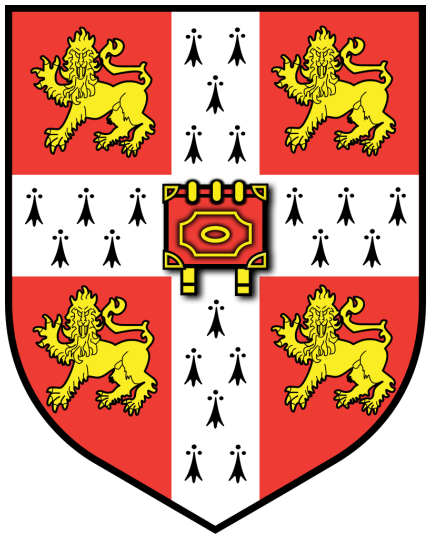
Accuracy and Social Motivations Shape Judgements of (Mis)Information

Steven James Rathje



Supervisor: Sander van der Linden
Trinity College
University of Cambridge
August 2022

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration: This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Summary

Why do people believe in and share misinformation? Some theories focus on social identity and politically motivated reasoning, arguing that people are motivated to believe and share identity-congruent news. Other theories suggest that belief in misinformation is not shaped by motivated reasoning, but is instead shaped by other factors, such as prior knowledge, lack of reflection, or inattention to accuracy.

Integrating multiple perspectives, this thesis argues that the spread of (mis)information is shaped by two (often competing) motivations: accuracy and social motivations, in combination with other factors, such as personality variables and information exposure. Through a variety of methods, including analyses of large-scale social media datasets, online experiments, network analysis, and a digital field experiment, this thesis illustrates how accuracy motivations, social motivations, and other variables shape the belief and spread of (mis)information.

Chapter 2 takes a big data approach to test whether online content that fulfills political identity motivations, such as out-group derogation and in-group favoritism, tends to receive more engagement online across eight large-scale datasets containing a total of 2.7 million tweets and Facebook posts. *Chapter 3* experimentally manipulates accuracy and social motivations for believing in and sharing true and false news headlines in a series of four online experiments with 3,364 participants. *Chapter 4* examines partisan asymmetries in the effectiveness of a popular misinformation intervention, the accuracy nudge. *Chapter 5* links survey data to the Twitter data of 2,064 participants to examine how beliefs about the COVID-19 vaccine and politics are associated with following political elites online and interacting with low-quality news sources. Finally, *Chapter 6* examines how manipulating participants' online social networks in a naturalistic setting (e.g., incentivizing people to follow and unfollow specific accounts on Twitter in a randomized controlled trial) influences beliefs about the opposing political party and the sharing of misinformation.

Preface

A version of *Chapter 2* of the thesis is published in the *Proceedings of the National Academy of Sciences* with co-authors Jay Van Bavel and Sander van der Linden. The following citation can be used:

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), e2024292118.

A version of *Chapter 3* of the thesis has been revised and resubmitted to *Nature Human Behaviour* with co-authors Jon Roozenbeek, Jay Van Bavel, and Sander van der Linden. The following pre-print can be cited:

Rathje, S., Roozenbeek, R., Van Bavel, J. J., & van der Linden, S. (2022). Accuracy and Social Motivations Shape Judgements of (Mis)information. https://doi.org/10.31234/osf.io/hkqyv

A version of *Chapter 4* of the thesis is published in *Psychological Science* with co-authors Jon Roozenbeek, Cecilie Steenbuch Traberg, Jay Van Bavel, and Sander van der Linden. The following article can be cited:

Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the editors of psychological science: meta-analysis reveals that accuracy nudges have little to no effect for US conservatives: regarding Pennycook et al. (2020). *Psychological Science.*

A version of *Chapter 5* of this thesis has been revised and resubmitted to *PNAS Nexus*, with co-authors Jon Roozenbeek, James K. He, Jay Van Bavel, and Sander van der Linden.

A version of *Chapter 6* of this thesis will be submitted to an academic journal (pending a replication study) with co-authors James K. He, Jon Roozenbeek, Trisha Harjani, Sander van der Linden, and Jay Van Bavel.

Acknowledgements

First, I would like to thank my mentors, Professors Sander van der Linden and Jay Van Bavel:

To Sander: Thank you for guiding me through the difficult process of switching supervisors, giving me the opportunity to join your vibrant lab, and for encouraging me to continue in this PhD and career path. I have learned so much from you, including how to conduct good science, how to write strong papers, how to think like a scholar, and much more. Thank you, too, for your kindness, your wisdom, your leadership, your open-mindedness. I will always treasure the experience of working with you and your lab.

To Jay: Thank you for putting so much trust in me early on, for giving me so many opportunities, for believing in me and my ideas, for being so positive, kind, and enthusiastic, for teaching me about leadership, collaboration, and team science, and for being such a selfless mentor. I can't wait to start a post-doc with your lab.

Sander and Jay, you both are ideal mentors and role models for the mentor and lab director I will (hopefully) become one day. There is so much hidden knowledge about how to navigate academia. Most everything I know about academia I learned from the two of you (or from academic Twitter).

To my Mom and Dad: There is too much to thank you for. Thank you for always encouraging me to pursue my passions, for supporting me in everything I do, and for being the best parents I could ever ask for. To my amazing brother, Billy, to Dan, to all my incredible friends, to my collaborators, to the members of the Social-Decision Making Lab and the Social Identity and Morality Lab, thank you for your continuous encouragement and support throughout this process.

Table of Contents

“The internet is built to distend our sense of identity.”

~Jia Tolentino, “The I in Internet”

“Identity = Virality”

~Ezra Klein, “Why We’re Polarized”

“Until the incentives change, Facebook will not change.”

~Frances Haugen, Senate Committee Testimony

Chapter 1. Introduction

In 2010, less than one billion people used social media. Today, as of 2022, over 4 billion people – or about half the world's population – use social media (Statista, 2022a). On average, people spend roughly 147 minutes on social media per day (Statista, 2022b). The rapid growth of social media has sparked interest from researchers in understanding how information, as well as misinformation, is spread online (Pennycook & Rand, 2021c; Persily & Tucker, 2020; Van Bavel, Harris, et al., 2021; van der Linden, 2022).

During its relatively brief existence, social media has gone through many changes. In 2009, Facebook introduced its news feed, which allowed users to see recent posts from their friends. Shortly thereafter, Twitter invented the "retweet" button, which allowed users to re-post a tweet to their followers, which was followed by the very similar "share" button on Facebook. Soon after, Facebook and Twitter introduced algorithmic timelines that would prioritize showing people "viral" content, or content with which people frequently interacted. Some scholars have argued that the rapid propensity for information to go "viral" on social media might have damaging consequences for society (Haidt, 2022) – exacerbating polarization (Van Bavel, Rathje, et al., 2021a), undermining democratic institutions (Persily & Tucker, 2020), and leading to widespread belief in falsehoods (Johnson et al., 2020a; Vosoughi et al., 2018). As such, it is important to understand how (mis)information spreads online.

While social media may be a new and rapidly evolving technology, the study of social media can be broken down into enduring psychological questions. For example, why do people believe in or share true and false information? What are the consequences of seeing or sharing (mis)information? These questions will continue to be relevant as social media and society evolve. While this thesis provides insight into current issues, such as misinformation belief and sharing on social media, it also builds upon classic psychological theories, such as social identity theory (Tajfel et al., 1979a) and motivated cognition (Kunda, 1990), and gives insight into basic psychological processes that are relevant even outside the context of social media.

This thesis will frequently use the word (mis)information as an umbrella term to describe both false information and the misuse of true information. There has been much concern about exposure to "fake news" stories online, but studies suggest that people rarely are exposed to or interact with websites that publish completely fake news stories (A. Guess et al., 2019).

Although much research focus has been placed on blatantly false news, many have recently advocated for taking a broad definition of misinformation (Traberg, 2022; van der Linden, 2022), since misleading, biased, or divisive – but not completely false – information can also have deleterious consequences. Some have argued that lack of acceptance of true information may be as or more important than belief in false news (Acerbi et al., 2022). Others have noted that misleading headlines, such as the story published with the headline "A healthy doctor died two weeks after getting a COVID-19 vaccine" (which was viewed over 50 million times on Facebook) are also harmful. While it was factually true that a doctor died after getting the COVID-19 vaccine, this headline implies an unwarranted causal connection, and is thus highly misleading (van der Linden, 2022). Finally, several have connected misinformation belief and sharing to partisan polarization, as people are often motivated to share misinformation that denigrates the political out-group (Osmundsen et al., 2020). Thus, the term (mis)information will be used throughout the thesis to refer to false news, as well as related concepts, such as misleading content, polarizing content, conspiracy theories, or the dismissal of true news.

To help relate each chapter of this thesis to the current literature in the field, this introduction will 1) review literature on why social media posts spread online and propose that content that appeals to social-identity based motivations receives more online engagement (as elaborated on in *Chapter 2*). Then, it will 2) review work on the belief and sharing of misinformation, focusing specifically on the competition between accuracy and social motivations (as discussed in *Chapters 3* and *4*). While these first two sections focus on the antecedents of (mis)information belief and sharing, the final section of the introduction will look at 3) the consequences of exposure to (mis)information on social media, setting the stage for *Chapters 5* and *6* of the thesis.

Social Motivations for Believing and Sharing (Mis)Information

Why does content go "viral" online? The following section will review literature on why people share, engage with, and believe in social media posts. While most prior work has focused on how emotions drive social media sharing behavior, this section will propose that social identity-based motives (such as out-group derogation and in-group favoritism) also drive social media sharing. This will set the stage of *Chapter 2* of the thesis, which explores whether out-

group derogation and in-group favoritism predict virality online, and whether these factors better predict virality than previously explored factors, such as emotional language.

Emotion and Virality

A substantial amount of research on social media sharing focuses on how emotions drive sharing behavior. One study found that *New York Times* articles that evoke high-arousal positive emotions (awe) or high-arousal negative emotions (anger or anxiety) tend to be shared more (Berger & Milkman, 2012). An experimental study found that putting people into a state of arousal increases the propensity to share information with others (Berger, 2011). Fake news can evoke high-arousal emotions, such as surprise and disgust, which may explain its rapid spread on social media websites (Vosoughi et al., 2018).

In addition to high-arousal emotions driving sharing behavior, other work has found a *negativity bias* in the type of information that is shared and consumed on social media platforms. For instance, an analysis of A/B tests from the website Unworthy found that news stories were more likely to be clicked on when they contained a negative word in the headline (C. E. Robertson, Pröllochs, et al., 2022). Other work has found that emotions like anger (Fan et al., 2020) and negative sentiment more broadly (Schöne et al., 2021) spread further on Twitter. This may be due in part to the "negativity bias," or the general tendency for people to pay more attention to negative as opposed to positive information (Baumeister et al., 2001; Rozin & Royzman, 2001). A 17-country study found that exposure to negative news evokes more psychophysiological arousal than exposure to positive news (Soroka et al., 2019), which could help explain the quick spread of negative news.

