective predictions within each network condition (11 groups per condition $\times$ 10 events), we found that 77.27% of the static networks', 72.72% of the mean-extreme networks', 67.27% of the scheduled networks', and 65.45% of the polarise networks' predictions became more extreme post-communication. While this analysis is strictly exploratory and does not distinguish between accuracy-improving and accuracy-degrading changes in confidence, it further suggests that the network conditions influence calibration in different ways. Namely, the scheduled and polarise network conditions, which are designed to promote diversity, may reduce the probability of a network adopting a more extreme (confident) collective belief, as compared to the mean-extreme and static network conditions.

4.1.3 Discussion

The finding of a significant treatment effect on how communication influenced CES, and the qualitatively observed treatment effect on the calibration of collective predictions, suggests that mediating communication in online social networks with different rewiring algorithms can steer the accuracy of collective beliefs. As such, these findings can be taken as a proof of concept that encourages continued research. But on the other hand, our main preregistered hypothesis that there would be a statistically significant main effect between network treatments on post-communication CES was not supported, and our simulation results do not directly map onto the empirical results. In order to reconcile these findings, there are three key considerations for future work: (1) more closely controlling for the confounding effects of initial network structure and individuals' differences, (2) applying the rewiring algorithms to networks of more knowledgeable individuals, and (3) better accounting for potential context-dependent effects of each algorithm.

In the experimental design we originally conceived, we sought to control for the confounding effects of initial network structure and individuals' differences by randomly reassigning each participant into one of four identically-structured but differently treated networks between each round. Unfortunately, because this procedure involves running 64 participants simultaneously on a single server, and because our experiment necessarily

involves algorithmic computation between each stage of each round, we were unable to run this design with the software used because participants experienced significant lags and crashes. This unexpected obstacle forced us to adjust our design such that participants were randomly assigned to a network condition upon signing up for the experiment, and then sent to a separate server depending on the condition (i.e., one network per server at a time). Though this adjustment was necessary to ensure participants could provide quality responses, it means our analysis of a main effect between network conditions may be confounded. To remedy this in future work we could either use different software or increase the statistical power of our study with a larger sample size.

A second limitation of our experiment is that the participants did not possess much relevant knowledge on the events being predicted. This can be noted in the observation that for six of ten events, not a single group was able to produce an accurate binary prediction (i.e., a collective prediction greater than 0.5 if the event occurred in reality, and vice versa, Table 4.2; also see Table C.1 for the average post-communication CES for each event in each condition). In principle, this general poor performance of the participants is inconsequential, because random assignment balances incompetence across treatments and we then focus on between treatment effects. However, the underlying logic of rewiring algorithms assumes that there exists some relevant, varied information to be communicated amongst individuals in the group. While an examination of the rationales entered by participants suggests than a vast majority of individuals engaged in good faith participation, it seems that our participants did not possess many unique pieces of evidence that could be amplified or discounted by a rewiring algorithm. Future work could thus benefit from evaluating the effects of rewiring algorithms on networks of more knowledgeable individuals.

Another limitation of our experiment is its focus on one particular prediction context: probabilistic estimates on events where individuals initial estimates display little to no skew towards one alternative or another (Figure 4.6). Related ongoing research demonstrates that the optimal network structure for eliciting the wisdom of the crowd depends on the estimation context — the specific population of individuals faced with a specific estimation task (Almaatouq, Rahimian, et al., 2021). Almaatouq, Rahimian, et al. (2021) show that when a group's initial estimates are highly skewed then a centralised network structure can promote collective accuracy, whereas decentralised network structures might hinder collective accuracy in such contexts, and vice versa when initial estimates display low

Event ID	*Static*	*Mean-Extreme*	*Polarise*	*Scheduled*
uk_covid	0	0	0	0
youtube_subs	10	11	11	10
biden_approval	0	0	0	0
us_uk_vax	0	0	0	0
bitcoin	0	0	0	0
super_bowl	0	0	0	0
us_climate	11	9	10	11
sp500	11	11	11	11
epl	0	0	0	0
americas_covid	2	1	2	1

Table 4.2: Tally of groups in each treatment that made the correct binary prediction (0.5 cutoff) on each event post-communication. A correct prediction means that the group's collective prediction what greater than 0.5 if the true outcome was 1, and vice versa. Maximum of 11 per cell.

skewness. Given that our rewiring algorithms affect network centralisation in different ways — namely, the mean-extreme algorithm increases it while the polarise algorithm decreases it — this insight could explain our empirical results and why they differ from our simulations. In our simulations with optimal Bayesian agents, networks' initial estimates always display a skew towards the truth; but in our empirical study, initial estimates displayed displayed no such skew (Figure 4.6). Thus, the polarise algorithm may simply have been better suited to the particular prediction tasks considered in our empirical study, and the mean-extreme and scheduling algorithms may be better suited to other contexts, such as those simulated with our modelling. In the next section I present follow-up simulations in which we explored this point and tested our algorithms in numeric prediction contexts (e.g., predicting the number of ICU per week during a pandemic) rather than binary prediction contexts (e.g., predicting whether the number of ICU admissions per week will be greater than 1,000), which characteristically elicit highly right-skewed distributions of initial predictions.

4.1.4 Follow-up simulations of numeric estimation contexts

Following up on the experimental work, we conducted additional simulations to explore how the rewiring algorithms might perform in numerical estimation contexts — where the 16-agent networks estimate (or predict) some unknown positive number — rather than binary estimation contexts. Such tasks map onto classical crowd wisdom scenarios such as estimating the weight of an ox, as well as high-stakes, real-world scenarios like forecasting the number of ICU admissions per week during a pandemic.

We follow the procedure described in the previous section on "modelling and simula-

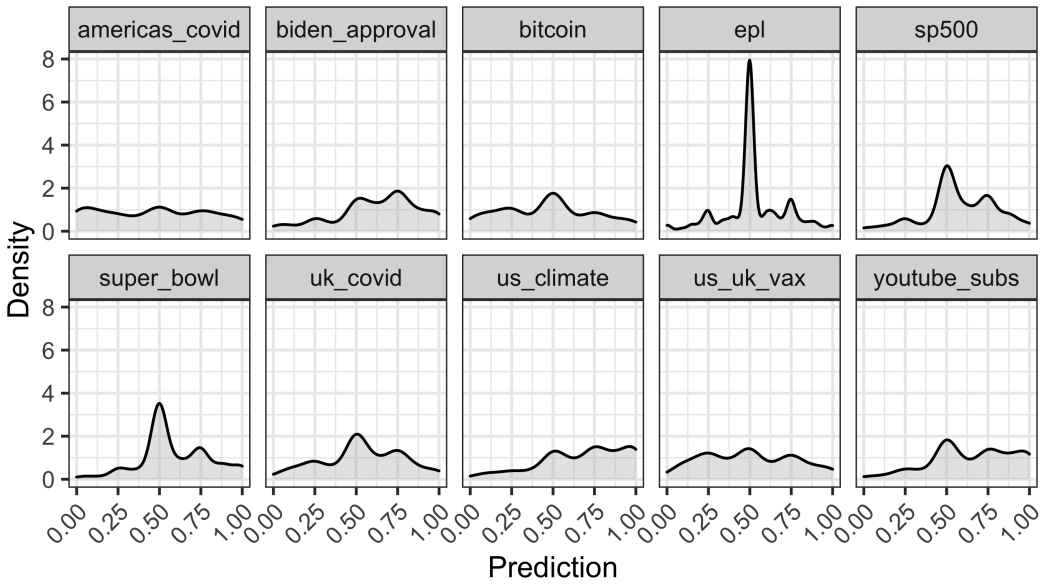

Figure 4.6: Aggregate distributions of participants' initial predictions for each event in the empirical study. See Table 4.1 to match event IDs to the actual event prompt.

tions" and initialise our model by randomly generating an undirected small-world network (Watts & Strogatz, 1998), have our agents follow the same updating rule borrowed from Becker et al. (2017), and consider the same four network conditions (static, mean-extreme, polarise, and scheduled). However, instead of having each agent integrate binary evidence via Bayes' theorem to establish their initial estimate, we assign each agent an initial estimate by sampling from a compilation of empirical data from four previously published experiments (Becker et al., 2017; Becker et al., 2019; Gürçay et al., 2015; Lorenz et al., 2011). This compiled dataset spans a total of 54 estimation tasks on which 2,885 individuals provided independent estimates (Almaatouq, Rahimian, et al., 2021). Each task — or "estimation context" — in this dataset is represented by a distribution of independent estimates and a true value. For example, one task contains 278 participants' estimates of the London population in July 2010, with the true value of 7,825,200 (Gürçay et al., 2015). Note, however, that we scale the estimates for each task to be between 0 and 1 in order to suit our belief updating rule and mean-extreme rewiring algorithm, while maintaining the distributions' shape.

Following 500 iterations of each of the 54 estimation tasks in which four matched networks are simulated (i.e., one of each network condition starting from an identical initial network), we assess collective accuracy by calculating the squared error of the mean estimate post-communication (CES). While other loss functions such as absolute error and square root error may be applicable in some task domains, our pattern of results is

consistent across these loss functions and we thus focus on CES for the sake of this paper; also because of the theoretical link of CES to the Diversity Prediction Theorem. Across all of the estimation tasks considered, the four network conditions' CES was nearly equal on average (static networks, $M = 0.016$, $SD = 0.031$; mean-extreme networks, $M = 0.017$, $SD = 0.033$; polarise networks, $M = 0.016$, $SD = 0.029$, scheduled networks, $M = 0.016$, $SD = 0.031$). However, these averages overlook potential context-dependent effects. Indeed, an analysis of CES task-by-task, rather than in aggregate, reveals that mean-extreme networks achieved the highest accuracy on 31 tasks, polarise networks achieved the highest accuracy on 15 tasks, and scheduled networks achieved the highest accuracy on 8 tasks. Static networks did not achieve the highest accuracy on any tasks. This observation further suggests that rewiring algorithms may serve as a viable strategy for boosting collective accuracy in social networks.

To better understand the context-dependent effects of the rewiring algorithms, we characterise each task by the skewness of the distribution of individuals' initial estimates, and then observed how each network condition's average CES varied across the skewness parameter space. As shown in Figure 4.7, the rewiring algorithms display a clear favouritism for certain regions of the skewness parameter space: mean-extreme networks were the most accurate for tasks with highly skewed estimate distributions ($n = 31$, $M = 9.47$, $SD = 9.33$), polarise networks were the most accurate for tasks with estimate distributions that display low skewness ($n = 15$, $M = 1.56$, $SD = 1.38$), and scheduled networks were the most accurate on tasks with mid-range skewness ($n = 8$, $M = 3.21$, $SD = 3.55$).

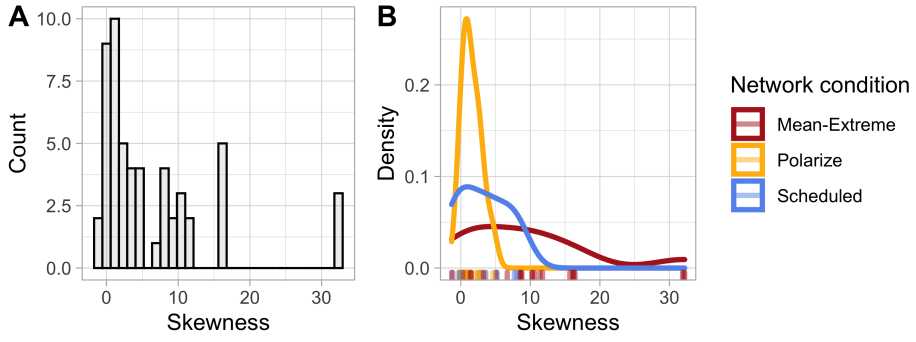

Figure 4.7: The skewness parameter space. (A) The distribution of skewness in the 54 estimation tasks considered. (B) The distribution of skewness where each network condition produced the lowest collective error as compared to the other conditions.

In Figure 4.8, we further investigate how the effects on collective accuracy produced by the rewiring algorithms track over skewness. Using the CES of static networks as

a baseline condition, we calculated three measures for each of the 54 estimation tasks for mean-extreme, polarise, and scheduled networks: the average effect on *CES* (i.e., the average change in error), the average relative effect on *CES* (i.e., the average change in error divided by the average error of matched static networks), and the probability of improvement (i.e., the proportion of the 500 iterations of each task where a given network condition was more accurate than a matched static network). This analysis suggests not only that the different rewiring algorithms prefer different estimation contexts, but that there is an important interaction: the mean-extreme algorithm actively increases collective error on tasks with low skewness and the polarise algorithm actively increases collective error on tasks with high skewness.

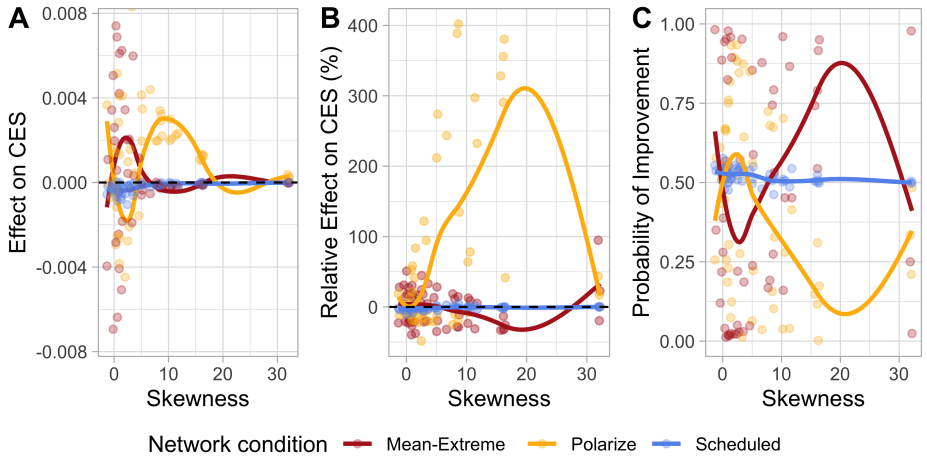

Figure 4.8: Network performance over skewness as compared to matched static networks. (A) The average effect on *CES* across skewness (i.e., the average change in *CES* compared to matched static networks). (B) The average relative effect on *CES* across skewness (i.e., the average change in error divided by the average error of matched static networks). (C) The probability of improvement across skewness (i.e., the proportion of the 500 iterations of each task where a given network condition was more accurate than matched static networks).

The results of these simulations suggest that it may be possible to identify distributional characteristics of judgements that allow one to select a rewiring algorithm capable of increasing the accuracy of social networks' collective estimations. Crucially, these are applicable in contexts where there is no track record of individuals' predictive success and the truth or falsity of individual estimates is not (yet) known. Where sufficient ground truth data on accuracy exists, such as in expert judgements of medical scans, that data can unquestionably be used to fine-tune networks of judges (Kurvers et al., 2019). But that leaves many of the most pressing real-world judgement tasks unaccounted for. In particular, we may want collective judgements to derive high-quality predictions for con-

sequential unique events, for which, by definition, ground truth data will be unavailable. A method that enhances collective accuracy in such contexts would thus provide a valuable prediction tool for many domains.