However, it should be noted that negativity does not go viral in all contexts: some work has argued that positive content tends to achieve more virality than negative content (Berger & Milkman, 2012; Kraft et al., 2020a; Milkman & Berger, 2014). But, interestingly, these studies did not look at social media, and instead looked at the spread of New York Times articles (Berger & Milkman, 2012; Kraft et al., 2020b) or science articles (Milkman & Berger, 2014). Additionally, there have been instances of "viral altruism," such as the ALS "Ice Bucket Challenge" in 2014, a viral social movement that prompted people to pour ice water on their heads in a social media video, make a charitable donation to support research for the disease amyotrophic lateral sclerosis (ALS), and nominate their friends to do the same. However, an

analysis of this movement found that, while this movement was viral and raised a substantial amount for ALS research, it was relatively short-lived, meaning that viral altruism may be difficult to sustain on social media (Van Der Linden, 2017). In sum, while negativity may be shared more in general on social media, there are contexts in which positivity goes "viral" as well, and more research is needed to explore the contexts and incentive structures that promote the spread of negativity and positivity online.

Another body of work has focused on the role that *moral emotions* play in driving the sharing of online content. For example, one study found that each moral-emotional word (such as "bad" or "blame") added to a Twitter post led to an estimated 20% increase in retweets (Brady et al., 2017). While this "moral contagion" effect has not been found in every context (Burton et al., 2021), a meta-analysis found that the estimated effect size of this effect across 27 studies is a 12% increase in retweets per moral-emotional word (Brady & Van Bavel, 2021a). Experimental studies find that moral and emotional words capture our attention more than non-moral and non-emotional words (Brady et al., 2020), which may help increase their spread in an "attention economy," where several social media posts are competing for our attention (Williams, 2018). Related work has suggested that moral "outrage," or anger and disgust at moral violations, is likely to go viral online (Crockett, 2017). Outrage expression can also be reinforced by the incentive structure of the social media environment: people who have received positive social feedback (e.g., favorites, retweets, etc.) for posting an outrage evoking tweet in the past are more likely to post more outrage-evoking tweets in the future. Furthermore, people whose online network expressed outrage were more likely to express outrage themselves (Brady et al., 2021). Relatedly, people are more likely to share controversial news (Kim & Ihm, 2020) or social media posts from politicians that express "indignant disagreement" (Messing & Weisel, 2017). The most politically extreme politicians also have the most Twitter followers, potentially because they post more controversial, attention-grabbing content (Hong & Kim, 2016).

Beyond Emotion: Social Identity and Virality

While prior work has focused on how emotion drives social media virality, there are limitations to this approach. Much of this work looks at the use of emotional words out of context, by, for instance, counting the number of words representing emotions using dictionary-based approaches (Messing & Weisel, 2017), or estimating whether social media posts contain

discrete emotions such as "outrage" through machine learning classifiers (Brady et al., 2021). A limitation of these approaches is that they fail to capture the context of social media posts, which is potentially problematic, since people can express strong emotions about serious political events or toward trivial occurrences. Moving beyond the approach of looking at how emotions drive virality, I argue that content that appeals to social identity-based motivations, such as praising one's ingroup or derogating one's outgroup, will receive more engagement online.

This hypothesis is based in part on the predictions of social identity theory, a classic social psychological theory that states that people are motivated to belong to social groups. Specifically, people categorize themselves and similar others into in-groups, or groups to which they belong, and categorize dissimilar others into out-groups, or groups to which they do not belong (Tajfel et al., 1979). Social identity theory can help explain why people engage in in-group favoritism, or preferential treatment toward ingroups, as well as out-group derogation, or the expression of negative emotions and behaviors toward outgroups. This inter-group bias, or preferential treatment of the ingroup over the outgroup, may fulfill a number of fundamental needs, such as the need for belonging (Baumeister & Leary, 1995), self-esteem (Rubin & Hewstone, 2016), and the need to feel positive distinctiveness (Jans et al., 2011) or certainty about the world (Hogg, 2000).

Relatedly, self-categorization theory (Turner et al., 1987) suggests that the social groups to which people belong are highly flexible, and when a certain social identity is activated or made salient, people will act in accordance with that identity (Turner et al., 1987). Classic "minimal group" experiments have demonstrated that even when people are assigned to arbitrary groups – for instance, groups based on liking paintings by the artist Paul Klee versus the artist Wassily Kandinsky – they express in-group favoritism and out-group derogation (Pechar & Kranton, 2017; Tajfel et al., 1971). In other words, merely belonging to a group, whether or not that group is important, can cause inter-group bias.

Social identity and self-categorization theory can help explain why people engage with content online. Scholars have argued that social media is an environment where group identities are highly salient (Brady, Crockett, et al., 2019), which should, according to self-categorization theory, make people act in accordance with them. Indeed, work has found that people are increasingly defining themselves with political identity language (e.g., "Democrat" or "MAGA") in their Twitter bios (Rogers & Jones, 2021), supporting the idea that political identities, which

are a form of social identity (Van Bavel & Pereira, 2018), are highly salient on the social media platform Twitter. Other design features of social media, such as greater anonymity and distance from others, have been suggested to promote deindividuation and conformity to group norms, which may lead people's individual identities to be subsumed by group identities online (Brady, Crockett, et al., 2019).

Emotional expression online may fulfill social identity-related goals. For instance, outrage is often directed toward outgroup members, acting as a form of outgroup derogation (Brady, Crockett, et al., 2019). Supporting the idea that emotions fulfill group-related functions, outrage-evoking tweets tend to be shared within – but not between – groups (Brady et al., 2017). Experimental studies reveal that people who express outrage tend to be viewed as good ingroup members, but are perceived as less open-minded to outgroup members (Brady & Van Bavel, 2021b). Others have proposed that sharing high-arousal emotional content (whether or not this content is positive or negative) might encourage social bonding, as shared emotional experiences can bring people together (Milkman & Berger, 2014). However, while emotions have been proposed to fill a number of inter-group functions (Mackie et al., 2016), much of the current work on social media virality still examines emotional language out of context rather than looking into social identity motivations in particular.

As an anecdotal example of social identity-related content predicting online engagement, executives at Buzzfeed, a website specializing in creating viral content for social media platforms, reportedly noticed that identity-related content was successful online. Headlines written with the following structure: "X things only X would understand" would do particularly well. For example, titles like "13 Struggles All Left-Handers Know to Be True," "27 Struggles You'll Only Understand If You Were Born Before 1995" or "14 Things Only Anxious People Will Understand" would gain a lot of engagement (Klein, 2020). Interestingly, these headlines seem to appeal to relatively unimportant identities, such as being left-handed. This resembles the classic "minimal group" experiments, whereby people would identify with seemingly arbitrary and unimportant groups (Pechar & Kranton, 2017; Tajfel et al., 1971).

Empirical work has also supported the idea that social identity motives might drive engagement with media. For example, when in-group members were presented in a more positive light in a television pilot proposal, the proposal was given more positive ratings (Joyce & Harwood, 2020). People tend to believe in and share identity-congruent true and false news in

survey experiments (Pereira et al., 2018; Van Bavel & Pereira, 2018). Similarly, people tend to engage more with content that seems personally relevant to them. As an example, people are much more likely to highlight passages in Kindle books that contain the word "you" in them (Orvell et al., 2020). A large amount of literature on *selective exposure* suggests that people prefer to consume media that confirms (as opposed to contradicts) their beliefs (Frimer et al., 2017; Hart et al., 2009).

People also self-select into identity-congruent social networks. People are more likely to follow back individuals who share their political identity on Twitter (Mosleh, Martel, et al., 2021a). Additionally, a large amount of work suggests that people cluster into "echo chambers" online (Bakshy et al., 2015; Barberá et al., 2015; Cinelli et al., 2021; Pariser, 2011), seeing content that reinforces their existing beliefs and identities. However, there is debate over the strength of the echo chamber phenomenon, with some arguing that many people are exposed to more politically diverse news – or less political news in general – than many might expect (Eady et al., 2019; A. Guess et al., 2018). It should also be noted that people exist in offline echo chambers (Brown & Enos, 2021; McPherson et al., 2001a), sorting into neighborhoods with like-minded others, and it is unclear whether online or offline echo chambers are stronger. Indeed, some work suggests that social media exposes people to more voices that they would not normally encounter in their offline networks (Bor & Petersen, 2022), bursting people's echo chambers open. However, there is also evidence that online and offline echo chambers interact with one another (Bastos et al., 2018). Thus, as the world becomes increasingly online, it may become difficult to disentangle online and offline echo chambers.

Social media sharing is a public, self-conscious activity, so it is likely to reflect motivations to present oneself in a positive way. One might do this by, for instance, sharing content that reflects well on oneself, is liked by one's group, or gives one higher social status. One analysis found that people were more likely to share scientific articles if they thought these articles reflected well on them (Milkman & Berger, 2014). There may be various ways to gain social status, including sharing interesting, surprising, emotional, or useful content, or content that appeals to people's group biases (derogating the out-group or praising the in-group). Status motivations can also help explain the sharing of hostile, false, or otherwise problematic content. This is perhaps counterintuitive, as most people report that sharing fake news hurts their reputation (Altay et al., 2019). However, one study found that people who score high in the

personality trait "status-driven risk-taking" are most likely to engage in hostile political discussions online (Bor & Petersen, 2022). Additionally, people who are highly partisan and hate the opposing party are most likely to share fake news about the opposing party (Osmundsen et al., 2021). In sum, people may share both helpful and harmful content to gain social status and receive approval from their online social groups.

In-Group Favoritism vs. Out-Group Derogation

What types of identity-related expressions should succeed most online? Specifically, should in-group favoritism or out-group derogation receive more engagement? There has been debate over whether in-group favoritism is stronger than out-group derogation (Brewer, 2017). Some have argued that in-group love might be a more important driver of inter-group behavior than out-group hate (Amira et al., 2021; Appiah et al., 2013; Mummendey & Otten, 1998). However, inter-group competition can increase out-group derogation (Halevy et al., 2008), partially by increasing the salience of social identity (Cikara et al., 2011). Realistic group conflict theory, another classic social psychological theory, states that intergroup hostility can develop because of conflict over shared resources (Relations & Sherif, 1961). Out-group derogation may also be particularly strong when the out-group is perceived as immoral (Parker & Janoff-Bulman, 2013).