4.2 Chapter conclusion

The combined simulation and empirical results presented in this chapter speak to the issue of belief accuracy in a digital world on two levels. First, the findings here encourage a more positive, active approach towards (re)building online information environments that cater to and enhance human reasoning and decision making. While substantial research efforts have sought to identify psychological and structural mechanisms that hinder people's ability to form accurate beliefs online, work that leverages those findings to develop digital tools for supporting belief accuracy has only recently gained traction. Moreover, much of the existing literature that does evaluate such digital tools focuses on interventions that seek to mitigate negative features of online information environments (e.g., *accuracy nudges*, Pennycook et al., 2021; *innoculation*, Roozenbeek & van der Linden, 2019). As the findings presented in this chapter demonstrate, research that goes further to explore how those same, supposedly negative features (in this case, content-curating algorithms) might be re-claimed and re-designed to actively enhance human reasoning may prove fruitful. Such inquiry could inform the design of new civic technologies that are capable of realising the promise of a digital, democratic future. From a second level, the research presented in this chapter could be taken as evidence that the reliance on content-curating algorithms in the digital world can in fact undermine belief accuracy, inadvertently or otherwise. Although we designed algorithms specifically for the purpose of enhancing the wisdom of the crowd, we observed how even these well-meaning algorithms can lead users astray in certain contexts. It thus seems reasonable to infer that algorithms designed for other purposes — such as maximising user engagement on social media — may have unintended side-effects on the accuracy of users' beliefs under some circumstances.

Chapter 5

General conclusion

Has the digital world led to or catalysed the emergence of a "post-truth" society? Amid a backdrop of polarisation, anti-intellectualism, and conspiratorial thinking, understanding the ways in which online information environments influence people's ability to form accurate beliefs is a contemporary venture that warrants serious investigation. By adopting the conceptual lens of bounded rationality (H. A. Simon, 1955, 1957, 1971, 2000), I organised my approach to this topic by considering potential threats to belief accuracy related to both cognitive capacities "in the mind" and structural features "in the (digital) world," and further proposed how the two can be (re)aligned with the support of novel digital tools.

In Chapter 2, I began with a brief review of the general account of confirmation bias and motivated reasoning, whereby people are said to sample, reason, and update in ways that unduly favour information that is concordant with pre-existing beliefs over information that is discordant. While such psychological distortion has worrying implications for the accuracy of people's beliefs in an information-rich, digital world, this chapter emphasised the difficulty of identifying true bias — as a systematic deviation from accuracy — in contrived experimental set-ups. Specifically, I probed the optimistic belief updating phenomenon by conducting three variations of the standard methodology to test for asymmetric belief updating with neutral, non-valenced stimuli. If there is indeed a directional, motivational bias as predicted by the optimism account, then there should be no "bias" in belief updating with such stimuli. Yet, this chapter's main analysis demonstrates a thrice-replicated asymmetry in belief updating with neutral stimuli (Figure 2.2), and further investigation with proposed "fixes" for the standard methodology's flaws displays uninterpretable variability across samples and analytic techniques (Table 2.5). These findings

run counter to the narrative that people are irrational, motivated reasoners who use information to "believe what they want to believe" (Kunda, 1990, p. 480), thereby pushing for a greater appreciation of normative models if findings of "biased" cognition are to speak to the issue of belief accuracy in a digital world.

In Chapter 3, I opened with a brief overview of structural features of online information environments that have been cited as inherent threats to people's belief accuracy. Despite warranted and widespread concerns about attention-oriented incentive structures, manipulative bots, and content-curating algorithms, this chapter emphasised the difficulty of drawing meaningful conclusions on their effects with non-experimental methods. As a demonstration of this, I challenged recently presented evidence of a moral contagion effect, which is said to support the claim that the attention economics of social media result in moral-emotional content being shared regardless of its informational quality. By applying widely-accepted techniques for analysing large-scale, observational social media data, results show not only that the moral contagion effect does not consistently generalise, but, even more worryingly, that those same analytic techniques can support patently absurd findings, such as the XYZ contagion (Table 3.2). Moreover, specification curve analyses revealed how seemingly arbitrary decisions on outliers and covariates can lead materially different results to emerge from the same data (Figures 3.1 to 3.2). Altogether, these findings question whether conventional methodological tools for studying the digital world are providing the solid, meaningful insights needed to build upon.

Lastly, in Chapter 4, I raised the question of how belief accuracy might be supported online by re-appraising features of the digital world for epistemic benefit. Whereas there is growing interest in designing interventions for assisting users' navigation and evaluation of information on existing digital infrastructures (e.g., on social media platforms), I sought to encourage the research community to go further and imagine what alternative online information environments might look like. To do so, I designed, deployed, and evaluated a novel tool for enhancing collective belief accuracy online: rewiring algorithms that dynamically manipulate the structure of online social networks based on the distribution of beliefs reported. Here, agent-based modelling and an online multiplayer experiment showed how mediating communication with such algorithms can steer the accuracy of collective beliefs in different, context-dependent ways, thereby laying a foundation for continued research to inform the development of new civic technologies and more socially responsible digital media.

In the remainder of this chapter I conclude by describing three key insights provided by this thesis as a whole, touching on methodological, theoretical, and practical elements. Each of these aims to encourage a deeper reflection on not only *what* existing research tells us about belief accuracy in the context of a digital world, but also *how* researchers organise their study and methodologies. In doing so, directions for future work are also highlighted.

5.1 The value of absurd science

In Chapters 2 and 3 of this thesis, there was a focus on methodological challenges that limit our understanding of belief accuracy in a digital world. While these chapters took aim at substantively different domains of study — cognitive psychology and computational social science, respectively — the underlying approach of both of the studies presented is shared. That is, both the investigation of optimistic belief updating with neutral stimuli and the investigation of moral contagion via the XYZ contagion are examples of what can be called "absurd science."

Absurd science is not unique to this thesis' topic of belief accuracy in a digital world, and the inspiration for this approach is in fact taken from past studies across disciplines. These include the famous brain imaging study by Bennett et al. (2009) apparently showing the neural activity of a dead salmon engaged in a perspective-taking task, the analysis of longitudinal public health records by Cohen-Cole and Fletcher (2008) seemingly suggesting that acne, height, and headaches are contagious, and the study of human decision making by Hilbig (2010) that produced evidence in line with the conclusion that people accurately judge cities' population size by ranking them in alphabetical order. The value of such absurd scientific results is, of course, not in presenting a new substantive finding to be built upon. Just as neither my co-authors nor I believe that people consider which direction they are updating a belief on the probability scale before deciding the magnitude of their revision, or that people count the number of Xs, Ys, and Zs in a message before choosing whether to share it on social media, the researchers referenced above did not pursue their findings for the purpose of advancing a theory. Rather, such studies provide a broader critique of the methods they employ.

For a method to be informative it must be able to both display an effect when there is a meaningful effect of interest, *and* display no effect when there is no meaningful effect of interest. In the language of Bayes' theorem, this is to say that a method must have

adequate *diagnosticity*. The usefulness of any given research method within this frame can be judged much in the same way as a medical test or a machine learning classifier. For any of these, performance is defined by the ability to discriminate between alternative hypotheses; by the ability to balance sensitivity, $P(e|h)$, with specificity, $P(\neg e|\neg h)$. Consider now the results presented in Chapters 2 and 3 of this thesis: If past studies of optimistic belief updating (e.g., N. Garrett et al., 2014; Moutsiana et al., 2013; Sharot et al., 2011) are to be taken as evidence of a motivational, valence-driven bias, then the method used should not display "bias" in the absence of valence. If the analysis by Brady et al. (2017) is to be taken as evidence of a true moral contagion effect (also see, e.g., De Choudhury & De, 2014; Rathje et al., 2021; Stieglitz & Dang-Xuan, 2013, for studies using similar methods), then the method used should not display a "contagion" effect for nonsensical, causally irrelevant factors.

To be clear, it would be incorrect to conclude that people are not biased updaters, that social media platforms are not conducive to the spread of low-quality information, or that the digital world at large does not undermine belief accuracy based on the methodological problems highlighted in Chapters 2 and 3. In just the time it has taken to write this thesis, there have been several real-world events and news stories that seem to demonstrably document both instances of motivated reasoning and maliciously designed online infrastructures. For example, individuals who fell prey to conspiracy theories online were seemingly unable to accurately update their beliefs about COVID-19 risks as the pandemic unfolded, and eventually contracted the disease and died after failing to follow health guidance (Spring, 2020). Even more recently, an ex-Facebook employee on the "civic integrity" team revealed that despite internal research finding that their news feed algorithm promotes divisive content, the company chose not to act on the finding because of concerns it would reduce user engagement (Hagey & Horwitz, 2021)[1]. The studies presented in this thesis by no means downplay issues such as these. On the contrary, the studies in Chapters 2 and 3 further highlight the pressing need to develop proper methods and analytical techniques for evaluating such issues so that they may be effectively addressed.

While absurd science studies appear sporadically and garner considerable attention, continued work is needed to better define the scope and scale of problems they highlight. This is perhaps most clear in the domain of computational social science, which has seen the rapid development of machine learning tools to exploit large-scale, observational digital

[1] At the time of writing, however, it is unclear what data and methods were used to draw this conclusion in Facebook's internal research.

trace data for research. As demonstrated in Chapter 3 with the XYZ contagion, such data is inherently confounded and prone to display spurious correlations, meaning analytic tools built to find associations will not necessarily be able to find meaning in the data (Butler, 2013; Khoury & Ioannidis, 2014; Lazer et al., 2009). This prompts a range of questions that are the subject of ongoing work: Is correlation ever enough for explanatory research? If so, under what conditions? What is the probability that a significant correlation is indicative of a theoretically interesting relationship? Answering these questions seems necessary if researchers are to draw meaningful, generalisable conclusions from observational digital trace data.

5.2 The digital world as a hybrid system

Much of the existing discourse around the influence of the digital world on the accuracy of people's beliefs has struggled to disentangle user- and environment-driven effects. In Chapters 2 and 3, we saw the limitations of conventional methods for studying either psychological bias "in the mind" or structural bias "in the (digital) world" in isolation, whilst in Chapter 4 we saw how algorithmically mediated communication can indeed influence collective beliefs. Taken together, these results support a view of the digital world as a complex, hybrid, human-machine system, which raises the question of whether attempts to separate user- and environment-driven effects are worthwhile or theoretically informative.

Conceptualising the digital world as a hybrid system means examining the feedback processes among its human users and computer-mediated environments as the central units of analysis. This perspective is perhaps best articulated by Rahwan et al. (2019) in their call for a research programme on so-called *machine behaviour*, where they explain that:

> "many of the questions that relate to hybrid human–machine behaviours must necessarily examine the feedback loops between human influence on machine behaviour and machine influence on human behaviour simultaneously... there remains an urgent need to further understand feedback loops in natural settings, in which humans are increasingly using algorithms to make decisions and subsequently informing the training of the same algorithms through those decisions" (p. 483).

Through this view, the beliefs of individuals and collectives alike can be understood as emergent phenomenon produced by repeated interactions between cognitive components of the mind and structural components of the digital world. Of course, in many ways, this realisation simply brings us back to the basic proposition of bounded rationality that provided the framework for this thesis. Nevertheless, the findings presented here push for a more explicit acceptance of this conceptual point and call for further methodological innovation.

The development of new methodological approaches to study the digital world is in fact already ongoing. For example, researchers have recently proposed guidelines for digital field experiments (e.g., Mosleh, Pennycook, et al., 2021), algorithm auditing (e.g., Sandvig et al., 2014), and virtual lab experiments (e.g., Almaatouq, Becker, et al., 2021). Of particular interest to this thesis is the latter, which was utilised in Chapter 4. By translating the basic model of experimentation into web-based applications, virtual lab experiments (often referred to as "online multiplayer experiments" when more than one participant is involved at a time) enable two key functionalities that are especially relevant to the study of belief accuracy in a digital world. First, virtual lab experiments allow researchers to construct customised, realistic tasks and immersive online environments (Almaatouq, Becker, et al., 2021). This means that instead of speculating how cognitive processes observed in a contrived, artificial experiment might situate in the digital world, researchers can simply place (large groups of) participants in controlled online infrastructures in miniature, manipulate both task and environmental features, and observe interactions in real time. Second, the virtual lab allows researchers to run macro-level experiments where collective entities — such as algorithmically mediated social networks — are the unit of analysis, rather than individuals (Almaatouq, Becker, et al., 2021). This functionality sits well with this thesis' emphasis of the digital world as a complex, hybrid system by permitting researchers to explore questions that studies of micro-level processes alone can not address. Given these functionalities, virtual lab experiments present a ripe methodological opportunity for studying the emergence of beliefs in "algorithmically infused societies" (Wagner et al., 2021).

5.3 How to build a digital world that supports belief accuracy

Finally, the findings presented in this thesis provide some practical notes for how a digital world that actively supports belief accuracy might be built. Perhaps the most clear contribution here is a call for caution. As the findings in Chapters 2 and 3 show, existing, widely-cited studies that claim to identify mechanisms undermining belief accuracy might be less meaningful than they seem to be. Therefore, designing new tools and interventions to target those mechanisms could lead to not only ineffective investments and opportunity costs, but potentially adverse consequences given the complex, connected nature of the digital world as a hybrid system. Even when digital tools are consciously designed to enhance human reasoning and decision making in some contexts, those same tools can be expected to be damaging in others, as shown in Chapter 4.

Beyond this caution, however, this thesis encourages researchers to not restrict themselves to the design space that is delimited by existing online infrastructures. Instead of only considering ways to mitigate harmful effects of the digital world, designing and evaluating entirely new civic technologies can both generate practical tools and lead to the conceptualisation of new empirical questions for study.

Despite the methodological and conceptual limitations on our current understanding of belief accuracy highlighted in this thesis, it seems reasonable to conclude that the "solution" — or, the path to an epistemically responsible digital world — cannot be reduced down to any one thing. The solution so to speak will involve a concert of many different measures, be it psychologically-inspired digital tools, education for digital literacy, or policy-level regulation. What this thesis emphasises is that when implementing these (imperfect) measures together, researchers, technologists, and policy-makers alike must anticipate complex interactions and feedback processes. Belief accuracy in a digital world is neither the result of psychological capacities "in the mind" nor structural features "in the (digital) world," but rather a product of the alignment between them.

Appendix A

Supplementary information for Chapter 2

A.1 Supplementary analyses

A.1.1 Accounting for post-treatment bias

As is the case for existing studies that use the update method and life events of varying valence (e.g., N. Garrett & Sharot, 2014, 2017), there is a possibility that a post-treatment bias may influence our models' estimates (see Montgomery et al., 2018 for a detailed exposition of post-treatment bias). Since participants provide their ratings of valence for each life event after having received the BR, the provision of the BR might influence the subsequent valence rating and the subsequent belief update. To remedy this potential problem in our main analysis, we re-ran the analysis as if every event were rated as neutral by the participants. Given that we aimed to compile a set of life events that could plausibly be rated as neutral by participants, this analysis is consistent with our research objective of detecting an asymmetry with valence-neutral events, despite its neglect of the variability in participants' perceptions of event valence. The results of this analysis mirror those of the main analysis in the main text, albeit slightly attenuated, meaning that an asymmetry was observed in upwards versus downwards updating across all life events.