One instance where out-group negativity may supersede in-group positivity is the conflict between Republicans and Democrats in the United States. It has been noted that partisan conflict tends to abide by "negative partisanship," whereby partisan identities are more defined by dislike of the opposing party as opposed to favoritism toward one's own party (Abramowitz & Webster, 2016, 2018). In the United States, out-party hate has been increasing over time, and has become a bigger predictor of voting than in-party love (Finkel et al., 2020). Additionally, out-party hostility is one of the strongest predictors of sharing fake news online (Osmundson et al., 2021). Some scholars have called this strong partisan conflict in the United States political sectarianism, or a sense of othering, aversion, and moralization toward the opposing party (Finkel et al., 2020). In the United States, where negative partisanship is strong, out-group animosity may be a better predictor of virality than in-party favoritism.

Chapter 2 of the thesis helps advance the perspective that social identity motivations contribute to social media sharing decisions using large-scale, real-world datasets of social media

posts from US congress members and partisan news sources on Facebook and Twitter. The study examines whether words about the political out-group, words about the political in-group, positive emotion words, negative emotion words, and moral-emotional words predict engagement on social media. Replicating prior work, positive language lead to fewer shares and retweets, whereas negative language and moral-emotional language lead to more shares and retweets (Brady et al., 2017; C. E. Robertson, Pröllochs, et al., 2022). Words about the political in-group do not predict increased engagement online; however, each additional word about the political out-group leads to a 67% increase in retweets and shares. Out-group language also predicted "angry" reactions and "haha" reactions, indicating that mentions of the out-group likely reflect out-party animosity. Mirroring findings on negative partisanship (Abramowitz & Webster, 2018) and political sectarianism (Finkel et al., 2020), out-group derogation appears to be a better predictor of virality than in-group favoritism. This study advances the literature on social media virality by demonstrating how social identity motivations contribute to virality. Alarmingly, it also suggests that social media may be incentivizing politicians and news media organizations to create content that derogates the out-group, which could potentially contribute to affective polarization, or animosity toward the opposing party (Druckman & Levendusky, 2019).

The Interaction Between Accuracy and Social Motivations

Related to why people share polarizing information online is the question of why people tend to share *misinformation* online. Misinformation is an umbrella term that is often used to refer to fabricated news sites, conspiracy theories, political rumors, disinformation, or other forms of false content (Van Bavel, Harris, et al., 2021). Many people report wanting to be accurate (Pennycook, Epstein, et al., 2021a) and believe that sharing fake news would hurt their reputation (Altay et al., 2020). So, why does false content appear to spread rapidly online (Vosoughi et al., 2018)?

There are several different theories of why people believe in and share misinformation. Some focus on cognitive factors, suggesting that prior beliefs, knowledge, media literacy skills, reflection, and attention to accuracy all predict belief in fake news (Pennycook & Rand, 2021c). Others focus on motivational factors, suggesting that motivated reasoning, partisan polarization,

and ideology all predict the belief and sharing of fake news (Van Bavel, Harris, et al., 2021; van der Linden, 2022). The following section will review prior research on why people share misinformation online and potential interventions for reducing the spread of misinformation. It will address key debates in the literature and outline how *Chapters 3-4* aim to reconcile these debates.

Motivated Reasoning Account

According to the theory of motivated reasoning, while people are often motivated to be accurate, they can also have directional motivations that interfere with their accuracy motivations (Kunda, 1990), such as directional motivations to protect one's partisan identity (Bolsen et al., 2014). Partisans have been shown to uncritically accept information that supports their pre-existing beliefs and be skeptical information that contradicts their beliefs (Taber & Lodge, 2006). Similar models, such as the Identity-Based Model of Political Belief, suggest that people's accuracy goals compete with other identity-related goals, such as status and belonging goals (Van Bavel & Pereira, 2018). Many scholars have suggested that people fall for fake news because they are motivated to believe that falsehoods that support their worldview or identity are true (Van Bavel, Harris, et al., 2021).

One more extreme version of the motivated reasoning account of belief suggests that people who are better at reasoning, are, ironically, more likely to engage in motivated reasoning, using their superior reasoning skills to search for reasons to support their desired beliefs (Kahan, 2012). This conceptualization of motivated reasoning has sometimes been called "motivated numeracy," (Kahan et al., 2013) "motivated reflection," (Batailler et al., 2021) "identity-protective cognition," (Kahan et al., 2007) or "motivated reasoning system 2" (Pennycook & Rand, 2018). As an example, one study found that people higher in numerical skills are better at solving politically-neutral math problems, but, counterintuitively, were worse at solving math problems that had answers that went against their political beliefs. For instance, the study found that liberals with better numerical skills were more likely to solve math problems incorrectly if the math problem's answer showed that gun control increased crime (Kahan et al., 2013). Other work has shown that more educated individuals tend to be more polarized about controversial science topics, such as climate change (Drummond & Fischhoff, 2017; Kahan et al., 2012).

These findings can be interpreted through the perspective that people use their superior reasoning ability to support conclusions that they want to believe rather than seeking out the truth.

Challenges to the Motivated Reasoning Perspective

There have been numerous challenges to the motivated reasoning account. For example, while many instances of partisan differences in opinion can be attributed to motivated reasoning, they can also be explained by selective exposure to different information (Druckman & McGrath, 2019; Tappin et al., 2020a; van der Linden, 2022). For instance, if a Republican does not believe in climate change, it may not be because of motivated reasoning; instead, it might reflect selective exposure to information doubting climate change from partisan media sources. Indeed, when presented with evidence about the scientific consensus surrounding climate change, both Republicans and Democrats updated their beliefs about climate change, even though their prior beliefs about climate change were highly polarized (van der Linden et al., 2018) The fact that people generally change their beliefs in accordance with evidence has been said to challenge the motivated reasoning perspective. Indeed, people tend to change their beliefs along with evidence even when the evidence contradicts partisan cues (Tappin et al., 2022). With regards to fake news specifically, it has been argued that the partisan differences in belief in fake news, instead of reflecting motivated reasoning, may reflect "unbiased rational (e.g., Bayesian) inference built on prior factual beliefs" (Pennycook & Rand, 2021c).

There have also been a number of challenges to the "motivated numeracy" or "identity-protective cognition" account. For example, the classic experiment showing that people high in numerical skills show more partisan bias has faced replication challenges (Persson et al., 2021). Furthermore, numeracy and cognitive reflection, rather amplifying partisan bias, tend to protect people from falling for fake news (Pennycook & Rand, 2018; Roozenbeek et al., 2020). This has been said to support a "classical reasoning" perspective, whereby greater reasoning ability helps people seek out the truth instead of helping them justify their beliefs (Pennycook & Rand, 2018).

While there is limited evidence for the "motivated numeracy" or "identity-protective" cognition account, this does not rule out other theories of motivated reasoning. A key challenge for future research is to isolate the *causal* role accuracy and directional motivations play in shaping belief (van der Linden, 2022). This will help rule out alternative explanations for partisan bias in belief, such as the idea that partisan bias simply arises because people are

exposed to different news sources and thus have different beliefs (Druckman & McGrath, 2019; Tappin et al., 2020a). Some prior work has already shown that motivating people to be accurate by, for instance, paying them for accurate responses can reduce biased perceptions of the economy (Prior et al., 2015). Similarly, priming one's group identity or moral values can also influence beliefs (Bayes et al., 2020a; Bolsen et al., 2014). However, studies that have used incentives to motivate accurate beliefs about fake news have yielded mixed results. For instance, while one study found that accuracy incentives can improve the detection of scientific misinformation (Panizza et al., 2021), another study found that accuracy incentives backfired, increasing belief in fake news items (Aslett et al., 2021). Because of these conflicting results, a systematic investigation is needed to explore how various motivations causally influence the belief in and sharing of fake news.

To address these questions, *Chapter 3* of the thesis experimentally manipulates people's motivations to be accurate by providing them with financial incentives to correctly identify true versus false headlines in four large online experiments. It also manipulates people's social or partisan-identity motivations to share content that will appeal to one's in-group online and examines how these interact with accuracy motivations. Altogether, the studies find that accuracy incentives causally improve accuracy and social motivations causally decrease accuracy. However, accuracy incentives primarly increase the perceived accuracy of politically-incongruent true news, and do not have an effect on false news. Manipulating social motivations also led people to report greater intentions to share politically-congruent true and false articles. This chapter then addresses how these results fit in with other accounts of fake news sharing, which will be described below. While motivation is an important part of understanding why people believe in and share fake news, it is only one piece of the puzzle, and effort should be made to understand how motivation in combination with other factors contribute to misinformation belief and spread.

Inattention Account

The inattention account is an account of misinformation sharing that proposes that people often share fake news because they are not attending to accuracy (Pennycook, Epstein, et al., 2021a; Pennycook & Rand, 2021c). Contrary to the perspective that people purposely share fake news, surveys suggest that most people report being highly motivated to share accurate content

online (Pennycook, Epstein, et al., 2021a). However, people's judgements of which headlines they intend to share online do not always align with the headlines people rate as accurate. This dissociation between accuracy judgements and sharing intentions can be explained, in part, by people not thinking about accuracy. For example, experiments show that when people are "primed" or "nudged" to think about accuracy, they subsequently report intentions to share more accurate headlines (Pennycook, McPhetres, et al., 2020; Pennycook, Epstein, et al., 2021a). Other studies have shown that simply asking people to stop and think before sharing a news headline improves the quality of news people intend to share (L. Fazio, 2020). Similar findings suggest that performance on the cognitive reflection test, which measures, in part, a willingness to stop and reflect (Thomson & Oppenheimer, 2016), but also captures other factors, such as intelligence (Otero et al., 2022), correlates with the ability to identify fake news (Pennycook & Rand, 2018). *Chapter 3* of the thesis includes a measure of cognitive reflection to examine how this variable predicts the ability to discern between true and false headline in comparison to other variables.

While the "accuracy nudge" intervention has faced replication challenge (Roozenbeek et al., 2021), it has been shown to have small effects on sharing discernment in a larger meta-analysis (Pennycook & Rand, 2021a). Importantly, the inattention account has been said to contradict the idea that fake news sharing is driven by motivated reasoning and partisanship (Pennycook, Epstein, et al., 2021a). *Chapter 4* of the thesis suggests that inattention may interact with factors such as motivation and partisanship. This chapter presents a meta-analysis of prominent papers on the accuracy nudge or accuracy prime intervention (Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2021a; Roozenbeek et al., 2021), finding that the accuracy nudge is less effective for US conservative or Republican participants as compared to US liberal or Democrats participants. While it is unclear exactly why this pattern exists, this chapter suggests that cognitive accounts of fake news sharing cannot be completely separated from accounts that focus on motivation or partisanship.

Media Literacy and Inoculation

Beyond motivational factors and inattention, one additional reason people might fall for fake news is that they lack knowledge and media literacy skills. Studies suggest that both political knowledge (Jardina & Traugott, 2019) and media literacy knowledge (Jones-Jang et al.,

2021) can protect against fake news belief. Furthermore, interventions that teach media literacy or that "inoculate" people against common manipulation tactics can help protect people against fake news. For instance, a short digital media literacy intervention improved discernment between true and false news in the United States and India (A. M. Guess, Lerner, et al., 2020a). "Inoculation" and "pre-bunking" interventions that expose people to the common manipulation tactics of those who peddle fake news (polarization, false dichotomies, and emotional language, etc.) can have long-lasting effects on people's ability to detect fake news (Maertens et al., 2020; Roozenbeek et al., 2022; Roozenbeek & van der Linden, 2019a). One of the benefits of "pre-bunking" misinformation through inoculation interventions is that "de-bunking" or fact-checking does not completely eradicate belief in misinformation, due its continued influence, or the finding that people will often continue to rely on misinformation even after that misinformation has been corrected (Lewandowsky et al., 2012). Inoculation interventions also have relatively large effect sizes (Banas & Rains, 2010). Thus, while some interventions should aim to motivate people to be more accurate, motivation can only be so effective if people lack knowledge, media literacy, or an understanding of the manipulation techniques present in misinformation. *Chapter 3* contains a measure of political knowledge to understand how knowledge compares to other factors in predicting misinformation belief and spread.

Partisan Polarization Account

The partisan polarization account of fake news sharing suggests that people are often motivated by partisan animosity when sharing fake news. Indeed, one large study found that the biggest predictor of fake news sharing was animosity toward the opposing party (Osmundsen et al., 2021). Furthermore, other studies suggest that extremely few people share fake news – indeed, 1% of users are responsible for 80% of the fake news sharing – and those who do share fake news tend to be older and conservative (Grinberg et al., 2019; A. Guess et al., 2019). This suggests that a small number of politically-motivated individuals may be spreading fake news to derogate the opposing party, even if they do not believe this news is true. Drawing on evolutionary psychology, some have proposed that disinformation (or deliberately spread false news) helps mobilize an in-group against an out-group (Petersen, 2020). In this sense, disinformation helps with group coordination, signaling one's loyalty to an in-group and helping to justify and coordinate conflict toward an out-group (Petersen et al., 2020). Supporting this

theory, experimentally manipulating perceptions of conflict leads people to want to share more fake news about an out-group (Osmundsen et al., 2022). *Chapter 3* of the thesis includes a measure of affective polarization to help address this account of fake news sharing and compare it to other accounts.

Partisan Asymmetries

Many studies have found that conservatives, at least in the US, tend to believe in and share more misinformation than liberals (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Lawson & Kakkar, 2021; Roozenbeek et al., 2022b; Roozenbeek & van der Linden, 2019a; Van Bavel & Pereira, 2018). This is even true with representative samples of popular fake news stories (Garrett & Bond, 2021) and apolitical fake news (Pereira et al., 2018). Conservatives also tend to score higher on non-political measures of conspiracy mentality (van der Linden, Panagopoulos, et al., 2021), which is related to misinformation belief, and this pattern can be observed across cultures (Imhoff et al., 2022). However, one recent paper does not find reliable evidence for an association between conservatism and conspiracy belief (Enders et al., 2022), instead arguing that this asymmetry is influenced by the conspiracy theories employed by researchers in survey questions and the socio-political context in which conspiracy beliefs are measured. Nonetheless, the majority of articles have found a robust association between conservatism and belief in misinformation and conspiracy theories, raising the question of why this association might exist.

There are multiple possible reasons why many studies report that conservatives are more likely to believe in misinformation and conspiracies. One potential reason is that conservatives have lower media literacy or less accurate knowledge about the world. This could be possible due to the information environments to which conservatives are exposed. One study found that Donald Trump was perhaps one of the largest sources of coronavirus misinformation during the early stages of the pandemic (Evanega et al., 2020). Other work has shown that conservatives tend to be exposed to more fact-checked false claims on Twitter (Mosleh & Rand, 2021). Thus, it is possible that US conservatives have worse information diets and are thus more likely to believe false claims. Another interpretation behind this asymmetry focuses more on motivation. Some scholars suggest conservatives have a greater desire to believe in and share misinformation due to asymmetries in certain psychological motivations (Jost et al., 2003, 2018). For instance,

conservatives are more likely to prioritize conformity, desire a shared reality with like-minded others, and maintain homogenous social networks, which are all conducive to the spread of misinformation (Jost et al., 2018).

Chapter 3 helps investigate potential explanations behind these asymmetries. It finds that incentivizing conservatives to be accurate can close a substantial portion of the gap in accuracy between liberals and conservatives. This suggests that conservatives' greater tendency to believe in and share misinformation is not due to lack of knowledge or ability alone, but can also be explained, in part, by a lack of motivation to be accurate. *Chapter 4* of this thesis extends on the finding that conservatives believe in and share more misinformation, showing that misinformation-reduction interventions may also be less effective for conservatives.

Toward an Integrated Account of Fake News Belief and Sharing

Some scholars have tried to construct integrated accounts of misinformation belief and sharing, focusing on a number of risk factors, such as motivation, ideology, reflection, morality, personality, and other factors (Van Bavel, Harris, et al., 2021). An investigation using computational modeling found that the integrated account of misinformation belief developed by Van Bavel et. al (2021) best predicted misinformation belief. Thus, cognitive factors (such as inattention, lack of reflection, prior beliefs, familiarity, knowledge, media literacy skills, numerical ability, etc.) as well as motivational factors (partisan identity, polarization, motivated reasoning, emotion, etc.) likely both play a role in misinformation belief and sharing. Other work suggests that heuristics such as processing fluency can promote belief in misinformation, which explains why repeated misinformation appears more accurate (L. K. Fazio et al., 2015; Pennycook et al., 2018a). Lastly, personality factors may also explain misinformation sharing, since conservatives who score low in the personality trait conscientiousness (Lawson & Kakkar, 2021) and those who score high in a personality trait called need for chaos, or a desire to incite chaos and destruction in society (Arceneaux et al., 2021), tend to share more misinformation. Researchers should continue to integrate disparate factors into an integrated account to better understand the phenomena of misinformation spread and design misinformation-reduction interventions. *Chapter 3* helps contribute to the effort to form an integrated account, demonstrating how motivation, ideology, cognitive reflection, and affective polarization all distinctly contribute to the belief and sharing of misinformation.

The Consequences of Online (Mis)Information Exposure

The first two sections cover the antecedents of (mis)information sharing, focusing specifically on accuracy and social identity motivations (which *Chapters 2-4* of the thesis address). This next section focuses on the consequences of online (mis)information exposure (which *Chapters 5* and *6* of the thesis address). I argue that people use their social media feeds to infer social norms and to acquire information about the world. But, because social media feeds amplify certain types of information (e.g., moral and emotional content, negative content about one's out-group, etc.), they may present people with distorted information about the world, which can, in turn, distort people's beliefs. They may do this through changing perceptions of social norms, exposing people to polarizing party cues, or by repeating misperceptions and changing people's knowledge about the world. I discuss innovations in measuring how (mis)information exposure influences belief through linking social media data to survey data (*Chapter 5*) and conducting social media field experiments (*Chapter 6*).

How Social Media Use is Related to (Mis)information Belief and Polarization

People are highly sensitive to social norms. Classic social psychological studies suggest that people can claim implausible things are true if they feel pressure to conform with others (Asch, 1951), and more recent work suggests that social norm interventions can be highly effective for inducing behavior change (Cialdini & Goldstein, 2004; Goldstein et al., 2008; Melnyk et al., 2010). Social media feeds can provide people with the chance to infer norms about their community. Recent studies indicate that people engage in norm learning on social media (Brady et al., 2021; Wojcieszak et al., 2020). For example, people's expression of outrage online is influenced by norm learning (Brady et al., 2021).

However, because social media tends to amplify certain types of content, such as negative content (C. E. Robertson, Pröllochs, et al., 2022), content that expresses animosity toward the out-party (Rathje, Van Bavel, & van der Linden, 2021c; Yu et al., 2021), or surprising (but false) content (Vosoughi et al., 2018), it may contribute to false norms. For instance, people who are consistently exposed to negative information about the opposing party online might develop misperceptions about the other party. Supporting the idea that social media usage contributes to (false) norms, one study found that heavy social media usage was correlated with

greater false consensus effects (Bunker & Varnum, 2021) – or the misguided belief that others hold the same beliefs as oneself (Ross et al., 1977).

The promotion of false social norms may have detrimental consequences for polarization and misinformation belief. Research suggests that "perceived" polarization, or false beliefs about how polarized others truly are, can exacerbate actual polarization (Ahler, 2014; Lees & Cikara, 2021). Partisans already hold misperceptions about the opposing party: for instance, people think that 32% of Democrats are LGBTQ (when the true amount is 6%), and people think that 38% of Republicans earn more than $250,000 per year (when the true amount is 2%). Affective polarization may be driven in part by misperceptions of the opposing party, since interventions that correct false beliefs about the opposing party can reduce affective polarization (Lees & Cikara, 2021; Ruggeri et al., 2021). Social media's tendency to amplify negative information about the out-party may promote false norms about how polarized people actually are, which could, in turn, increase polarization.

In addition to contributing to misperceptions about the opposing party, social media might contribute to other forms of (mis)perceptions, such as false beliefs about the COVID-19 vaccine, climate change, or other issues. A number of survey experiments illustrate that misinformation exposure can have deleterious consequences. For instance, exposure to climate change misinformation can reduce belief in climate change (Van der Linden et al., 2017a), and exposure to vaccine misinformation can reduce confidence in the COVID-19 vaccine (Loomba et al., 2021a). The mere repetition of misinformation can increase belief in it (Pennycook et al., 2018a), as well as the perceived morality of sharing it (Effron & Raj, 2020). Analysis of rumor cascades on Twitter suggests that the sharing of false news online is driven in part by "herding" behavior and conformity (Pröllochs & Feuerriegel, 2022), suggesting that seeing many people sharing misinformation may increase people's own tendency to share misinformation.

People are also sensitive to party elite cues (Bullock, 2020), or messages from politicians that they support. If people believe their own political party supports a given policy, they are much more likely to support that policy, regardless of its actual content (Cohen, 2003). For example, US Republicans, who tend to be more hesitant about getting the vaccine (Loomba et al., 2021a), report greater intentions to receive the vaccine if they see an endorsement of the vaccine by Donald Trump (S. L. Pink et al., 2021). A large-scale field experiment created a YouTube advertisement presenting Donald Trump endorsing the COVID-19 vaccine and showed

this advertisement to millions of people in 1,014 counties in the United States, and found an increase in vaccine uptake in the counties that were shown this advertisement (Larsen et al., 2022). Since people frequently follow and receive messages from politicians online, exposure to partisan cues on social media may contribute to belief in (mis)information.