In Study 1, there were 2,482 trials with an upwards direction of error ($M = 2.72$, $SD = 5.90$) and 2,336 with a downwards direction of error ($M = 9.36$, $SD = 14.31$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 4798) = 434.00$, $p < 0.001$), such that an upwards direction of error decreased update scores by approximately 6.43 percentage points (fixed effect estimate)

$\pm$ 0.31 (standard error), as compared to downwards direction of error.

In Study 2, there were 2,288 trials with an upwards direction of error ($M = 4.06$, $SD = 10.17$) and 2,459 with a downwards direction of error ($M = 9.16, SD = 18.22$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 4735) = 136.70$, $p < 0.001$), such that an upwards direction of error decreased update scores by about 5.02 percentage points (fixed effect estimate) $\pm$ 0.43 (standard error), as compared to downwards direction of error.

In Study 3, there were 2,429 trials with an upwards direction of error ($M = 4.04$, $SD = 9.60$) and 2,278 with a downwards direction of error ($M = 8.95$, $SD = 20.80$). An LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 4701) = 118.21$, $p < 0.001$), such that an upwards direction of error decreased update scores by about 4.78 percentage points (fixed effect estimate) $\pm$ 0.44 (standard error) as compared to downwards direction of error.

A.1.2 Adding stimuli as a random factor

In the LMM in our main analysis we included participants as a random factor to follow Marks and Baines (2017) and account for the nested structure of the data. Given that the main objective of the present work is to demonstrate that the update method — as it has been employed in the literature — can elicit asymmetric belief updating with neutral events, it was deemed crucial to follow analysis plans with precedent in the literature. However, it can be argued that the design of the update method warrants the inclusion of stimuli (life events) as a random factor, and that not doing so could inflate Type I error rates on the fixed effect estimates Judd et al. (2012) and Yarkoni (2019). As a check of robustness, we therefore conducted an additional analysis where we re-fit the LMMs in our main analysis with stimuli as a random factor[1]. In each study, the asymmetry in belief updating with neutral life events remained with slightly attenuated fixed effect estimates.

In Study 1, an LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 1507) = 222.13$, $p < 0.001$), such that an upwards direction of error decreased update scores by approximately 8.95 percentage points (fixed effect estimate) $\pm$ 0.60 (standard error), as compared to downwards direction of error.

In Study 2, an LMM determined that direction of error significantly affected the mag-

[1] We used the same procedure to select a model specification as described in the main analysis in the main text, which led us to reduce the complexity of the random effects structure to include only random intercepts by participant and random intercepts by stimuli. However, results also hold in the maximally complex model specifications.

nitude of participants' updating ($F(1, 1505) = 64.77$, $p < 0.001$), such that an upwards direction of error decreased update scores by about 5.96 percentage points (fixed effect estimate) $\pm$ 0.74 (standard error), as compared to downwards direction of error.

In Study 3, an LMM determined that direction of error significantly affected the magnitude of participants' updating ($F(1, 1442) = 61.05$, $p < 0.001$), such that an upwards direction of error decreased update scores by about 6.42 percentage points (fixed effect estimate) $\pm$ 0.82 (standard error) as compared to downwards direction of error.

A.2 Supplementary tables

ID	*Life event*	*BR (%)*
1	Be exactly the same weight in 10 years' time	26
2	Last the whole of next winter without catching a minor cold	20
3	Participate in a game of sport in the next four weeks	29
4	Clean the bathroom in the next four weeks	78
5	50 or more hours of sleep in a single week in the next four weeks	56
6	Fix a broken possession in the next four weeks	39
7	Get a haircut in the next four weeks	45
8	Have your photo taken in the next four weeks	75
9	Play a board game in the next four weeks	29
10	Shop for clothes in the next four weeks	56
11	Try a new hobby, craft, or sport in the next four weeks	31
12	Receive a utility bill in the next four weeks	78
13	Win a competitive game of sport in the next four weeks	22
14	Burn something that you are cooking in the next four weeks	41
15	Embarrass yourself in the next four weeks	60
16	Get lost in the next four weeks	26
17	Have a disagreement with a friend in the next four weeks	43
18	Have a headache in the next four weeks	82
19	Be ill one day because of over-drinking in the next four weeks	21
20	Stay up past 2 AM for school or work in the next four weeks	40
21	Get teased at/made fun of in the next four weeks	35
22	Get lied to in the next four weeks	60
23	Get stuck in traffic in the next four weeks	71
24	The next car that passes is a BMW	14
25	Have a vegan meal in the next four weeks	14
26	Make a purchase by contactless card in the next four weeks	29
27	Check your phone more that 100 times in one day in the next four weeks	45
28	The next car that passes is the colour black	20
29	Receive a phone call from an unknown number in the next four weeks	66
30	Buy a non-dairy milk alternative in the next four weeks	48
31	Spend more than £121 on dinners out over the next four weeks	19
32	Spend less than £89 on commuting over the next four weeks	33
33	Send fewer than 106 text messages over the next four weeks	15
34	Feel a phantom phone vibration in the next four weeks	80
35	Walk less than seven miles over the next four weeks	17
36	That your next flight will have a minor delay (i.e., 15 minutes or less)	26
37	That the next store you visit is air conditioned	30
38	Receive junk mail in the next four weeks	71
39	Drink between 56 and 84 cups of coffee over the next four weeks	43
40	Make your bed every day for the next four weeks	21
41	Use more than 3.7GB of mobile data over the next four weeks	17
42	Check your mobile data usage in your phone's settings in the next four weeks	13
43	Spend more than 40 hours online in the next week	81
44	The next car you ride in, other than your own, is the colour white	19
45	Take the Eurostar train service in the future	16
46	Own a pet	45
47	Live in a home that was originally built before 1900	20
48	Move homes more than 10 times in your lifetime	18
49	Enrol in private health insurance	11
50	Meet your future spouse through an online dating service	38
51	Marry someone with a different political affiliation to you	26

Table A.1: Set of life events and accompanying base rate (BR) statistics used as stimuli. Participants were asked to "Please estimate how likely this event is to happen to you," and to "Please estimate how likely this event is to happen to the average person." Events 1-2 are from Shah et al. (2016), 3-23 are from Garrett and Sharot (2017), and 24-51 and have not be previously used in research.

Study	*dfn*	*dfd*	*F*	*p-value*
1	1	112.57	131.46	<0.001
2	1	139.75	61.57	<0.001
3	1	107.53	45.64	<0.001

Table A.2: Results of linear mixed effects model with only neutral trials and the maximally complex random effects structure. This specification includes random slopes and intercepts by participant for direction of error, plus correlations between random effects. This model specification is singular, hence the reporting of a simpler specification in the main text. Statistics pertain to Type III tests of the fixed effect of the direction of error on belief updating. Degrees of freedom are approximated with Satterthwaite's method (dfn refers to the numerator degrees of freedom and dfd refers to the denominator degrees of freedom).

Study	*Fixed Factor*	*dfn*	*dfd*	*F*	*p-value*
1	Direction of Error	1	114.31	198.47	<0.001
	Event Valence	2	127.62	33.66	<0.001
	Interaction	2	160.89	49.19	<0.001
2	Direction of Error	1	123.41	104.90	<0.001
	Event Valence	2	136.27	29.82	<0.001
	Interaction	2	154.78	38.96	<0.001
3	Direction of Error	1	95.12	48.67	<0.001
	Event Valence	2	133.32	13.72	<0.001
	Interaction	2	143.60	47.26	<0.001

Table A.3: Results of linear mixed effects model with direction of error, event valence, and an interaction term and the maximally complex random effects structure. This specification includes random slopes and intercepts by participant for direction of error, event valence, and the interaction term, plus correlations between random effects. Fitting this model led to singularities and negative eigenvalues, hence the reporting of a simpler specification in the main text. Statistics pertain to Type III tests of the models' fixed effects. Degrees of freedom are approximated with Satterthwaite's method (dfn refers to the numerator degrees of freedom and dfd refers to the denominator degrees of freedom).

Study	*dfn*	*dfd*	*F*	*p-value*
1	1	246.67	19.47	<0.001
2	1	473.51	10.81	0.001
3	1	162.2	7.84	0.006

Table A.4: Results of linear mixed effects model with only neutral trials and the maximally complex random effects after accounting for misclassification. This specification includes random slopes and intercepts by participant for direction of error, plus correlations between random effects. This model specification is singular, hence the reporting of a simpler specification in the supplementary text. Statistics pertain to Type III tests of the fixed effect of the direction of error on belief updating. Degrees of freedom are approximated with Satterthwaite's method (dfn refers to the numerator degrees of freedom and dfd refers to the denominator degrees of freedom).

Study	*Fixed Factor*	*dfn*	*dfd*	*F*	*p-value*
1	Direction of Error	1	178.04	65.14	<0.001
	Event Valence	2	195.49	19.63	<0.001
	Interaction	2	168.84	2.33	0.101
2	Direction of Error	1	601.39	30.06	<0.001
	Event Valence	2	318.13	8.47	<0.001
	Interaction	2	815.00	8.13	<0.001
3	Direction of Error	1	217.40	5.77	0.017
	Event Valence	2	277.15	19.77	<0.001
	Interaction	2	211.56	5.95	0.003

Table A.5: Results of linear mixed effects model with direction of error, event valence, and an interaction term and the maximally complex random effects structure after accounting for misclassification. This specification includes random slopes and intercepts by participant for direction of error, event valence, and the interaction term, plus correlations between random effects. Fitting this model led to singularities and negative eigenvalues, hence the reporting of a simpler specification in the supplementary text. Statistics pertain to Type III tests of the models' fixed effects. Degrees of freedom are approximated with Satterthwaite's method (dfn refers to the numerator degrees of freedom and dfd refers to the denominator degrees of freedom).

A.3 Supplementary figures

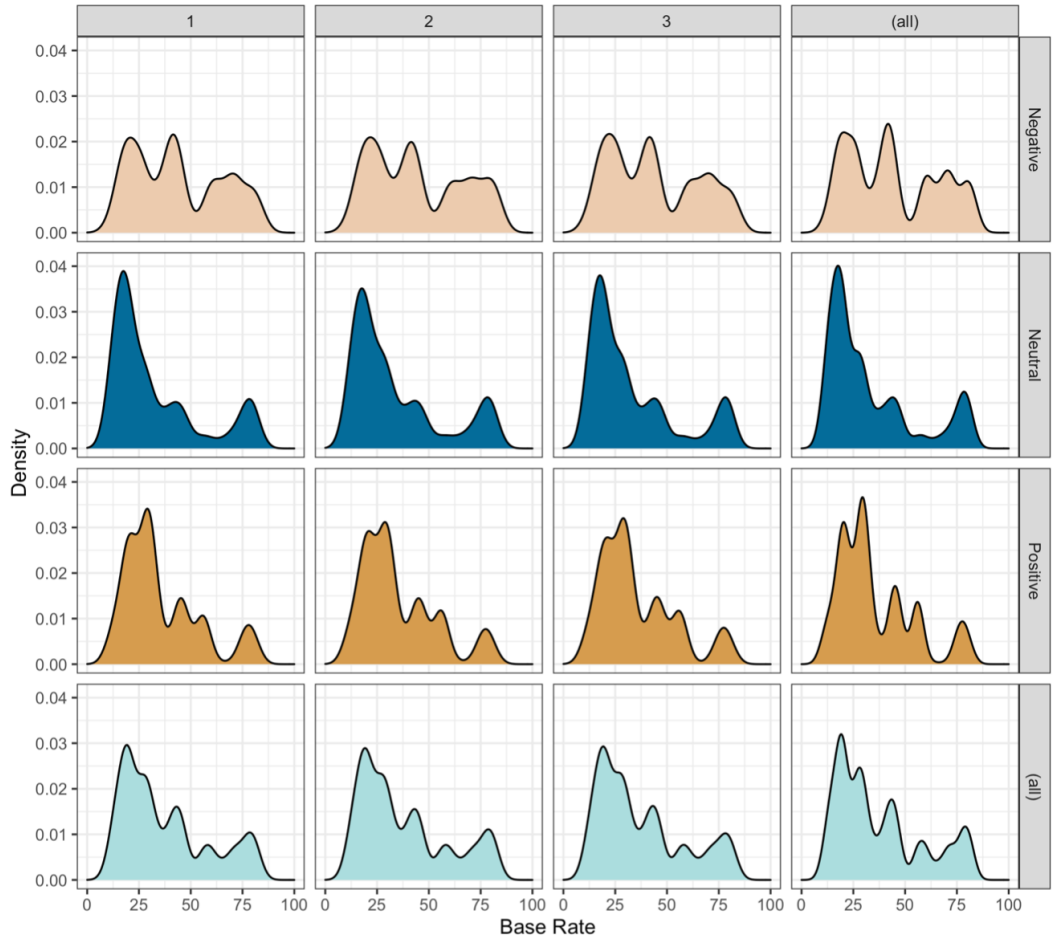

Figure A.1: Density plots displaying the distributions of event base rates across studies (top labels) and event valence (right labels). It should be noted, however, that because each participant self-rates the valence of each event, each participant is likely to encounter different distributions of base rate statistics.

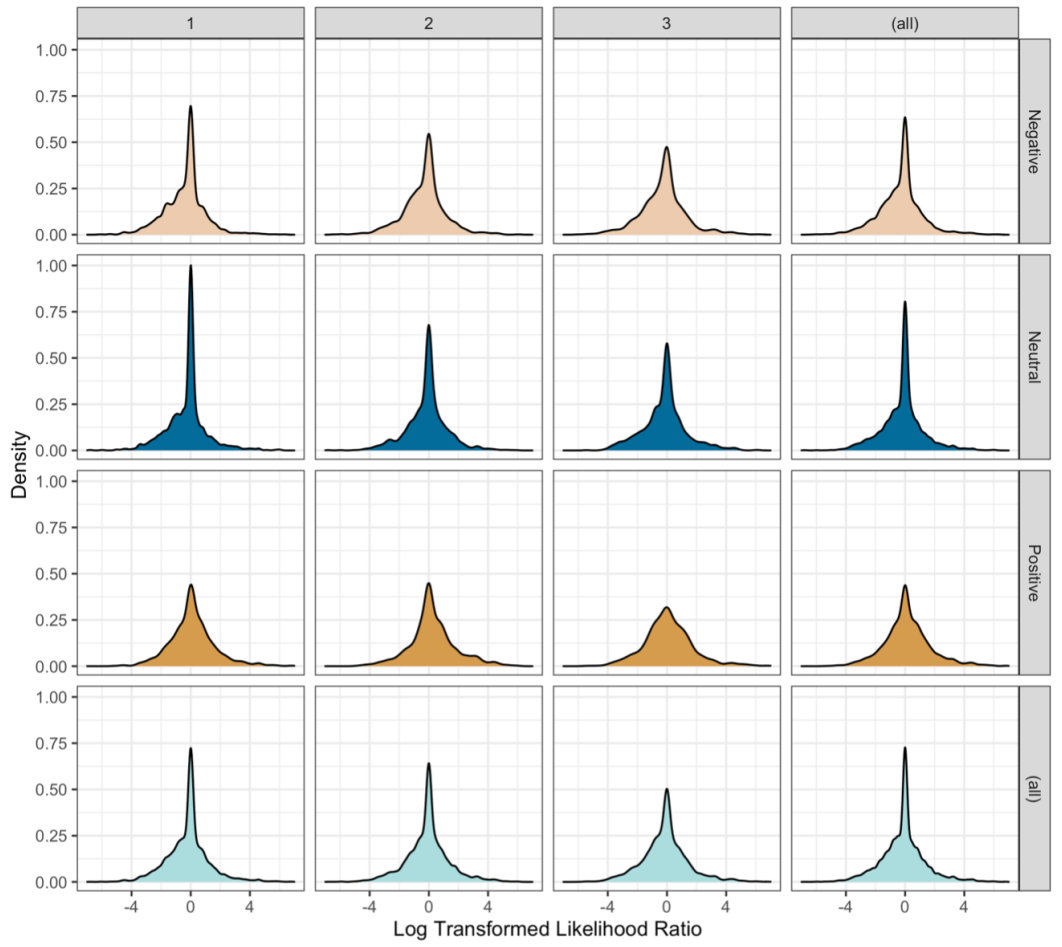

Figure A.2: Density plots displaying the distributions of log transformed likelihood ratios across studies (top labels) and event valence (right labels).

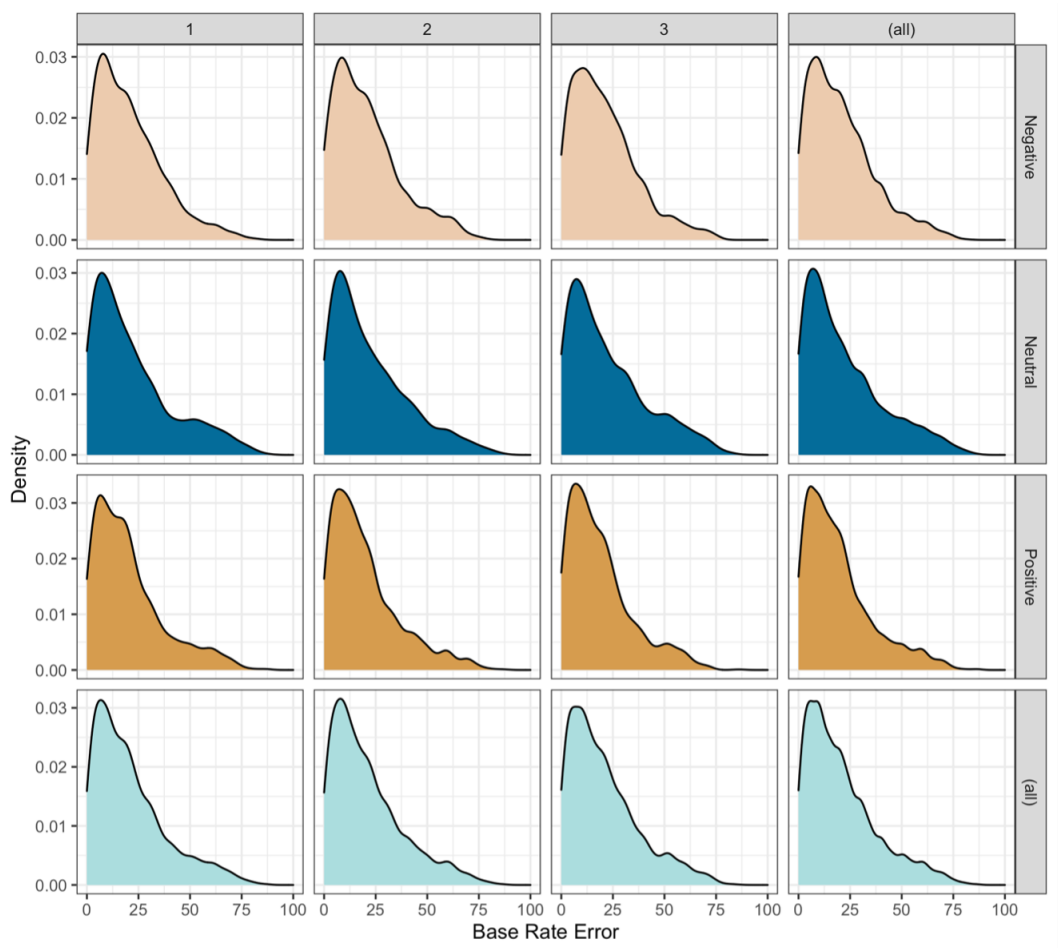

Figure A.3: Density plots displaying the distributions of "base rate error" across studies (top labels) and event valence (right labels)

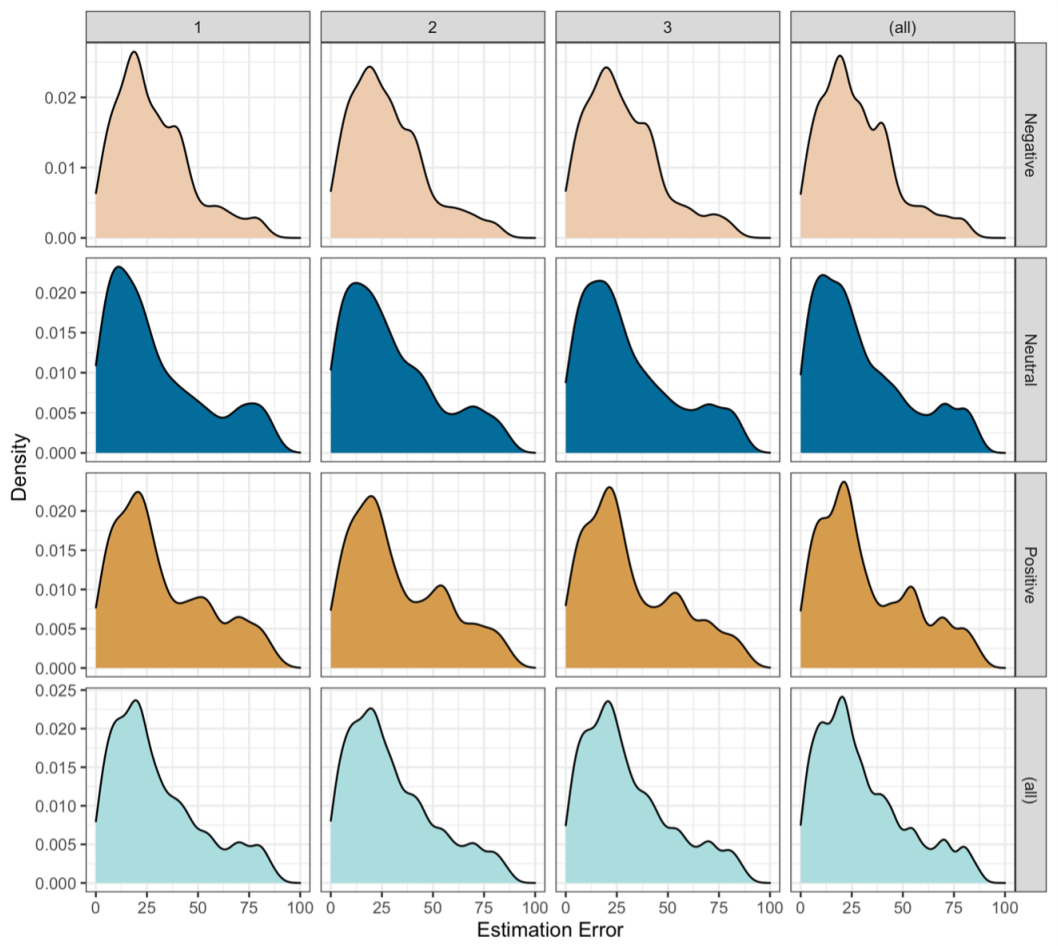

Figure A.4: Density plots displaying the distributions of "estimation error" across studies (top labels) and event valence (right labels).

Appendix B

Supplementary information for Chapter 3

B.1 Supplementary analyses

B.1.1 Evaluating Brady et al.'s dictionaries as predictors of human judgements of moral expression in the Moral Foundations Twitter Corpus

One possible explanation for the inconsistent moral contagion effects observed in the present work is measurement error. That is, the dictionaries used by Brady et al. (2017) might not be accurately measuring expressions of moral sentiment in tweets. Identifying moral sentiments in text is difficult because different types of moral sentiment can co-occur, they might only be implicitly signaled, and because the ground truth is inherently subjective (Hoover et al., 2019). In order to investigate how well Brady et al.'s (2017) dictionaries identify expressions of moral sentiment we conducted a supplementary analysis with the Moral Foundations Twitter Corpus (MFTC), which contains 34,987 tweets from seven topics of discourse displayed in Table B.1 ["All lives matter" (ALM), Baltimore protests, "Black lives matter" (BLM), hate speech messages from Davidson et al. (2017), 2016 US Presidential Election, #MeToo, and Hurricane Sandy] that have been manually annotated by three to five human annotators for moral sentiment (Hoover et al., 2019). Note that the #MeToo and the 2016 US Presidential Election corpora included in the MFTC and those addressed in the main analyses of the present work are different, despite sharing discourse topics.

Since the present work is not concerned with individual categories of moral sentiment

(e.g., purity, loyalty, authority, etc.), we collapsed the category labels such that we compared the total number of moral labels to the number of non-moral labels assigned by the annotators, in turn producing a binary classification of each tweet as moral or non-moral. We then applied four logistic regression classifiers with the moral-emotional, distinctly moral, and distinctly emotional dictionaries as predictors (one multiple logistic regression with all predictors included, and the three nested, single-variable logistic regressions) to see if the dictionaries' predicted classifications aligned with human judgements of moral expression.

Figure B.1 displays ROC curves and calculated AUC values for each logistic regression classifier as applied to each corpus included in the MFTC. Across the seven corpora the mean AUC for the multiple logistic regression classifier ranged from 51.7% in the Davidson hate speech corpus to 83.2% in the #MeToo corpus ($M_{AUC} = 72.2\%$). In line with the analysis reported by Hoover et al. (2019), we found classification performance to vary significantly by context. In addition, we calculated the logistic regression classifiers' precision, recall, and F1 metrics. Due to class imbalances in the data (Table B.1), we used repeated under-sampling whereby we randomly excluded observations from the majority class in each corpus and re-fit the classifiers and then averaged the calculations across 100 iterations (Table B.2). We found the logistic regression classifiers to have poor recall ($M_{Recall} = 53.9\%$). This suggests that the dictionary-based approach does not effectively identify all tweets in which human annotators find moral sentiment expressed, which raises an additional methodological concern about the specific measurements made in Brady et al. (2017).

Corpus	**Non-Moral**	**Moral**	*Total*
ALM	726	3,698	4,424
Baltimore	2,869	2,724	5,593
BLM	1,133	4,124	5,257
Davidson	3,825	1,048	4,873
2016 US Election	1,877	3,481	5,358
#MeToo	914	3,977	4,891
Hurricane Sandy	585	4,006	4,591
Total	11,929	23,058	34,987

Table B.1: Frequencies of moral and non-moral expression in the manually annotated twitter corpora comprising the Moral Foundations Twitter Corpus (MFTC).

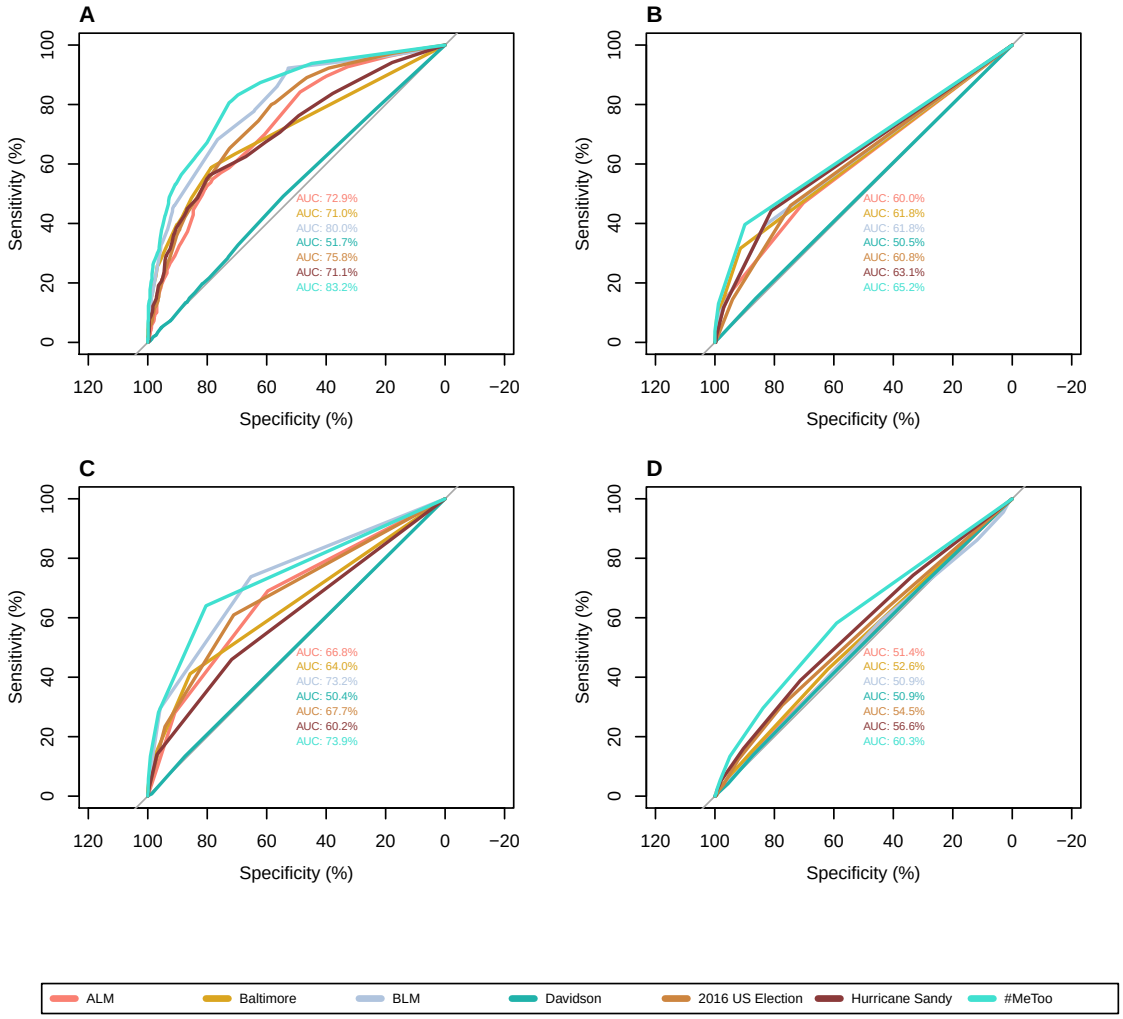

Figure B.1: ROC/AUC plots of dictionary logistic regression classifiers of moral expression when applied to the complete MFTC corpora. (A) Logistic regression classifier with all three dictionaries — moral-emotional, distinctly moral, and distinctly emotional — used as predictors of moral expression in tweets. (B) Logistic regression classifier with only the moral-emotional dictionary as a predictor. (C) Logistic regression classifier with only the distinctly moral dictionary as a predictor. (D) Logistic regression classifier with only the distinctly emotional dictionary as a predictor.