Much of the work on how social media usage affects polarization or belief in misinformation comes from independently using either survey experiments or social media data. More recent work, however, has started linking survey data about people's beliefs with their social media data to examine how real-world (as opposed to self-reported) social media usage is related to people's beliefs. For example, one study linked people's Facebook data to survey data, finding that older people and political conservatives share more misinformation on social media (A. Guess et al., 2019). Another study linked participants' Twitter data to their survey data, finding that those high in affective polarization were more likely to share misinformation online (Osmundsen et al., 2021). A similar study found that those who scored higher on the cognitive reflection test tended to share less misinformation on Twitter (Mosleh, Pennycook, et al., 2021). Studies such as these are a major methodological innovation, allowing social scientists to see how people's private beliefs correlate with public behavior on social media platforms (Al Baghal et al., 2019; Sloan et al., 2020).

Building on this methodological innovation, *Chapter 5* examines how people's beliefs about COVID-19 vaccination relate to social media behavior. In a sample of 2,064 total participants, it examines how social media favoriting, retweeting, and following behavior relate to vaccine hesitancy. The results reveal that following conservative politicians in the US (but not the UK), as well as tweeting, favoriting, and following low-quality news sources is associated with vaccine hesitancy. Additional network analysis finds that centrality within a "conservative" echo chamber in the US predicts vaccine hesitancy. Moving beyond survey research on misinformation and vaccine hesitancy or analysis of social media datasets on their own, these results demonstrate that different ways of interacting with misinformation and party elite cues online are associated with self-reported beliefs about the vaccine. A large sample of participants for this paper were collected from an app I designed called "Have I Shared Fake News" (https://newsfeedback.shinyapps.io/HaveISharedFakeNews/), in which people could enter their Twitter handles and answer free response questions in return for feedback about the quality of news they shared online. This app was shared widely online, and thousands of people have used

it, providing us with a very large sample of social media data linked with survey data. Apps like this one present a novel opportunity for researchers to collect massive datasets from participants around the world.

Testing the Causal Effects of Social Media Use

While research linking self-reported data to social media data can give us a precise view of how online behavior is correlated with beliefs, the causal role of social media usage is still unclear. For instance, while exposure to misinformation on social media may cause vaccine hesitancy, people who are vaccine-hesitant may also seek out information online that confirms their beliefs. Indeed, people generally abide by the homophily principle (McPherson et al., 2001b), choosing offline and online social networks consisting of people who are similar to them politically (Brown & Enos, 2021; Mosleh et al., 2020) and otherwise. Thus, it is difficult to tease apart whether people are engaging in homophily or are being influenced by their social networks (Aral et al., 2009a).

Some experimental studies have explored the causal effect of social media usage on beliefs and behavior. For instance, a large field experiment conducted in collaboration with Facebook found that increasing the amount of positive or negative posts in people's social media feeds had a causal effect on the language people used in their own social media posts (Kramer et al., 2014). Furthermore, a large-scale experiment found that a Facebook design feature showing that one's Facebook friends voted had a causal effect on voting behavior (Bond et al., 2012). Participants who were incentivized to follow a Twitter bot for one month that retweeted messages from the opposing political party showed increases in ideological polarization (Bail et al., 2018). However, contrary to these results, a separate field experiment encouraging people to like Facebook pages that share news from the opposing party reduced affective polarization (Levy, 2021a). Political advertising (Kalla & Broockman, 2018) and exposure to partisan media (A. M. Guess et al., 2021) often has minimal effects on belief, so it is possible that the causal effect of messages from social media is small. One analysis found that polarization on Reddit was primarily driven by conservative users joining the platform, rather than certain features of Reddit polarizing users (Waller & Anderson, 2021). Some statistical models have aimed to distinguish homophily-based explanations from online social influence (Aral et al., 2009a);

however, it is difficult to examine the causal role of one's online social network without experimental manipulation.

Some field experiments have examined the causal impact of deactivating social media platforms such as Facebook. However, these experiments have had conflicting results. For instance, one randomized controlled trial in the United States found that deleting Facebook for one month reduced issue polarization and marginally reduced affective polarization, with changes roughly 42% as large as the total change in polarization from 1996 to 2018 (Allcott et al., 2019). However, a replication of this study in Bosnia found that deleting Facebook during genocide remembrance week actually increased ethnic polarization, especially if participants had particularly homogenous offline social networks (Asimovic et al., 2021). These studies indicate that the causal effects of social media usage might differ greatly depending on the ways in which people use social media. For instance, some evidence suggests that social media may have more beneficial outcomes in the Global South or less established democracies, facilitating access to news and important information, but may have more harmful effects in established democracies such as the United States (Lorenz-Spreen et al., 2021). Other work has suggested that passive Facebook use (e.g., scrolling as opposed to commenting and messaging) is associated with worse subjective well-being (Verduyn et al., 2015). These studies suggest that one's social and cultural context, as well as the idiosyncratic ways in which people use social media, may moderate social media's impact on polarization, well-being, and other outcomes.

To help understand how specific ways of using social media causally contribute to polarization, *Chapter 6* of the thesis reports results from a randomized-controlled trial that incentivizes participants to unfollow highly partisan accounts and follow accounts about non-partisan topics (science, space, etc.) for one month. This experiment found that changing social media following behavior reduced affective polarization, improved well-being, and led people to report a more positive perception of their Twitter feed. These results indicate that polarization on social media is not entirely driven by self-selection and that changing the make-up of one's social media feed can have a causal effect on polarization.

Overview

This introduction presented an overview of the antecedents and consequences of (mis)information exposure and an analysis of how the upcoming chapters fit into the prior literature. *Chapter 2-4* focus on how accuracy and social motivations drive (mis)information belief and sharing. Specifically, *Chapter 2* focuses on how social identity motivations (in-group favoritism and out-group derogation) drive virality online (Rathje, Van Bavel & van der Linden, *Proceedings of the National Academy of Sciences*). *Chapter 3* experimentally manipulates accuracy and social identity motivations and examines how these motivations causally impact the belief and sharing of (mis)information (Rathje et. al, R&R at *Nature Human Behaviour*). *Chapter 4* further explores the interaction between accuracy and social motivations by exploring partisan asymmetries in the success of a popular misinformation-reduction intervention (Rathje et. al, *Psychological Science*). *Chapters 5-6* then turn to the consequences of (mis)information exposure on social media. *Chapter 5* examines how specific ways of using social media (favoriting, retweeting, and following politicians and misinformation sources) predict vaccine hesitancy in unique datasets of survey data linked to Twitter data (Rathje et. al, R&R at *PNAS Nexus*). *Chapter 6* then presents the results of a large-scale social media experiment and finds that changing social media following and unfollowing behavior has a causal impact on affective polarization and well-being. *Chapter 7* discusses the conclusions and limitations of this research and presents future directions for the study of social media and (mis)information belief and sharing.

Chapter 2. Outgroup Animosity Drives Engagement on Social Media

A version of this chapter was published in the *Proceedings of the National Academy of Sciences* and can be cited as follows:

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), e2024292118.

Introduction

According to a recent article in the Wall Street Journal, a Facebook research team warned the company in 2018 that their "algorithms exploit the human brain's attraction to divisiveness." This research was allegedly shut down by Facebook executives, and Facebook declined to implement changes proposed by the research team to make the platform less divisive (Seetharaman, 2020). This article is consistent with concerns that social media might be incentivizing the spread of polarizing content. For instance, Twitter CEO Jack Dorsey has expressed concern about the popularity of "dunking" (i.e., mocking or denigrating one's enemies) on the platform (Wagner, 2019). These concerns have become particularly relevant as social media rhetoric appears to have incited real-world violence, such as the recent storming of the US Capital (Van Dijcke & Wright, 2021). We sought to investigate whether outgroup animosity was associated with increased virality on two of the largest social media platforms: Facebook and Twitter.

A growing body research has examined the potential role of social media in exacerbating political polarization (Persily & Tucker, 2020; Sunstein, 2018). A large portion of this work has centered on the position that social media sorts us into "echo chambers" or "filter bubbles" that selectively expose people to content that aligns with their pre-existing beliefs (Barberá, 2014; Cinelli et al., 2021; Del Vicario et al., 2016; Madsen et al., 2018; Pariser, 2011; Wojcieszak et al., 2021). However, some recent scholarship questions whether the "echo chamber" narrative has been exaggerated (Bakshy et al., 2015; Eady et al., 2019). Some experiments suggest that social media can indeed increase polarization. For example, temporarily deactivating Facebook can reduce polarization on policy issues (Allcott et al., 2020). However, other work suggests that polarization has grown the most among older demographic groups, who are the least likely to use social media (Boxell et al., 2017) albeit the most likely to vote. As such, there is an open debate about the role of social media in political polarization and intergroup conflict.

Other research has examined the features of social media posts that predict "virality" online. Much of the literature focuses on the role of emotion in social media sharing. High-arousal emotions, whether they are positive (e.g., awe) or negative (e.g., anger or outrage), contribute to the sharing of content online (Berger, 2011; Berger & Milkman, 2012; Crockett, 2017; Fan et al., 2020; Van Der Linden, 2017). Tweets expressing moral and emotional content

are more likely to be retweeted within online political conversations, especially by members of one's political ingroup (Brady et al., 2017, 2021). On Facebook, posts by politicians that express "indignant disagreement" receive more likes and shares (Messing & Weisel, 2017), and negative news tends to spread further on Twitter (Hansen et al., 2011). Moreover, false rumors spread further and faster on Twitter than true ones, especially in the domain of politics, possibly because they are more likely to express emotions such as surprise and fear (Vosoughi et al., 2018).

Yet, to our knowledge, little research has investigated how social identity motives contribute to online virality. Group identities are hyper-salient on social media, especially in the context of online political or moral discussions (Brady, Crockett, et al., 2019). For example, an analysis of Twitter accounts found that people are increasingly categorizing themselves by their political identities in their Twitter bios over time, providing a public signal of their social identity (Rogers & Jones, 2021). Additionally, since sharing behavior is public, it can reflect self-conscious identity-presentation (Kraft et al., 2020b; Van Dijck, 2013). According to Social Identity Theory (Tajfel et al., 1979) and Self-Categorization Theory (Turner et al., 1987), when group identities are highly salient, this can lead individuals to align themselves more with their fellow in-group members, facilitating in-group favoritism and outgroup derogation in order to maintain a positive sense of group distinctiveness (Brewer et al., 1993). Thus, messages that fulfill group-based identity motives may receive more engagement online. As an anecdotal example, executives at the website Buzzfeed, which specializes in creating viral content, reportedly noticed that identity-related content contributed to virality, and began creating articles appealing to specific group identities (Klein, 2020).