Metric	ALM	Baltimore	BLM	Davidson	2016 US Election	Hurricane Sandy	#MeToo
moral_expression ~ME_words + moral_words + emotional_words							
AUC	0.729 (0.007)	0.710 (0.001)	0.801 (0.006)	0.519 (0.008)	0.758 (0.004)	0.712 (0.010)	0.832 (0.006)
F1	0.599 (0.009)	0.618 (0.001)	0.617 (0.006)	0.506 (0.004)	0.608 (0.004)	0.631 (0.011)	0.652 (0.007)
Precision	0.669 (0.008)	0.640 (0.001)	0.732 (0.007)	0.505 (0.004)	0.677 (0.004)	0.601 (0.010)	0.739 (0.007)
Recall	0.514 (0.008)	0.527 (0.001)	0.508 (0.011)	0.511 (0.006)	0.545 (0.005)	0.566 (0.011)	0.604 (0.007)
moral_expression ~ME_words							
AUC	0.613 (0.013)	0.652 (0.000)	0.715 (0.017)	0.425 (0.080)	0.671 (0.008)	0.638 (0.014)	0.727 (0.023)
F1	0.529 (0.014)	0.451 (0.000)	0.501 (0.011)	0.295 (0.148)	0.538 (0.007)	0.543 (0.019)	0.530 (0.014)
Precision	0.660 (0.010)	0.530 (0.000)	0.708 (0.007)	0.425 (0.080)	0.642 (0.005)	0.527 (0.016)	0.696 (0.009)
Recall	0.521 (0.012)	0.473 (0.001)	0.513 (0.052)	0.436 (0.051)	0.399 (0.008)	0.464 (0.018)	0.586 (0.009)
moral_expression ~moral_words							
AUC	0.722 (0.009)	0.734 (0.002)	0.737 (0.016)	0.522 (0.012)	0.708 (0.004)	0.694 (0.039)	0.770 (0.014)
F1	0.611 (0.008)	0.787 (0.002)	0.702 (0.006)	0.520 (0.017)	0.644 (0.004)	0.700 (0.010)	0.798 (0.006)
Precision	0.631 (0.005)	0.743 (0.002)	0.680 (0.003)	0.522 (0.012)	0.678 (0.002)	0.619 (0.010)	0.765 (0.003)
Recall	0.508 (0.007)	0.530 (0.001)	0.506 (0.017)	0.511 (0.008)	0.574 (0.006)	0.575 (0.012)	0.588 (0.005)
moral_expression ~emotional_words							
AUC	0.533 (0.020)	0.587 (0.000)	0.697 (0.041)	0.373 (0.113)	0.638 (0.015)	0.593 (0.036)	0.691 (0.052)
F1	0.467 (0.016)	0.316 (0.000)	0.390 (0.011)	0.261 (0.259)	0.462 (0.008)	0.443 (0.022)	0.397 (0.014)
Precision	0.691 (0.015)	0.412 (0.000)	0.738 (0.012)	0.373 (0.113)	0.609 (0.007)	0.459 (0.019)	0.639 (0.012)
Recall	0.535 (0.017)	0.427 (0.000)	0.525 (0.089)	0.390 (0.101)	0.306 (0.008)	0.389 (0.020)	0.583 (0.013)

Table B.2: Performance metrics of dictionary-based logistic regression classifiers. Values indicate the mean of a given metric following 100 iterations of under-sampling, with standard deviations in parentheses. Classification threshold set to 0.5.

B.1.2 Bootstrap resampling

Bootstrap resampling was also conducted as a robustness check to keep with the procedures of Brady et al. (2017). This technique involves regenerating variations of a dataset by

sampling with replacement, meaning that certain datapoints may be duplicated and others may be omitted. By iteratively repeating this procedure and re-fitting each model in question (500 iterations in this case), a distribution of effect sizes is produced along with a 95% confidence interval, which is considered indicative of the reliability of an effect within a sample. Specifically, an observed effect may be deemed stable if the confidence interval does not straddle zero.

It should be noted, however, that this procedure will only ever speak to the robustness of an effect within a sample, when the critical issue of interest is whether what has been found in a sample is indicative of the population at large (e.g., is the observed moral contagion effect generalizable to political tweets or political communications?). While it is true that an effect that is not even stable within a sample provides poorer evidence vis-à-vis the wider population than one that is, the fact that an effect is stable within a sample is insufficient to determine whether it extends beyond that sample. Moreover, the concerns that correlational analyses of big data raise for spurious factors are evidently not assuaged by bootstrap resampling: the XYZ contagion passes this robustness check in the three largest datasets analysed Figure B.2. Only out-of-sample prediction can address this issue, as conducted in the present study.

B.1.3 Specification curve analyses of Brady et al.'s (2017) data

In the main text, we report the results of specification curve analyses (SCA) of the COVID-19, #MeToo, and #MuellerReport corpora. To supplement these analyses, we applied SCA to Brady et al.'s (2017) data. Across model specifications that considered their chosen covariates and three arbitrary (but defensible) increments of outliers (the top 10, 100, or 1,000 most retweeted messages), we find the moral contagion effect to be particularly robust in the climate change corpus (median $B = 0.14$, $SD = 0.06$), and positive but variable in the gun control corpus (median $B = 0.08$, $SD = 0.10$). However, the moral contagion effect appears notably unstable in the same-sex marriage corpus with a negative median regression coefficient (median $B = -0.04$, 0.09). Figure B.3 further shows how supposed "outliers" can influence results, which is expected to an extent given the fat-tailed distribution of retweet data.

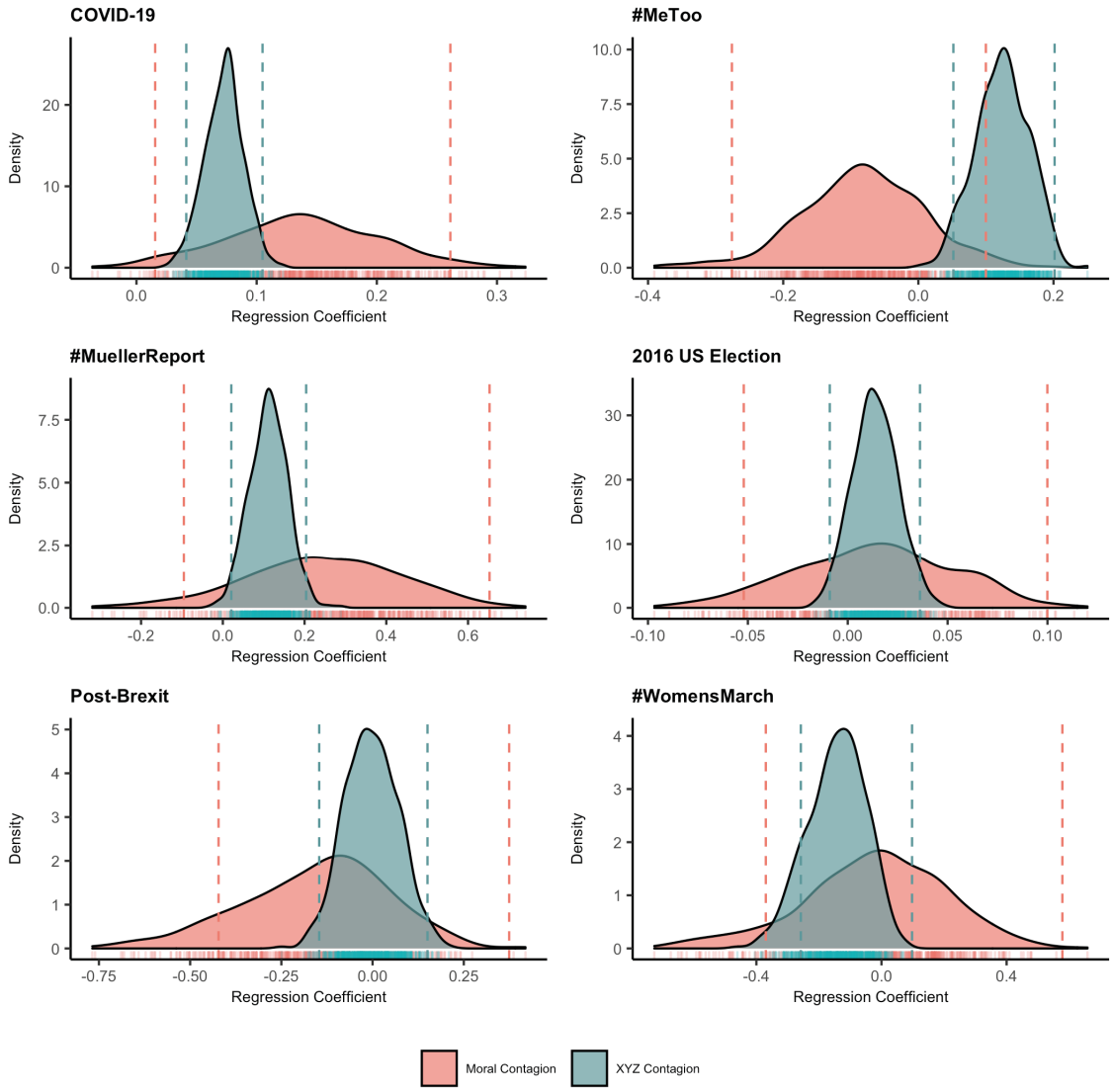

Figure B.2: Density plots of bootstrap resampling results in each corpus. Each plot displays 500 iterations (per model) of resampling. Dotted lines indicate the 95% confidence intervals for the respective effects.

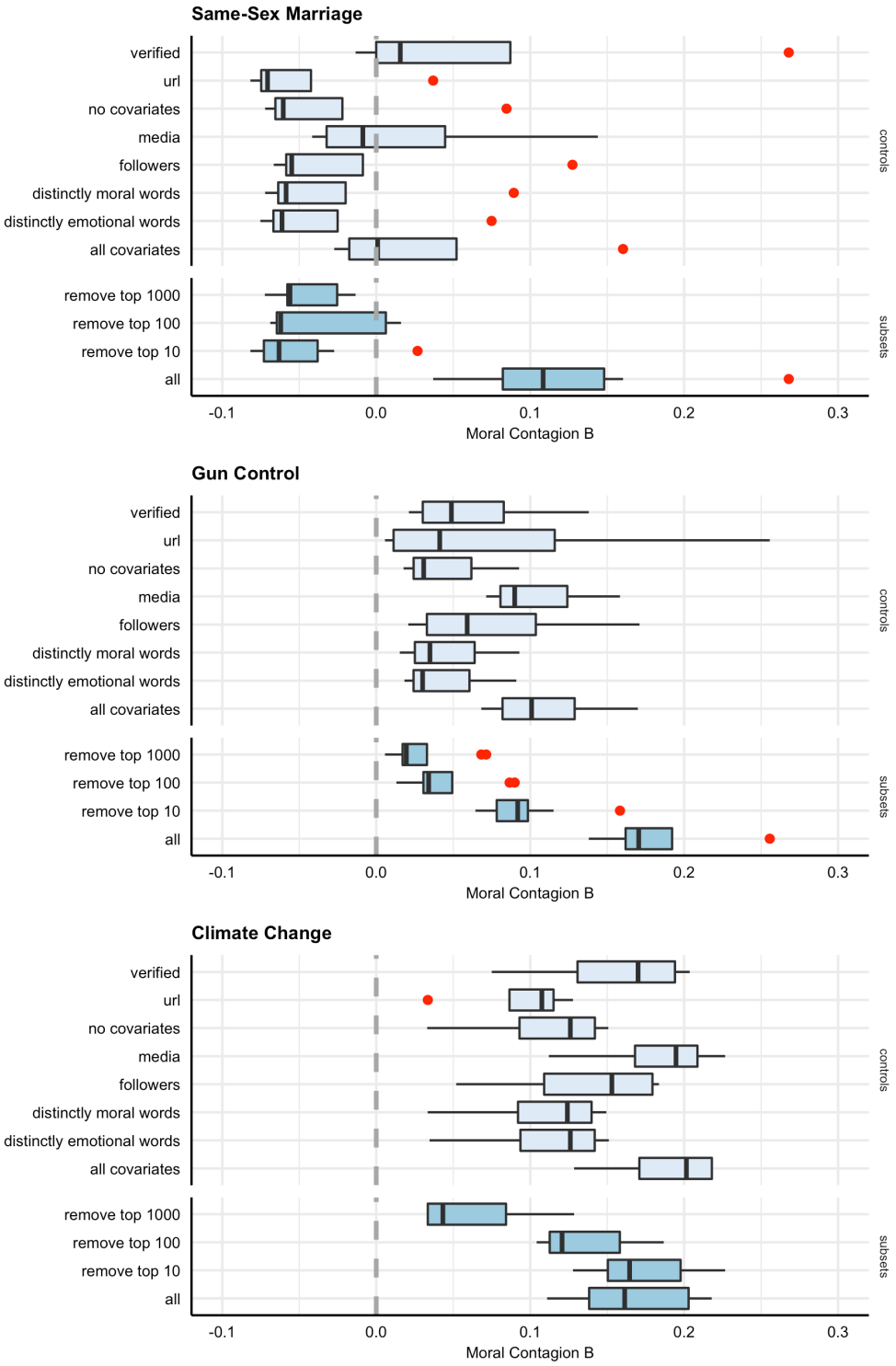

Figure B.3: Summary plot of specification curve re-analysis of Brady et al. (2017). Boxplots show the distribution of unstandardised negative binomial regression coefficients produced by model specifications accounting various covariates and outliers (y-axis).

Appendix C

Supplementary information for Chapter 4

C.1 Rewiring algorithm schematics

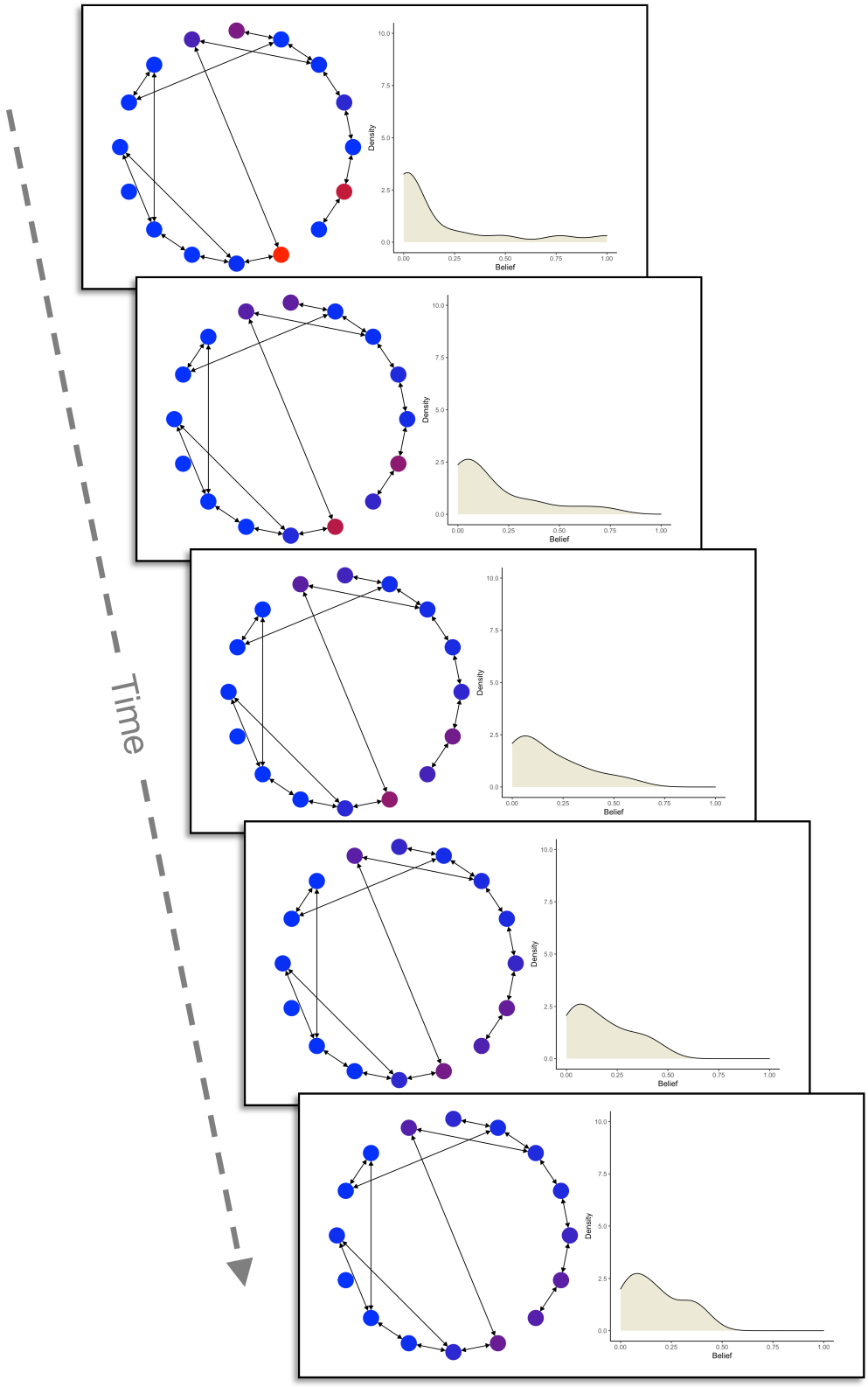

Figure C.1: Example case of the *static* network treatment across the five stages (an initial starting network plus four stages of communication). Nodes are coloured on a gradient based on their current belief: bright red for 1 and bright blue for 0. The distribution of these beliefs are shown to the right of each network at each stage.

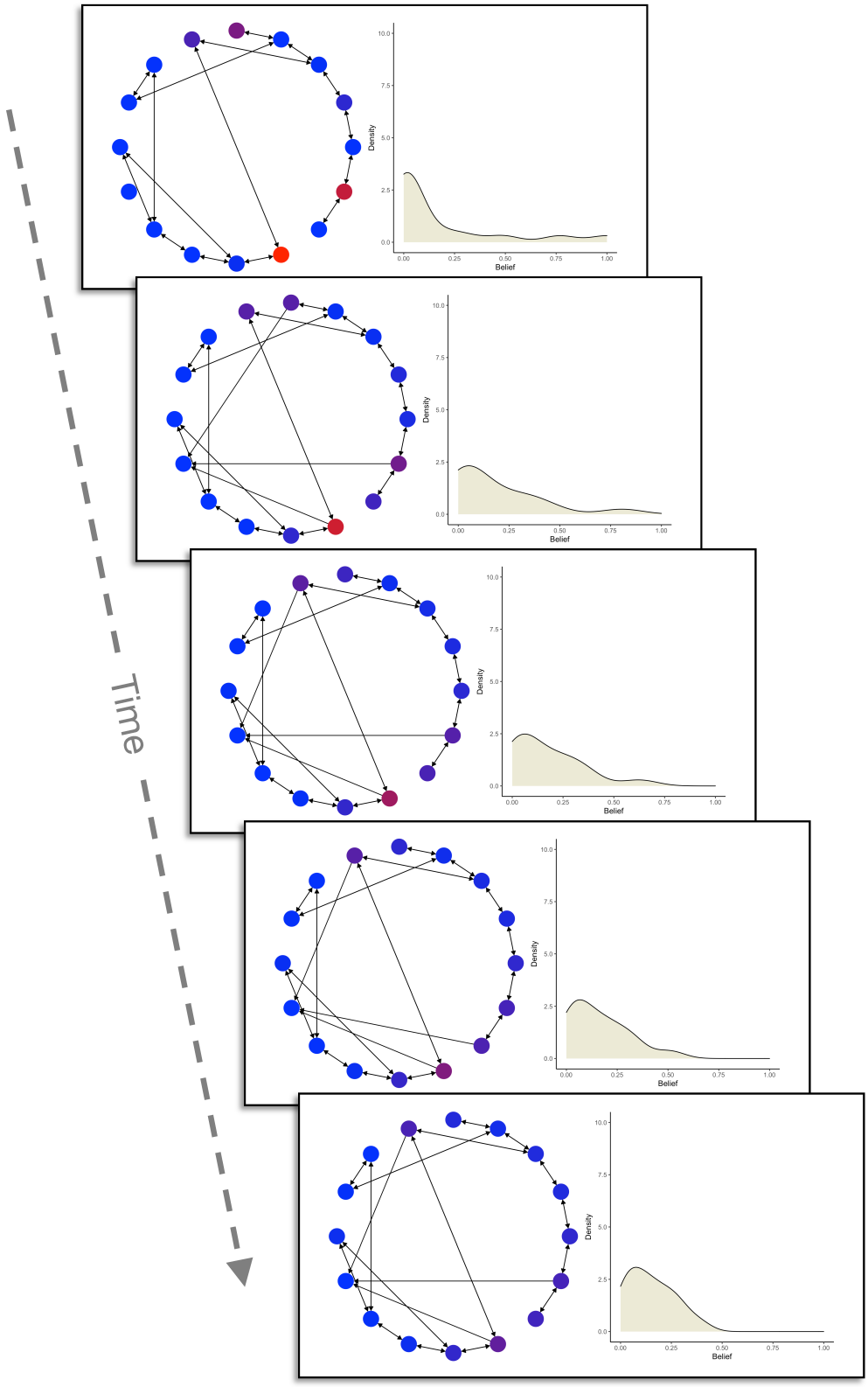

Figure C.2: Example case of the *mean-extreme* network treatment across the five stages (an initial starting network plus four stages of communication). Nodes are coloured on a gradient based on their current belief: bright red for 1 and bright blue for 0. The distribution of these beliefs are shown to the right of each network at each stage.

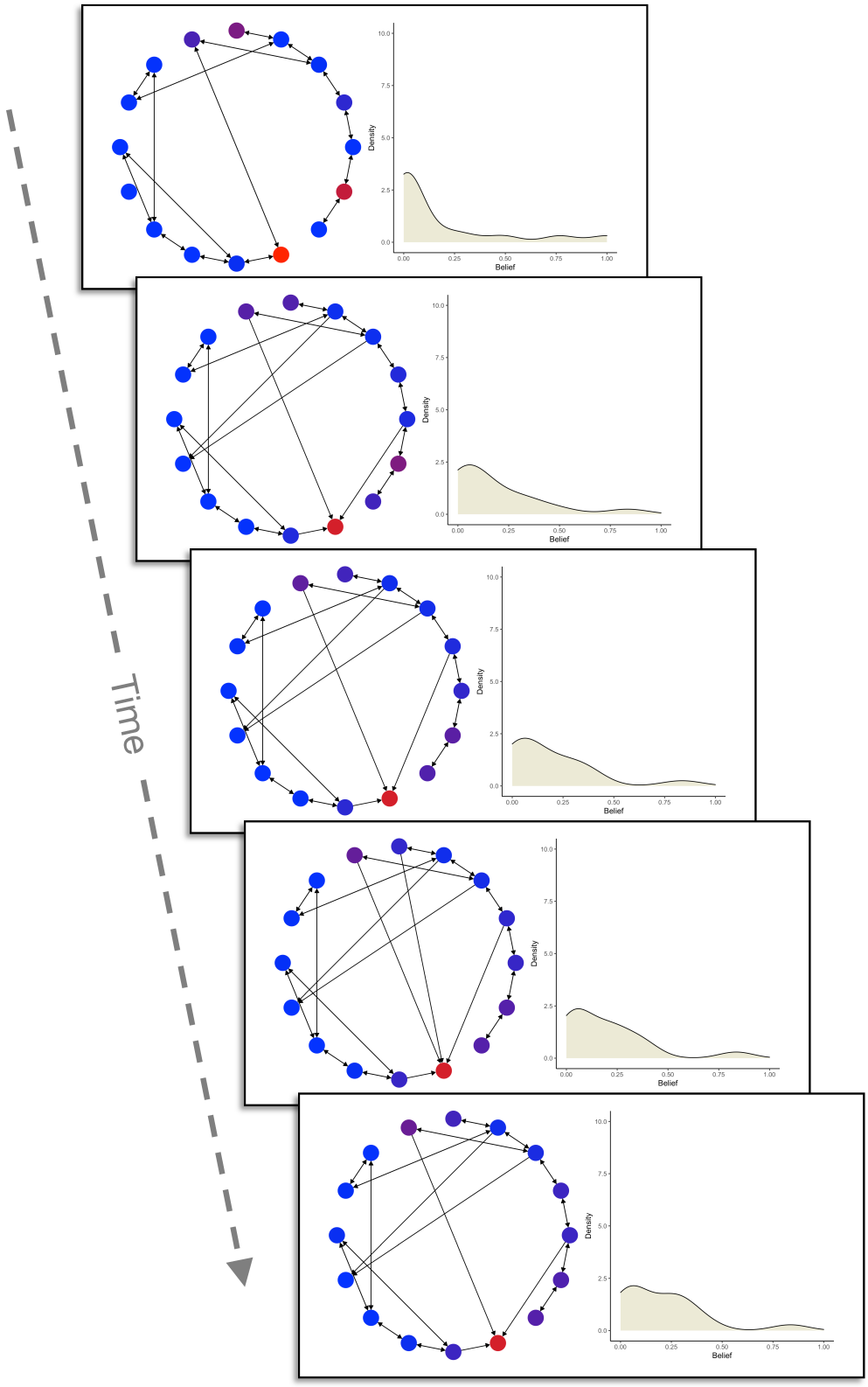

Figure C.3: Example case of the *polarise* network treatment across the five stages (an initial starting network plus four stages of communication). Nodes are coloured on a gradient based on their current belief: bright red for 1 and bright blue for 0. The distribution of these beliefs are shown to the right of each network at each stage.

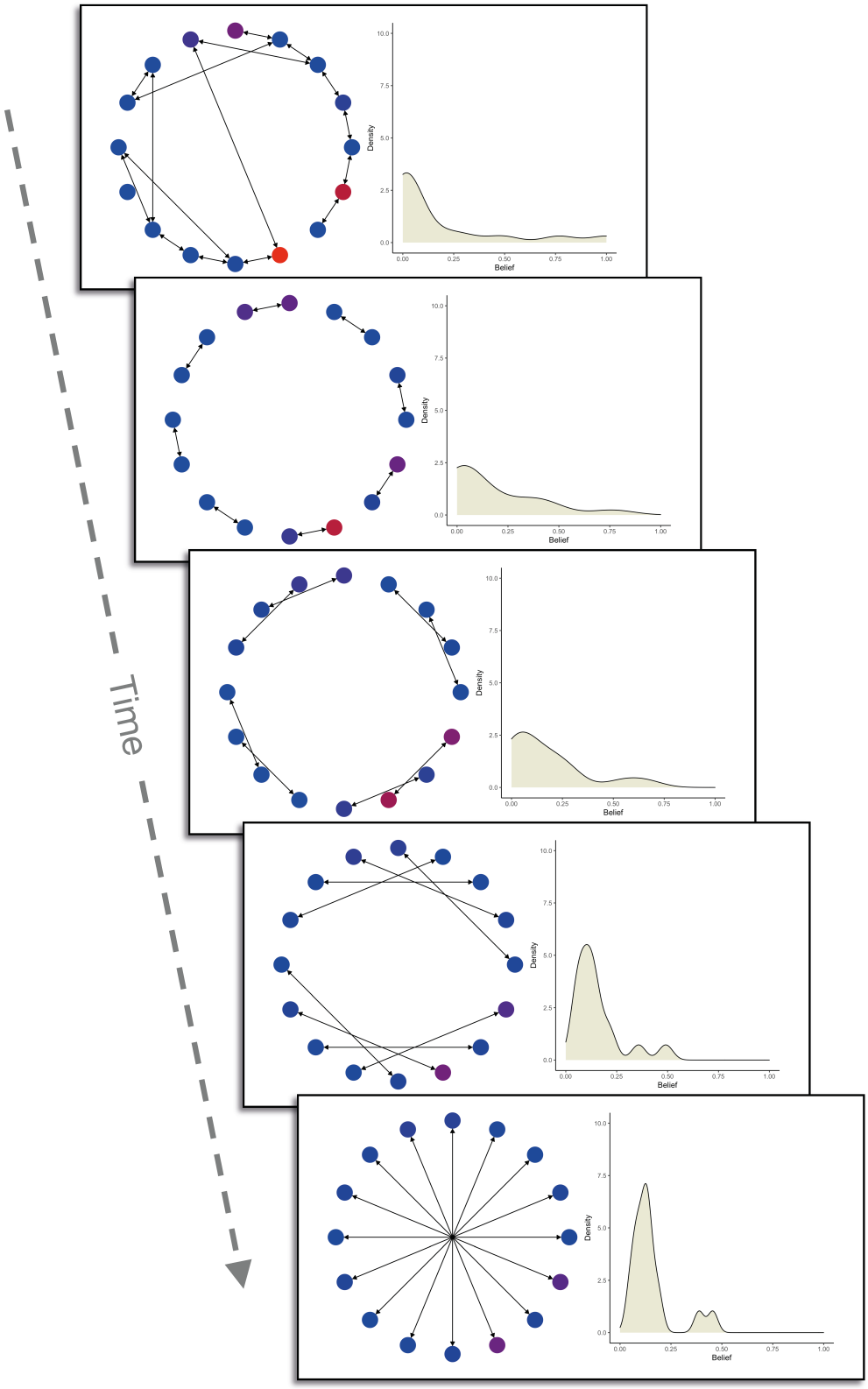

Figure C.4: Example case of the *scheduled* network treatment across the five stages (an initial starting network plus four stages of communication). Nodes are coloured on a gradient based on their current belief: bright red for 1 and bright blue for 0. The distribution of these beliefs are shown to the right of each network at each stage.