People may process information in a manner that is consistent with their partisan identities, prior beliefs, and motivations, a process known as motivated cognition (Kahan, 2015; Kunda, 1990; Taber & Lodge, 2006; Van Bavel & Pereira, 2018). Scholars noted early on that the degree to which individuals identify with their political party "raises a perceptual screen through which the individual tends to see what is favorable to his [or her] partisan orientation" (Campbell et al., 1960). Partisan motivations have been hypothesized to influence online behavior, such as the sharing of true and false news online (Iyengar et al., 2018; Mosleh, Pennycook, et al., 2021; Shin & Thorson, 2017). Accordingly, we suggest that just as people engage in motivated cognition—processing information in a way that supports their beliefs—people may also engage in *motivated tweeting* (or sharing, liking, or retweeting), selectively

interacting with and attending to content that conforms to their partisan identity motivations. There is already evidence suggesting that people selectively follow (Mosleh, Martel, et al., 2021a) and retweet (Shin & Thorson, 2017; Wojcieszak et al., 2021) in-group members at much higher rates than out-group members.

In polarized political contexts, outgroup animosity may be a more successful strategy for expressing one's partisan identity and generating engaging content than ingroup favoritism. Political polarization has been growing rapidly in the United States over the past few decades. Affective polarization, which reflects dislike of people in the opposing political party as compared to one's own party, has most strikingly increased (Iyengar et al., 2019) and ideological polarization may have increased as well (though this is still a topic of debate) (Lelkes, 2016). This growth in affective polarization is driven primarily by increasing out-party animosity (rather than increasing in-party warmth) – a phenomenon known as "negative partisanship" (Abramowitz & Webster, 2018). According to recently released American National Election Studies (ANES) data, affective polarization grew particularly steeply from 2016 to 2020, reaching its highest point in 40 years. Out-party animosity, more-so than in-party warmth, has also become a more powerful predictor of important behaviors, such as voting behavior (Finkel et al., 2020) and the sharing of political fake news (Osmundsen et al., 2021). When out-party animosity is strong, partisans are motivated to distinguish themselves from the out-party (by, for instance, holding opinions that are distinct from the out-party) (Druckman et al., 2013). While some research suggests that out-group cues might be more powerful than in-group cues (Nicholson, 2012), there is still debate about the extent to which partisan belief and behavior is driven by in-party versus out-party cues (Fowler & Harris, 2020) or in-group favoritism versus out-group derogation (Lelkes & Westwood, 2017). A limitation of prior research is that most of it is based on self-report surveys, and so it remains unknown how expressions of in-group favoritism or out-group animosity play out in a social media context – or whether one might be a more powerful contributor to virality than the other.

We investigated the role that political ingroup and outgroup language, as well as discrete emotions, play in predicting online engagement in a large sample of posts from news media accounts and US congressional members (n = 2,730,215). We sought to examine this on both Facebook and Twitter since they are two of the world's largest and most influential social media companies and constitute around three billion users out of nearly four billion total social media

users worldwide (Clement, 2020). Specifically, we were interested in *(a)* how political ingroup and outgroup language compared to other established predictors of social media diffusion, *(b)* whether ingroup or outgroup language was a better predictor of shares and retweets, and *(c)* whether outgroup terms were associated with negative emotions (as measured by the six Facebook "reactions"), and whether ingroup terms were associated with positive emotions, reflecting patterns of out-party derogation and ingroup favoritism. Finally*, (d)* we wanted to see if these findings applied to both news sources and political leaders, who often have an outsized influence on social discourse as well as policy change.

Results

To analyze these questions, we examined large datasets of tweets and Facebook posts from liberal media sources and conservative media sources (as defined by AllSides.com, see ***Methods*** and **Fig. S3**), as well as liberal (i.e., Democrat) and conservative (i.e., Republican) members of congress. Specifically, we counted how many words in each tweet or Facebook post referred to a liberal, b) conservative, or included c) negative emotion, d) positive emotion, or e) moral-emotional language. To measure reference to a liberal or conservative, we use a list of the Top 100 most famous Democrat and Republican politicians as defined by YouGov, a list of all the Democrat and Republican congressional members, and a list of liberal and conservative identity terms (e.g., "left-wing," "conservative," "far-right"), which have been used in prior research (Osmundsen et al., 2021; Rogers & Jones, 2021). We also used previously validated dictionaries of negative affect, positive affect, and moral-emotional language (Brady et al., 2017; Tausczik & Pennebaker, 2010). Adapting prior methods used in similar studies (Brady, Wills, et al., 2019), we fit regression models to examine how language about the outgroup, language about the ingroup, as well as language expressing various emotions (positive affect, negative affect, and moral-emotional language) predicted retweet rates, controlling for various factors known to be correlated with retweet or sharing rate, such as whether a tweet is a retweet (for the Twitter datasets only), whether a message contained a URL or media, and how many followers or likes

the account had. More details are in the methods section. Data and code are available on the OSF at: https://osf.io/py9u4/.

Study 1: Major Media Outlets

In Study 1, we looked at liberal (e.g., New York Times, MSNBC) and conservative (e.g., Fox News, Breitbart) media accounts from Facebook (n = 599,999 posts) and Twitter (n = 227,229 posts). First, we looked at the effect of emotional language on diffusion. Controlling for all other factors, each additional negative affect word was associated with a 5-8% increase in shares and retweets, except in the conservative media Facebook dataset, where it decreased shares by around 2% (exp(b) = 0.98, 95% CI = [0.98, 0.99], $p < 0.001$). Positive affect language was consistently associated with a *decrease* in shares and retweet rates by about 2%-11% across datasets. This largely replicates prior work on the negativity bias in news headlines (Hansen et al., 2011). Additionally, moral-emotional words consistently increased shares and retweets in all datasets by 10%-17%, replicating prior work on the moral contagion effect with similar effect sizes, and extending on this work by showing novel evidence that moral contagion operates on multiple social media platforms, including Facebook (Brady et al., 2017).

To test our primary questions, we looked at how political *ingroup* language predicted diffusion. In the liberal news media accounts on Twitter, political ingroup words were associated with increased retweet rate (exp(b) = 1.10, 95% CI = [1.09, 1.12]). On Facebook, however, there was no equivalent effect of political ingroup language (exp(b) = 1.00, 95% CI = [0.99, 1.00]). In the conservative media Twitter accounts, political ingroup (conservative) words increased retweet rate (exp(b) = 1.23, 95% CI = [1.20, 1.26], $p < 0.001$), and this effect was similar on Facebook (exp(b) = 1.37, 95% CI = [1.35, 1.38]), $p < 0.001$). In sum, political ingroup words led to an estimated 0-37% increase in diffusion per word across all four new media datasets.

We then looked at the effects of political *outgroup* language. In the liberal media Twitter accounts, outgroup language was a strong predictor of retweets (exp(b) = 1.46, 95% CI = [1.44, 1.48]). This effect was similar on Facebook, with outgroup language leading to increased shares (exp(b) = 1.57, 95% CI = [1.55, 1.58]). In the conservative media Twitter accounts, outgroup language increased retweet rate (exp(b) = 1.29, 95% CI = [1.26, 1.31], $p < 0.001$), and this effect was similar on Facebook (exp(b) = 1.35, 95% CI = [1.34, 1.36], $p < 0.001$). Thus, across datasets, outgroup language led to a 35-57% increase in diffusion per additional outgroup word.

Descriptively, the effect sizes of political outgroup language are generally larger than those of ingroup language, and considerably larger than those of any of the emotional dictionaries. The full regression models are reported in *SI Appendix Table S1*, and are plotted visually in **Fig. 1**. The results were similar when the control variables were removed (*SI Appendix Table S3*) and when the models were re-run with cluster-robust standard errors with each media account representing a different cluster (*SI Appendix Table S4*). To further probe the importance of each predictor in the model, we calculated a relative importance analysis (*See SI Appendix Table S4*). In each of the models, outgroup words had the highest "lmg" values (an estimate of the R^2 contributed by each predictor) of all five of the key predictors. Thus, political outgroup language appears to be the most powerful predictor of engagement of all factors measured.

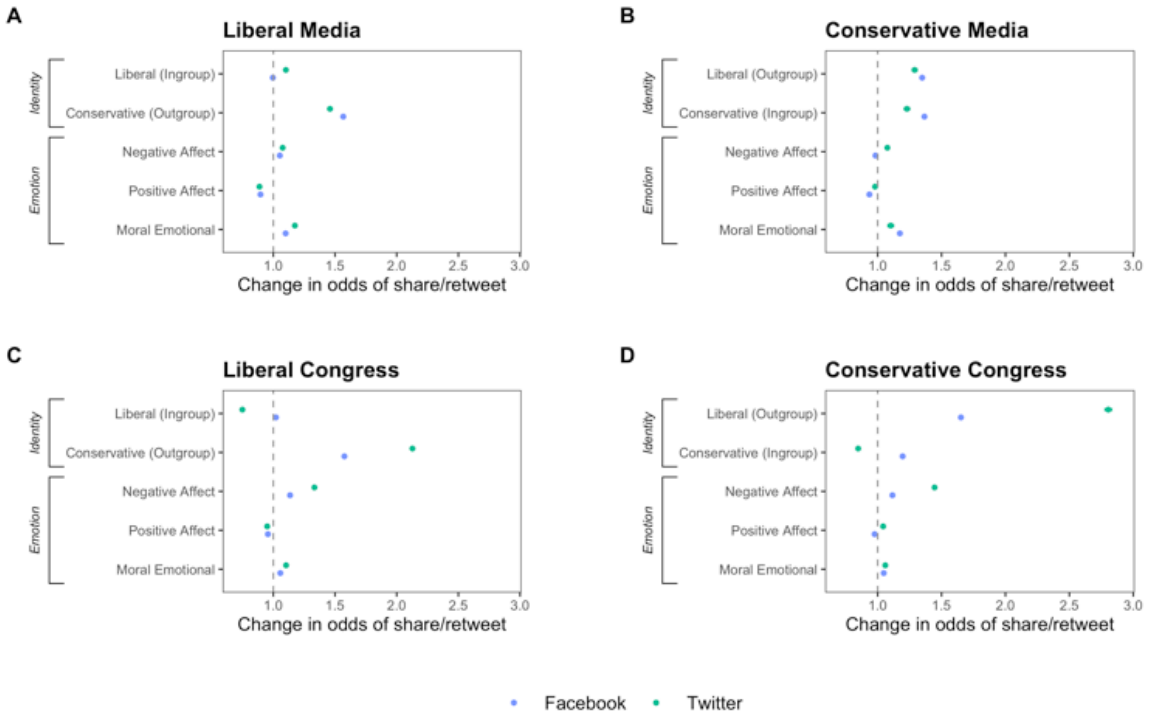


Figure 1. Political out-group words were the strongest predictors of shares and retweets from both liberal and conservative news media accounts (A and B, respectively) and liberal and conservative congress member accounts (C and D, respectively) on Facebook and Twitter. By comparison, political in-group words, as well as measures of discrete emotions, such as positive emotion words, negative emotion words, and moral-emotional words, were relatively weak predictors of shares and retweets. Error bars represent 95% CI (though error bars are small)

We next assessed the valence generated by posts with political outgroup language. We expected posts about the outgroup to evoke negative emotions such as anger or outrage, and posts about the ingroup to evoke positive emotions. Examples of some of the most popular tweets and Facebook posts containing outgroup terms are in *Table 1*. Descriptively, these posts appeared to be very negative. To assess the valence of these posts more systematically, we examined how outgroup language predicted each of the six "reactions" (like, love, haha, sad, wow, and angry) available on Facebook. We assumed that the "angry" reaction was a reasonable proxy for feelings of outgroup animosity, outrage and anger, and the "love" reaction was a reasonable proxy for feelings of in-group love. These results are plotted in **Fig 2**, and full regression models are shown in *SI Appendix, Tables S6-S7*.