C.2 Experimental interface

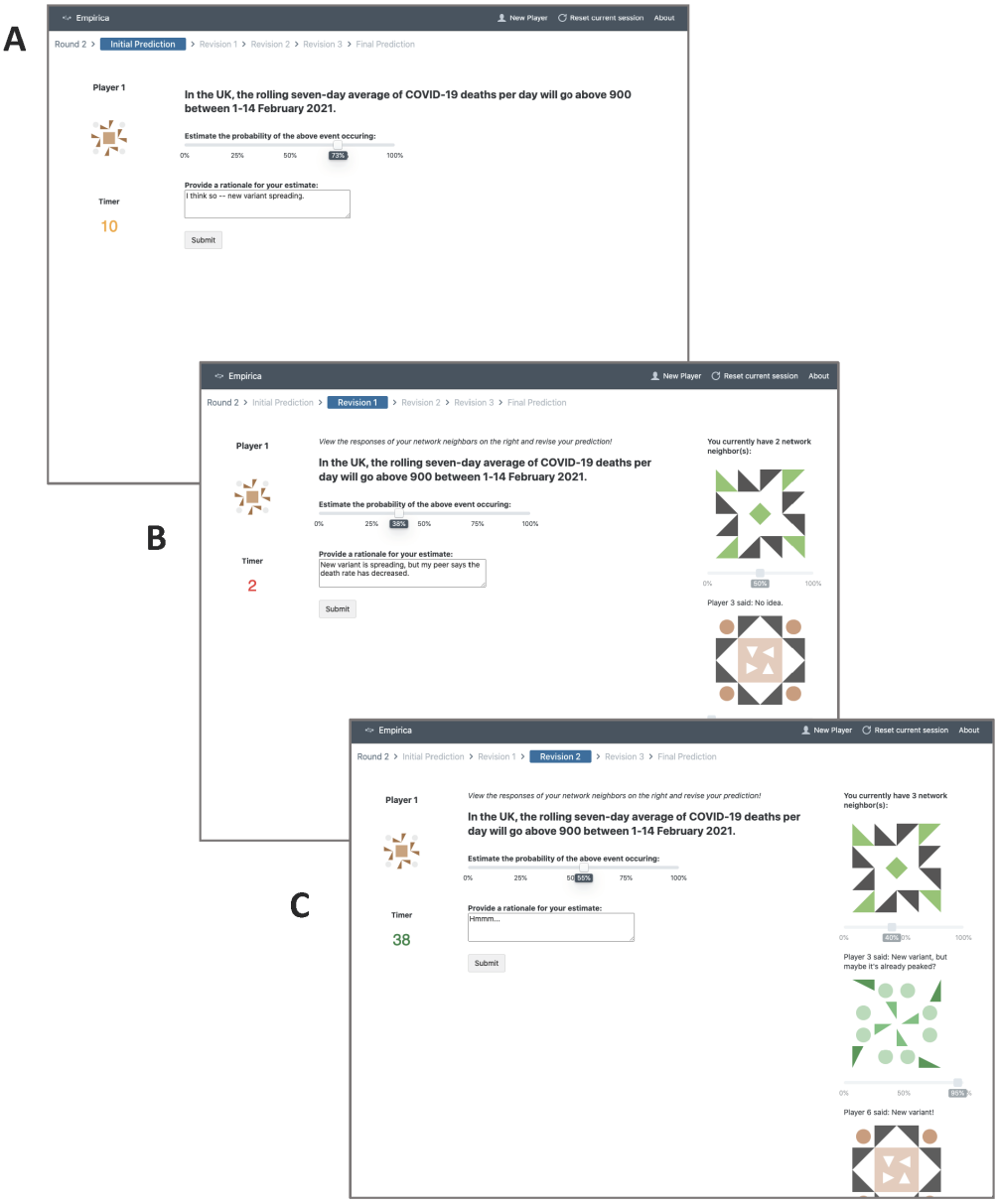

Figure C.5: Screenshots of the experiment's user interface. (A) At the first stage of each round, participants provide an initial prediction and rationale in the absence of social information. (B) After initial, independent responses have been provided, participants view the responses of their network neighbours in a right-side column and revise their own responses. (C) In the experimental network treatments, participants' network neighbours may change between stages. This procedure is repeated for ten rounds (i.e., ten events being predicted) with five stages per round (three of five stages pictured above). Each stage is limited to a duration of 60 seconds.

C.3 Supplementary empirical results

Table C.1: Average collective error squared (CES) of groups in each treatment for each event. Standard deviations are in parentheses.

Event ID	*Static*	*Mean-Extreme*	*Polarise*	*Scheduled*
uk_covid	0.31 (0.12)	0.46 (0.08)	0.30 (0.11)	0.30 (0.10)
youtube_subs	0.11 (0.11)	0.06 (0.05)	0.06 (0.05)	0.09 (0.11)
biden_approval	0.49 (0.10)	0.52 (0.10)	0.44 (0.14)	0.44 (0.11)
us_uk_vax	0.25 (0.13)	0.14 (0.08)	0.20 (0.09)	0.19 (0.11)
bitcoin	0.20 (0.10)	0.15 (0.07)	0.20 (0.12)	0.16 (0.08)
super_bowl	0.45 (0.13)	0.47 (0.19)	0.37 (0.10)	0.46 (0.08)
us_climate	0.07 (0.05)	0.13 (0.10)	0.11 (0.10)	0.09 (0.06)
sp500	0.12 (0.06)	0.11 (0.05)	0.13 (0.05)	0.12 (0.06)
epl	0.27 (0.11)	0.28 (0.10)	0.27 (0.09)	0.28 (0.15)
americas_covid	0.41 (0.13)	0.43 (0.21)	0.34 (0.14)	0.38 (0.08)

Bibliography

Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, *70*(2), 542–555.

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of us elections in the 21st century. *Electoral Studies*, *41*, 12–22.

Adams, J. S. (1961). Reduction of cognitive dissonance by seeking consonant information. *The Journal of Abnormal and Social Psychology*, *62*(1), 74.

Adhokshaja, P. (2017). #inauguration and #WomensMarch. *data.world*. https://data.world/adhokshaja/inauguration-and-womensmarch

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, *191*, 104254.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 1–14.

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379–11386.

Almaatouq, A., Rahimian, M. A., Burton, J. W., & Alhajri, A. (2021). When social influence promotes the wisdom of crowds. *PsyArXiv*, *Version 3*.

Amador, J., Oehmichen, A., & Molina-Solana, M. (2017). Fakenews on 2016 US elections viral tweets (november 2016 - march 2017). *Zenodo*.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al. (2008). The promise of prediction markets. *Science*, *320*(5878), 877.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243–250.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235.

Barron, K. (2018). Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? *WZB Discussion Paper*, 68.

Bartley, N., Abeliuk, A., Ferrara, E., & Lerman, K. (2021). Auditing algorithmic bias on twitter. *13th ACM Web Science Conference 2021*, 65–73.

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1).

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious Mixed Models [arXiv: 1506.04967]. *arXiv:1506.04967 [stat]*.

BBC. (2019). Jair bolsonaro: 'poop every other day' to protect the environment. https://www.bbc.co.uk/news/world-latin-america-49304358

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*(26), E5070–E5076.

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, *116*(22), 10717–10722.

Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. *Neuroimage*, *47*(Suppl 1), S125.

Ben-Nun Bloom, P., & Levitan, L. C. (2011). We're closer than i thought: Social network heterogeneity, morality, and political persuasion. *Political Psychology*, *32*(4), 643–665.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, *21*(11-7).

Bishop, B. (2009). *The big sort: Why the clustering of like-minded america is tearing us apart*. Houghton Mifflin Harcourt.

Bolin, J. L., & Hamilton, L. C. (2018). The news you choose: News media preferences amplify views on climate change. *Environmental Politics*, *27*(3), 455–476.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2021). *Cross-country trends in affective polarization* (tech. rep.). National Bureau of Economic Research.

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

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*(1), 17–33.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107–117.

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4).

Burton, J. W., Almaatouq, A., Rahimian, M. A., & Hahn, U. (2021). Rewiring the wisdom of the crowd. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, 1–7.

Burton, J. W., Hahn, U., Almaatouq, A., & Rahimian, M. A. (2021). Algorithmically mediating communication to enhance collective decision-making in online social networks. *Proceedings of the 9th ACM Collective Intelligence Conference*.

Burton, J. W., Harris, A. J. L., Shah, P., & Hahn, U. (2022). Optimism where there is none: Asymmetric belief updating observed with valence-neutral life events. *Cognition*, *218*.

Buser, T., Gerhards, L., & Van Der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, *56*(2), 165–192.

Butler, D. (2013). When google got flu wrong. *Nature News*, *494*(7436), 155.

Caddick, Z., & Rottman, B. M. (2019). Politically motivated causal evaluations of economic performance. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 182–188.

Caplan, R., & boyd, d. (2016). Who controls the public sphere in an era of algorithms. *Mediation, Automation, Power*, 1–19.

Center for Humane Technology. (n.d.). Take control. https://www.humanetech.com/take-control

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots reveal political bias on social media. *Nature Communications*.

Chowdhury, R., Sharot, T., Wolfe, T., Düzel, E., & Dolan, R. J. (2014). Optimistic update bias increases in older age. *Psychological Medicine*, *44*(9), 2003–2012.

Cohen-Cole, E., & Fletcher, J. M. (2008). Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *BMJ*, *337*, a2533–a2533.

Colley, T., & Moore, M. (2020). The challenges of studying 4chan and the alt-right:'come on in the water's fine'. *New Media & Society*, 1461444820948803.

Condorcet, M. J. A. N. C. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Chelsea.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*(2), 331–335.

Corner, A., Whitmarsh, L., & Xenias, D. (2012). Uncertainty, scepticism and attitudes towards climate change: Biased assimilation and attitude polarisation. *Climatic Change*, *114*(3), 463–478.

Corney, D. (2021). How does automated fact checking work? https://fullfact.org/blog/2021/jul/how-does-automated-fact-checking-work/

Coscia, M., & Rossi, L. (2020). Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, *17*(167), 20200020.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, *22*(2), 369–395.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, *110*(15), 5791–5796.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business School Press, Boston, USA.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1).

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 3267.

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Eighth international AAAI conference on weblogs and social media*.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*(3), 366–375.

Denny, M., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

Desai, S. A. C., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453.

DeSilver, D. (2020). The polarized congress of today has its roots in the 1970s. https://www.pewresearch.org/fact-tank/2014/06/12/polarized-politics-in-congress-began-in-the-1970s-and-has-been-getting-worse-ever-since/

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745.

Dunlap, R. E., McCright, A. M., & Yarosh, J. H. (2016). The political divide on climate change: Partisan polarization widens in the us. *Environment: Science and Policy for Sustainable Development*, *58*(5), 4–23.

Dunn, A. G., Surian, D., Dalmazzo, J., Rezazadegan, D., Steffens, M., Dyda, A., Leask, J., Coiera, E., Dey, A., & Mandl, K. D. (2020). Limited role of bots in spreading vaccine-critical information among active twitter users in the united states: 2017–2019. *American Journal of Public Health*, *110*(S3), S319–S325.

Ecker, U. K., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 1–25.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. John Wiley & Sons.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Einstein, K. L., & Glick, D. M. (2015). Do I think BLS data are BS? The consequences of conspiracy theories. *Political Behavior*, *37*(3), 679–701.

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–11.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545.

European Parliament. (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Facebook. (n.d.). How is facebook addressing false information through independent fact-checkers? https://www.facebook.com/help/1952307158131536/

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).

Feinerer, I., Hornik, K., & Feinerer, M. I. (2015). Package 'tm'. *Corpus*, *10*(1).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Ferrara, E., & Yang, Z. (2015). Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, *1*, e26.

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552.

Fisher, E., & Shapiro, S. A. (2020). *Administrative competence: Reimagining administrative law*. Cambridge University Press.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. *International Workshop on Agent Theories, Architectures, and Languages*, 21–35.

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339.

Galton, F. (1907). Vox populi.

Garcia, D., & Rimé, B. (2019). Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, *30*(4), 617–628.

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.

Garrett, N., & Sharot, T. (2014). How Robust Is the Optimistic Update Bias for Estimating Self-Risk and Population Base Rates? (M. Pessiglione, Ed.). *PLoS ONE*, *9*(6), e98848.

Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*, *50*, 12–22.

Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience*, *8*, 639.

Garrett, R. K. (2017). The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, *6*(4), 370–376.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–466.

Gentzkow, M. (2016). Polarization in 2016. *Toulouse Network for Information Technology*, 1–23.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, *126*(4), 1799–1839.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Goldhaber, M. H. (1997). The attention economy and the net. *First Monday*.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological review*, *109*(1), 75.

Golebiewski, M., & boyd, d. (2018). Data voids: Where missing data can easily be exploited.

González-Bailón, S., & De Domenico, M. (2021). Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences*, *118*(11).

Good, I. J. (1950). *Probability and the weighing of evidence* (tech. rep.). C. Griffin London.

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, *12*(2), 225–248.

Gotthard-Real, A. (2017). Desirability and information processing: An experimental study. *Economics Letters*, *152*, 96–99.

Grofman, B., Owen, G., & Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*(3), 261–278.

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*.

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*, *117*(27), 15536–15545.

Gürçay, B., Mellers, B. A., & Baron, J. (2015). The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making*, *28*(3), 250–261.

Hagey, K., & Horwitz, J. (2021). Facebook tried to make its platform a healthier place. it got angrier instead. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline

Hahn, U., Hansen, J. U., & Olsson, E. J. (2018). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, 1–31.

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of learning and motivation* (pp. 41–102). Elsevier.

Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science*, *11*(1), 194–206.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web*, 527–538.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., & Etter, M. (2011). Good friends, bad news-affect and virality in twitter. *Future information technology* (pp. 34–43). Springer.

Hardman, D. (2009). *Judgment and decision making: Psychological perspectives* (Vol. 11). John Wiley & Sons.

Harris, A. J. L., & Hahn, U. (2021). Problems with optimistic belief updating tasks compromise behavioural and neuroscientific conclusions. *In preparation*.

Harris, A. J. L., Shah, P., Catmur, C., Bird, G., & Hahn, U. (2013). Autism, optimism and positive events: Evidence against a general optimistic bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 555–560.

Hart, W., Albarracin, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555.

Heath, B. (2020). *Americans divided on party lines over risk from coronavirus: Reuters/ipsos poll* [Reuters]. https://www.reuters.com/article/us-health-coronavirus-usa-polarization/americans-divided-on-party-lines-over-risk-from-coronavirus-reuters-ipsos-poll-idUSKBN20T2O3

Herman, E. S., & Chomsky, N. (2010). *Manufacturing consent: The political economy of the mass media*. Random House.

Hern, A. (2017). Netflix's biggest competitor? sleep. https://www.theguardian.com/technology/2017/apr/18/netflix-competitor-sleep-uber-facebook

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed) [OCLC: ocn694679188]. Cambridge University Press.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65.

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, *17*(6), 923–930.

Hills, T. T. (2019). The dark side of information proliferation. *Perspectives on Psychological Science*, *14*(3), 323–330.

Hodas, N. O., & Lerman, K. (2014). The simple rules of social contagion. *Scientific Reports*, *4*(1), 4343.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Holst, A. (2021). Total data volume worldwide 2010-2025. https://www.statista.com/statistics/871513/worldwide-data-created/

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin,

J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2019). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*.