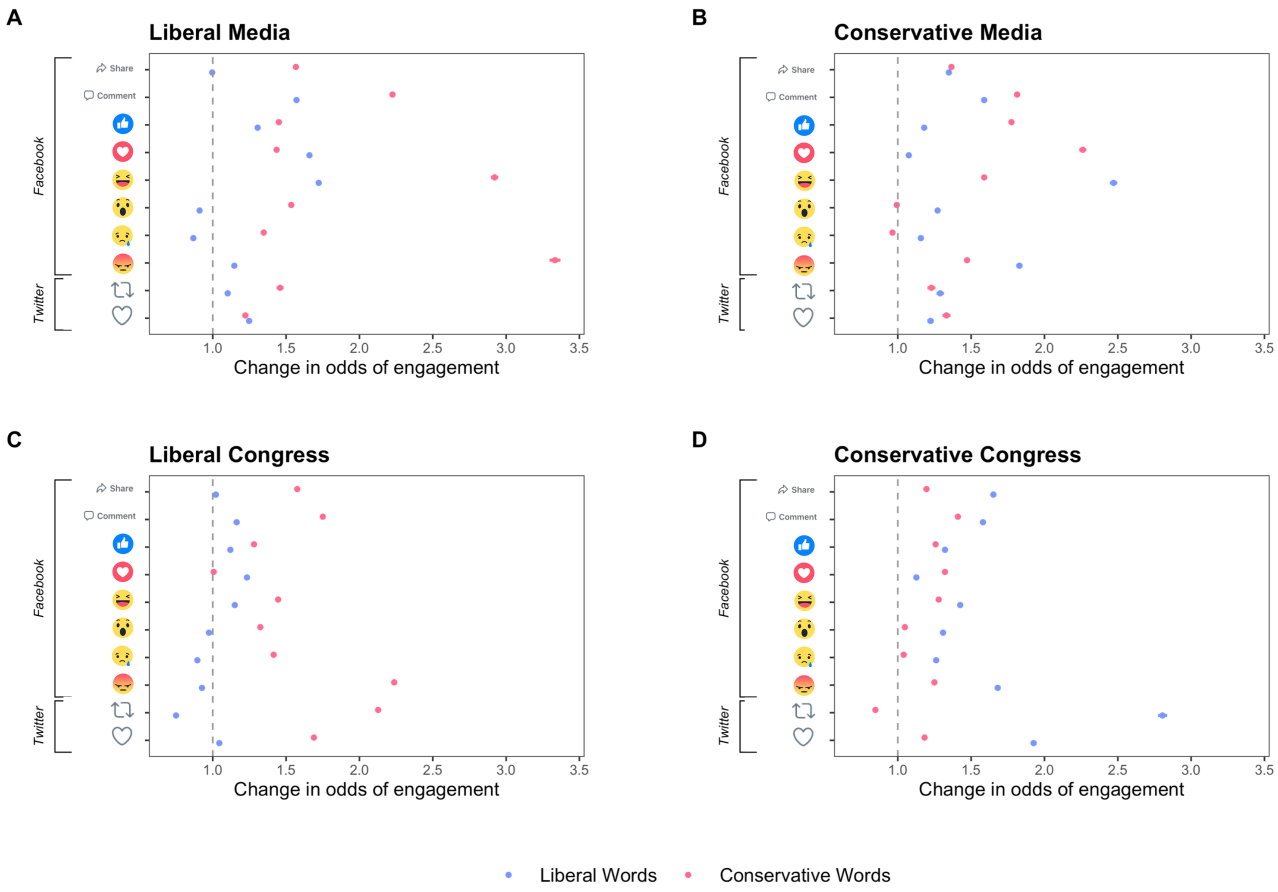


Figure 2. In-group and out-group words predicted different types of engagement in both the liberal and conservative newsmedia accounts (A and B, respectively) and congress accounts (C and D, respectively). Political out-group words were strong predictors of shares, comments, "haha," and "angry" reactions, whereas in-group words were strong predictors of "love" reactions. Reactions are shown as they are shown on the Facebook and Twitter platforms (from Top to Bottom: share, comment, like, heart, haha, wow, sad, angry, retweet, and favorite). Error bars represent 95% CI (though error bars are small).

Table 1: Example Tweets and Facebook posts.

Dataset	Liberal	Shares/Retweets	Conservative	Shares/Retweets
Media (Facebook)	BREAKING: PRESIDENT **TRUMP** HAS BEEN IMPEACHED.	82,886	Reported **Antifa** Protester tries a sucker punch and it doesn't go so well…	71,482
Media (Twitter)	Vice President **Mike Pence** blatantly lied to reporters about the trajectory of COVID-19 cases in Oklahoma, where President **Trump** is scheduled to hold a large campaign rally on Saturday.	8,793	Every American needs to see Joe **Biden**'s latest brain freeze.	15,354
Congress (Facebook)	Donald **Trump** has lied more than 3,000 times since taking office but **Republicans** refuse to say **Trump** is a liar. What's going on?	29,737	**Democrats** just passed a bill that would make it harder for American innovators to develop a COVID-19 vaccine. Here's what you need to know:	10,354
Congress (Twitter)	**Republicans** are saying they are being barred from the "secret" hearings. But here's a list of every **Republican** who is allowed into the hearings.	41,541	RT to tell **Chuck Schumer** and **Nancy Pelosi** to STOP blocking critical funding for small businesses. The Paycheck Protection Program is about to run out of money—millions of jobs are hanging in the balance. Congress MUST ACT!	37,872

Note. Examples of some of the most popular posts from each dataset, along with their shares and retweets at the time of data collection. Political outgroup language is bolded (and color-coated so that red equals conservative/Republican and blue equals liberal/Democrat).

As expected, political outgroup language was a very strong predictor of "angry" reactions for both liberals (exp(b) = 3.33, 95% CI = [3.30, 3.37], $p < 0.001$) and for conservatives (exp(b) = 1.83, 95% CI = [1.81, 1.85], $p < 0.001$). Outgroup words were also strong predictors of "haha" reactions for both groups ($\exp(b)_{liberals} = 2.92$, 95% CI = [2.90, 2.95], $p < 0.001$; $\exp(b)_{conservatives} = 2.47$, 95% CI = [2.45, 2.50], $p < 0.001$). Thus, posts about the outgroup may generate engagement by inspiring negative emotions such as anger, outrage, or mockery. Strikingly, descriptive statistics (S9) show that, on average, the angry reaction was the most popular of the six reactions for both liberals and conservatives in the news media accounts, consistent with the perspective that outrage is popular on online social networks (Crockett, 2017). On the flipside, ingroup words, as expected, strongly predicted love reactions for both liberals (exp(b) = 1.66, 95% CI = [1.64, 1.68], $p < 0.001$) and conservatives (exp(b) = 2.26, 95% CI = [2.24, 2.26], $p < 0.001$).

Study 2: Congress Members

In Study 2, we replicated the above results in a different context: tweets (n = 1,078,562) and Facebook posts (n = 825,424) by Democratic and Republican Congressional members. Given growing levels of polarization in Congress (Bisgaard & Slothuus, 2018a) and because political elites are often agenda setters who frame political debates and influence public opinion (Bisgaard & Slothuus, 2018a; Druckman et al., 2013), we thought this was an important additional context to investigate the virality of social media posts.

First, we looked at the effect of emotional language on virality. Negative affect language consistently increased retweet rate and shares across all datasets by 12-45% per negative affect word, with the effect size being largest in the conservative Twitter dataset (exp(b) = 1.45, 95% CI = [1.44, 1.45], $p < 0.001$). Similarly, moral-emotional language had a consistent positive effect across all datasets, increasing retweets and shares by roughly 5-10%. Positive affect language slightly decreased shares by roughly 2-5%, except in the conservative Twitter accounts (exp(b) = 1.04, 95% CI = [1.04, 1.05], $p < 0.001$). Replicating the results from Study 1, negative language and moral-emotional language were once again positively associated with diffusion, whereas positive affect language was negatively associated with it.

Next, we again looked at the effects of political ingroup language. In the liberal congressional accounts, political ingroup language decreased retweet rate on Twitter (exp(b) = 0.75, 95% CI = [0.75, 0.75], $p < 0.001$), and only slightly increased shares on Facebook (exp(b) = 1.02, 95% CI = [1.01, 1.03], $p < 0.001$). Similarly, in the conservative dataset, political ingroup language decreased retweet rate on Twitter (exp(b) = 0.85, 95% = [0.84, 0.85], $p < 0.001$), and slightly increased shares on Facebook (exp(b) = 1.20, 95% = [1.19, 1.20], $p < 0.001$). In sum, political ingroup language led to a mixed pattern of results across all four congressional datasets.

Replicating our findings with media accounts, political outgroup language was a very large predictor of retweets in the liberal congressional Twitter accounts (exp(b) = 2.13), 95% CI = [2.11, 2.15], $p < 0.001$), and of shares in the liberal congressional Facebook accounts (exp(b) = 1.58, 95% CI = [1.57, 1.59], $p < 0.001$). The same was true in the conservative congressional Twitter accounts (exp(b) = 2.80, 95% CI = [2.77, 2.84], $p < 0.001$) and Facebook accounts (exp(b) = 1.65, 95% CI = [1.64, 1.67], $p < 0.001$). This effect translates into an estimated 65%-180% increase in the odds of being shared per outgroup word across datasets. Descriptively, these effect sizes are very large, and larger than those found in the news media accounts. This might be due to the fact that members of congress are explicitly identified with a political party and have a large partisan following.

To further explore the importance of political outgroup language, we conducted another relative importance analysis (see *SI Appendix Table S13*). In each model, outgroup language had the highest estimated R^2 ("lmg") value compared to the other key predictors (political ingroup, negative, positive, and moral-emotional language). In other words, it was once again the most important predictor in each model.