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). The textcat package for n-gram based text categorization in r. *Journal of Statistical Software, Articles*, *52*(6), 1–17. https://www.jstatsoft.org/v052/i06

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences*, *118*(32).

Hotez, P. J. (2020). Combating antiscience: Are we preparing for the 2020s? *PLoS Biology*, *18*(3), e3000683.

House Select Subcommittee on the Coronavirus Crisis. (2021). Select subcommittee releases new evidence of trump administration's political meddling in coronavirus guidance, testing and treatments. https://coronavirus.house.gov/news/press-releases/select-subcommittee-releases-new-evidence-trump-administration-s-political

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Hu, M., Rao, A., Kejriwal, M., & Lerman, K. (2021). Socioeconomic correlates of anti-science attitudes in the us. *Future Internet*, *13*(6), 160.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707.

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush limbaugh and the conservative media establishment*. Oxford University Press.

Jarvstad, A., Hahn, U., Rushton, S. K., & Warren, P. A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proceedings of the National Academy of Sciences*, *110*(40), 16271–16276.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Jolley, D., & Douglas, K. M. (2014). The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology*, *105*(1), 35–56.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power. *Annual Review of Psychology*, *68*(1), 601–625.

Jungherr, A., Rivero, G., & Gayo-Avello, D. (2020). *Retooling politics: How digital media are shaping democracy*. Cambridge University Press.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision making*, *8*(4), 407–24.

Kahan, D. M. (2016). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. *Emerging trends in the social and behavioral sciences*, *29*.

Kahan, D. M., Braman, D., Slovic, P., Gastil, J., & Cohen, G. (2009). Cultural cognition of the risks and benefits of nanotechnology. *Nature Nanotechnology*, *4*(2), 87–90.

Kahan, D. M., Hoffman, D. A., Braman, D., & Evans, D. (2012). They saw a protest: Cognitive illiberalism and the speech-conduct distinction. *Stan. L. Rev.*, *64*, 851.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54–86.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735.

Kahneman, D., & Tversky, A. (1973). On the Psychology of Prediction. *Psychological Review*, *80*(4), 237–251.

Kappes, A., Faber, N. S., Kahane, G., Savulescu, J., & Crockett, M. J. (2018). Concern for others leads to vicarious optimism. *Psychological Science*, *29*(3), 379–389.

Kearney, M. W. (2019). Rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software*, *4*(42), 1829.

Khoury, M. J., & Ioannidis, J. P. (2014). Big data meets public health. *Science*, *346*(6213), 1054–1055.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Klein, C., Clutton, P., & Dunn, A. G. (2019). Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS One*, *14*(11), e0225098.

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016). *Bots and automation over twitter during the u.s. election* (Data memo). Oxford, UK.

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592.

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.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P. A., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), eaaw9011.

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162.

Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. *Frontiers in Psychology*, *8*, 1087.

Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. *The Journal of Neuroscience*, *38*(37), 7996–8010.

Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617–634.

Lanham, R. A. (2006). *The economics of attention: Style and substance in the age of information*. University of Chicago Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lazer, D. (2015). The rise of the social algorithm. *Science*, *348*(6239), 1090–1091.

Lazer, D., 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.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196.

Le, H., Maragh, R., Ekdale, B., High, A., Havens, T., & Shafiq, Z. (2019). Measuring political personalization of google news search. *The World Wide Web Conference*, 2957–2963.

Lee, N. (2021). Do policy makers listen to experts? evidence from a national survey of local and state policy makers. *American Political Science Review*, 1–12.

Lerman, K., Yan, X., & Wu, X.-Z. (2016). The" majority illusion" in social networks. *PloS One*, *11*(2), e0147617.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106–131.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). Nasa faked the moon landing—therefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, *24*(5), 622–633.

Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2009). Misinformation and the "war on terror": When memory turns fiction into fact. In W. G. Stritzke, S. Lewandowsky, D. Denemark, J. Clare, & F. Morgan (Eds.), *Terrorism and torture: An interdisciplinary perspective* (pp. 179–203). Cambridge University Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 986–1005.

Llewellyn, S. (2020). Covid-19: How to be careful with trust and expertise on social media. *BMJ*, *368*.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, *4*(11), 1102–1109.

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, *10*(1), 1–9.

Macdougall, R. (1906). On secondary bias in objective judgments. *Psychological Review*, *13*(2), 97.

Mack, A., & Rock, I. (1998). Inattentional blindness: Perception without attention. *Visual Attention*, *8*, 55–76.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*.

Mance, H. (2016). Britain has had enough of experts, says gove. https://www.ft.com/content/3be49734-29cb-11e6-83e4-abc22d5d108c

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Cognitive Processing*, *11*(2), 103–121.

Marks, J., & Baines, S. (2017). Optimistic belief updating despite inclusion of positive events. *Learning and Motivation*, *58*, 88–101.

Merriam-Webster. (n.d.). Polarization. *Merriam-webster.com dictionary*. Retrieved August 18, 2021, from https://www.merriam-webster.com/dictionary/polarization

Miller, C. (2020). How taiwan's 'civic hackers' helped find a new way to run the country. https://www.theguardian.com/world/2020/sep/27/taiwan-civic-hackers-polis-consensus-social-media-platform

Mirsch, T., Lehrer, C., & Jung, R. (2018). Making digital nudging applicable: The digital nudge design method. *Proceedings of the 39th international conference on information systems (ICIS)*.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2014). Managing self-confidence. *NBER Working paper*, *17014*.

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396.

Moore, M. (2019). *Democracy hacked: Political turmoil and information warfare in the digital age*. Oneworld Publications.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021). Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Mosleh, M., Pennycook, G., & Rand, D. (2021). Field experiments on social media. *Version 3*.

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

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human frontal–subcortical circuit and asymmetric belief updating. *Journal of Neuroscience*, *35*(42), 14077–14085.

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, *341*(6146), 647–651.

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, *5*(3), 2056305119859294.

Nature. (2020). Why nature supports joe biden for us president. https://www.nature.com/articles/d41586-020-02852-x

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*(4), 480–494.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). Reuters institute digital news report 2019. *Reuters Institute for the Study of Journalism*.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2021). Reuters institute digital news report 2021. *Reuters Institute for the Study of Journalism*.

Newman, T. P., Nisbet, E. C., & Nisbet, M. C. (2018). Climate change, cultural cognition, and media effects: Worldviews drive news selectivity, biased processing, and polarized attitudes. *Public Understanding of Science*, *27*(8), 985–1002.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Obama, B. (2017). President obama's farewell address. https://obamawhitehouse.archives.gov/farewell

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

Pant, G., Srinivasan, P., & Menczer, F. (2004). Crawling the web. *Web dynamics* (pp. 153–177). Springer.

Pariser, E. (2017). *The filter bubble: What the internet is hiding from you*. Penguin Books.

Parker, C. (2017). Brexit tweets from the morning of its announcement. *Mendeley Data*.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., Fugelsang, J. A., et al. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision making*, *10*(6), 549–563.

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.

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.

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*, *116*(7), 2521–2526.

Persily, N. (2017). The 2016 us election: Can democracy survive the internet? *Journal of Democracy*, *28*(2), 63–76.

Persily, N., & Tucker, J. A. (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Peter, C., & Koch, T. (2016). When debunking scientific myths fails (and when it does not) the backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication*, *38*(1), 3–25.

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*(1), 117.

Pew Research Center. (2014a). Americans feel better informed thanks to the internet. https://www.pewresearch.org/internet/2014/12/08/americans-feel-better-informed-thanks-to-the-internet/

Pew Research Center. (2014b). *Political polarization in the american public* (tech. rep.). Pew Research Center.

Pew Research Center. (2017). Partisan conflict and congressional outreach. https://www.pewresearch.org/politics/2017/02/23/partisan-conflict-and-congressional-outreach/

Phillips, L. D., Hays, W. L., & Edwards, W. (1966). Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics*, *1*, 7–18.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, *72*(3), 346.

Politi, J. (2018). Salvini ramps up rhetoric with attack on mandatory vaccines. https://www.ft.com/content/e513740e-761a-11e8-b326-75a27d27ea5f

Pröllochs, N. (2021). Community-based fact-checking on twitter's birdwatch platform. *arXiv*, *Version 2*.

Raab, M. H., Auer, N., Ortlieb, S. A., & Carbon, C.-C. (2013). The sarrazin effect: The presence of absurd statements in conspiracy theories makes canonical information less plausible. *Frontiers in Psychology*, *4*, 453.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26).

Reijula, S., & Hertwig, R. (2020). Self-nudging and the citizen choice architect. *Behavioural Public Policy*, 1–31.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020). Auditing radicalization pathways on youtube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42.

Roozenbeek, J., Freeman, A. L., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook et al.(2020). *Psychological Science*, 09567976211024535.

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.

Rosa, H. (2013). *Social acceleration*. Columbia University Press.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Rutjens, B. T., Heine, S. J., Sutton, R. M., & van Harreveld, F. (2018). Attitudes towards science. *Advances in experimental social psychology* (pp. 125–165). Elsevier.

Rutjens, B. T., van der Linden, S., & van der Lee, R. (2021). Science skepticism in times of covid-19. *Group Processes & Intergroup Relations*, *24*(2), 276–283.

Saldivar, J., Parra, C., Alcaraz, M., Arteta, R., & Cernuzzi, L. (2019). Civic technology for social innovation. *Computer Supported Cooperative Work (CSCW)*, *28*(1), 169–207.

Salganik, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting critical concerns into productive inquiry*, *22*, 4349–4357.

Sansonnet, J.-P., Leray, D., & Martin, J.-C. (2006). Architecture of a framework for generic assisting conversational agents. *International Workshop on Intelligent Virtual Agents*, 145–156.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Anatomy of news consumption on facebook. *Proceedings of the National Academy of Sciences*, *114*(12), 3035–3039.

Scientific American. (2020). Scientific american endorses joe biden. https://www.scientificamerican.com/article/scientific-american-endorses-joe-biden1/

Seargeant, P., & Tagg, C. (2019). Social media and the future of open debate: A user-oriented approach to facebook's filter bubble conundrum. *Discourse, Context & Media*, *27*, 41–48.

Seifert, C. M. (2017). The distributed influence of misinformation. *Journal of Applied Research in Memory and Cognition*, *6*, 397–400.

Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, *90*, 71–127.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

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), 1–9.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945.

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33.

Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., & Dolan, R. J. (2012). How Dopamine Enhances an Optimism Bias in Humans. *Current Biology*, *22*(16), 1477–1481.

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences*, *109*(42), 17058–17062.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature*, *450*(7166), 102–105.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Simon, H. A. (1957). *Models of man; social and rational*. Wiley.

Simon, H. A. (1969). *The sciences of the artificial*. MIT Press.

Simon, H. A. (1971). Designing organizations for an information rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest* (pp. 37–72).

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, *1*(1), 25–39.

Simon, J., Bass, T., Boelman, V., & Mulgan, G. (2017). Digital democracy. *The Tools Transforming Political Engagement*, 87.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*(4), 644–649.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Nature Human Behaviour*.

Spring, M. (2020). Man who believed virus was hoax loses wife to covid-19. https://www.bbc.co.uk/news/world-us-canada-53892856

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440.

Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., & Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, *573*(7772), 117–121.

Stewart, L. G., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network. *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior mining on the web*, *70*.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248.

Strandberg, K., & Grönlund, K. (2018). Online deliberation. *The Oxford handbook of deliberative democracy*, 365–377.

Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, *30*(3), 341–366.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. *2010 IEEE Second International Conference on Social Computing*, 177–184.

Sunstein, C. R. (2018). *#republic: Divided democracy in the age of social media*. Princeton University Press.

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How People Update Beliefs About Climate Change: Good News and Bad News. *Cornell Law Review*, *102*, 1431–1444.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020a). Bayesian or biased? analytic thinking and political belief updating. *Cognition*, *204*, 104375.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020b). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General*.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020c). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, *34*, 81–87.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thomson-DeVeaux, A. (2019). The politics surrounding mueller have changed a lot since he started. https://fivethirtyeight.com/features/the-politics-of-the-mueller-report/

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167–171.

Tufekci, Z. (2013). Not this one social movements, the attention economy, and microcelebrity networked activism. *American Behavioral Scientist*, *57*(7), 848–870.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 505–514.

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Tufekci, Z. (2018). Youtube, the great radicalizer. *The New York Times*, *10*, 2018.

Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185.

Turner, A. (2018). 390,000 #MeToo tweets. *data.world*. https://data.world/balexturner/390-000-metoo-tweets

Twitter. (2020). Sharing an article can spark conversation, so you may want to read it before you tweet it [tweet]. https://twitter.com/twittersupport/%20status/1270783537667551233

Twitter. (2021). Introducing birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

Uscinski, J. E., & Parent, J. M. (2014). *American conspiracy theories*. Oxford University Press.

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., et al. (2020). Using social and behavioural science to support covid-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471.

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 Prooijen, J.-W., & Douglas, K. M. (2017). Conspiracy theories as part of history: The role of societal crisis situations. *Memory Studies*, *10*(3), 323–333.

Volzhanin, I., Hahn, U., Jönsson, M., & Olsson, E. J. (2015). Individual belief revision dynamics in a group context. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2505–2510.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197–204.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.

Wei, H., Pan, Z., Hu, G., Zhang, L., Yang, H., Li, X., & Zhou, X. (2018). Identifying influential nodes based on network representation learning in complex networks. *PloS One*, *13*(7).

Weinmann, M., Schneider, C., & Vom Brocke, J. (2016). Digital nudging. *Business & Information Systems Engineering*, *58*(6), 433–436.

Weinstein, N. D. (1980). Unrealistic Optimism About Future Life Events. *Journal of Personality and Social Psychology*, *39*(5), 806.

Weinstein, N. D., & Klein, W. M. (1996). Unrealistic optimism: Present and future. *Journal of Social and Clinical Psychology*, *15*(1), 1–8.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think (U. S. Tran, Ed.). *PLOS ONE*, *11*(3), e0152719.

WhatsApp. (2019). More changes to forwarding. https://blog.whatsapp.com/more-changes-to-forwarding

Wheeler, G. (2018). Bounded rationality. https://plato.stanford.edu/entries/bounded-rationality/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wilkes, A., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, *40*(2), 361–387.

Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, *121*(11), 1–40.

Wolfe, D., & Dale, D. (2020). All of the times president trump said covid-19 will vanish. https://edition.cnn.com/interactive/2020/10/politics/covid-disappearing-trump-comment-tracker/

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126.

Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 255.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163.

Woolley, S. C. (2016). Automating power: Social bot interference in global politics. *First Monday*.

Woolley, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. *International Journal of Communication*, *10*.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120.

Yarkoni, T. (2019). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37.

Yesilada, M., & Lewandowsky, S. (2021). A systematic review: The youtube recommender system and pathways to problematic content. *PsyArXiv*, *Version 1*.

YouGov. (2020). *YouGov-Cambridge Globalism Study* (tech. rep.). YouGov. https://docs.cdn.yougov.com/2ouu9vfd10/YouGov%5C%20-%5C%20Globalism%5C%20Study%5C%20and%5C%20conspiracies%5C%20Results.pdf