When examining different types of engagement (e.g., the six Facebook reactions, see S17-S18 for more detail), we once again saw similar patterns to media outlets. Posts about the outgroup strongly predicted negative reactions, such as "angry" reactions, for both liberals (exp(b) = 2.24, 95% CI = [2.22, 2.25], $p < 0.001$) and for conservatives (exp(b) = 1.68, 95% CI = [1.67, 1.69], $p < 0.001$). On the other hand, posts about the political ingroup predicted "love" reactions for both liberals (exp(b) = 1.23, 95% CI = [1.22, 1.24], $p < 0.001$) and for conservatives (exp(b) = 1.32, 95% CI = [1.31, 1.33], $p < 0.001$). Descriptive statistics (*See SI Appendix Tables S19-S20*) again found that the angry reaction was generally the most popular reaction, although the "love" reaction slightly surpassed the angry reaction in popularity in the

conservative dataset.

Internal Meta-Analysis

To estimate the average effect sizes across all eight datasets, we conducted a series of internal meta-analyses (**Fig. 3 (Panel A)** and *SI Appendix Table S22*). We computed a random-effects meta-analysis (because we expected this effect to vary across contexts), and used the Dersimonian-Laird (DL) estimator. Across datasets, each political outgroup word increased the odds of a retweet or share by about 67% (estimated exp(b) = 1.67, 95% CI = [1.43, 1.69], $p <$ 0.001)[1]. Political ingroup language, on the other hand, did not have a statistically significant effect on shares and retweets (exp(b) = 1.05, 95% CI = [0.90, 1.22], p = 0.563). Negative affect language increased diffusion by about 14% per word (exp(b) = 1.14, 95% CI = [1.05, 1.24], moral-emotional language increased diffusion by 10% per word (exp(b) = 1.10, 95% CI = [1.07, 1.13], $p < 0.001$), and positive affect language decreased diffusion by about 5% per word (exp(b) = 0.95, 95% CI = [0.93, 0.98], $p < 0.001$).

[1] Using different estimators for the meta-analysis did not yield different results. For instance, using the Empirical Bayes estimator led to the following estimated effect size: exp(b) = 1.67, 95% CI = [1.40, 2.00], p < 0.001, which is the same effect size with a slightly larger confidence interval.

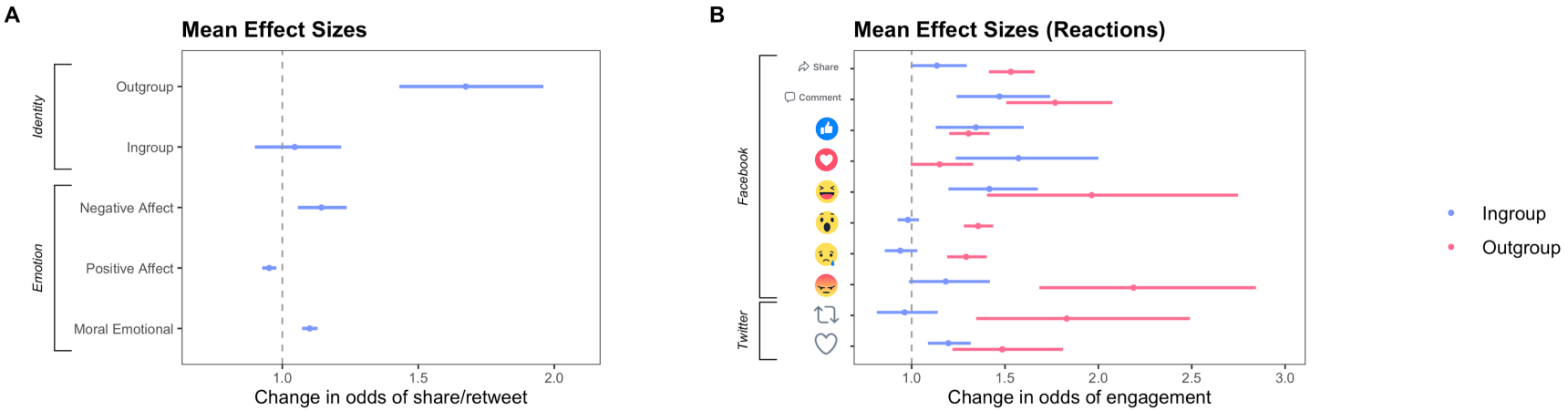


Figure 3. In-group and out-group words predicted different types of engagement in both the liberal and conservative news media accounts (A and B, respectively) and congress accounts (C and D, respectively). Political out-group words were strong predictors of shares, comments, "haha," and "angry" reactions, whereas in-group words were strong predictors of "love" reactions. Reactions are shown as they are shown on the Facebook and Twitter platforms (from Top to Bottom: share, comment, like, heart, haha, wow, sad, angry, retweet, and favorite). Error bars represent 95% CI (though error bars are small).

To put these effect sizes in context, the average percent increase in shares of political outgroup language was about *4.8 times* as large as that of negative affect language, and about *6.7 times* as large as that of moral-emotional language. While one might expect words that have clear political content (e.g., names of specific politicians) to be more predictive of social media shares than words that refer to general emotions (e.g., adjectives such as "bad"), this large effect size is notable, because negative emotion and moral-emotional language are well-established predictors of diffusion on social networks (Brady et al., 2017; Brady, Wills, et al., 2019; Fan et al., 2020), and have been the main focus of prior work looking at social media diffusion. Here, we show that the use of out-group terms (but not ingroup terms) is a much stronger predictor of diffusion than various measurements of moral or emotional language.

We then analyzed a set of moderator variables in the meta-analysis. The effect of political outgroup language on diffusion was not moderated by political orientation (exp(b) = 0.98, 95% CI = [0.75, 1.30], p = 0.187), nor by social media platform (exp(b) = 1.20, 95% CI = [0.91, 1.58], p = 0.910). However, it was moderated by whether the tweets came from members of congress as opposed to news media accounts (exp(b) = 1.39, 95% CI = [0.91, 1.58], p < 0.018).[2] Thus, we did not detect any ideological asymmetries in the internal meta-analysis, nor did we find any key differences between social media platforms. However, the effect was clearly stronger among politicians than in the news media accounts, possibly because of the more explicitly partisan rhetoric of political leaders. The estimated effect size for the media datasets only was exp(b) = 1.41, 95% CI = [1.30, 1.54], p < 0.001 (a 41% increase) and the estimated effect size for the congress datasets was exp(b) = 1.99, 95% CI = [1.53, 2.58], p < 0.001 (a 99% increase).

Because this analysis focused on the effect of each additional outgroup word, we also conducted additional analyses where we examined how much a post with at least one outgroup term diffused compared to a post with at least one ingroup term, controlling for all the same relevant variables. Posts with both ingroup and outgroup terms, as well as posts with no ingroup and outgroup terms, were excluded from this analysis. Thus, we could directly compare how much posts about only the outgroup (coded as 1) diffused compared to posts about only the

[2] For the moderation analysis, liberal was coded as 1, conservative was coded as 0, Twitter was coded as 1, Facebook was coded as 0, congress was coded as 1, and media was coded as 0.

ingroup (coded as 0), following the methods of past research that has looked at how the diffusion of false news compares to true news (Vosoughi et al., 2018). When meta-analyzed across all 8 datasets, posts with at least one outgroup word were more than twice as likely to be shared than posts with at least one ingroup word (estimated exp(b) = 2.32, 95% CI = [1.57, 2.47], $p < 0.001$) (a 132% increase).

We also conducted internal meta-analyses using the same methods to report average effect sizes for each of the Facebook reactions (**Fig. 3 (Panel B)**). While all reactions are shown in **Fig. 3**, we focused specifically on "anger" and "love" reactions, as these most clearly indicate outgroup animosity or in-group love. Outgroup language was a very large predictor of the "angry" reaction across datasets (exp(b) = 2.19, 95% CI = [1.68, 2.84], $p < 0.001$), but ingroup language was only marginally associated with the "angry" reaction (exp(b) = 1.18, 95% CI = [0.99, 0.42], $p = 0.07$). Furthermore, ingroup language was strongly associated with the "love" reaction across datasets (exp(b) = 1.57, 95% CI = [1.24, 2.00], $p < 0.001$), whereas outgroup language was not associated with the "love" reaction (exp(b) = 1.15, 95% CI = [1.00, 1.33], $p = 0.059$). Thus, outgroup language appears to reflect outgroup derogation, whereas ingroup language reflects ingroup favoritism. Furthermore, the effect size of outgroup language predicting "angry" reactions was more than twice as big as the effect size of ingroup language predicting "love" reactions, once again showing an outgroup bias.

We wanted to test whether the effect of political outgroup language was not driven by any specific words in particular (e.g., "Trump"). To examine this, we repeated our analysis with each of the three sub-dictionaries that made up the political outgroup language dictionary: 1) the dictionaries of the Democratic and Republican "identity" terms (i.e., "Democrat," "right-wing", "leftist"), 2) lists of the top 100 Democratic and Republican politicians as ranked by YouGov, along with their Twitter handles (or Facebook page names on Facebook) and 3) lists of all liberal and conservative congressional members, along with their Twitter handles (or Facebook page name on Facebook). We then meta-analyzed these results across all 8 datasets.

Looking only at the "identity" terms, political outgroup language led to an estimated 91% increase in the odds of being shared per word (exp(b) = 1.91, 95% CI = [1.38, 2.64], $p < 0.001$). An additional word from the top 100 most famous politicians dictionary led to an estimated 82% increase in the odds of being shared (exp(b) = 1.82, 95% CI = [1.53, 2.17], $p < 0.001$). Lastly, each additional word from the list of outgroup congressional members led to a 43% increase in

shares (exp(b) = 1.43, 95% CI = [1.23, 1.66], $p < 0.001$). In other words, whether referring to a general identity term, a famous politician, or a member of congress, outgroup language is a very strong predictor of diffusion. This helps validate that this phenomenon is not dependent on any one specific dictionary and is robust across specifications. The slightly smaller effect size of outgroup congressional words may be due to the fact that many congressional members are not as widely known as the most famous politicians. Because of this, including the full list of congressional members in the main analysis may have led to a conservative estimation of the true effect size.

Discussion

Across 2,730,215 total observations from Facebook and Twitter, we find that posts about the political outgroup are consistently more likely to be shared than those about the political ingroup. The effect of outgroup language was the most important predictor of sharing behavior in posts from both news media accounts and politicians — considerably stronger than the effects of political ingroup language or various discrete emotions, which have previously been the main focus when assessing what makes content go "viral" online (Berger, 2011; Brady et al., 2017). To contextualize this large effect, the percent increase in estimated shares associated with outgroup language was 4.8 times as big as that of negative affect language, and 6.7 times as big as that of moral-emotional language – previously established predictors of message diffusion online.

This outgroup effect was also robust against different ways of operationalizing the outgroup, suggesting that the pattern of results is not primarily driven by the mention of specific terms or particularly divisive politicians, such as Donald Trump. The effect was also not moderated by political orientation or by social media platform. However, the effect of outgroup language was considerably stronger among politicians than in the news media accounts, perhaps because of the more explicitly partisan rhetoric among political elites (Green et al., 2020; Hetherington, 2001) and their followers. Additionally, given prior concerns that much of social media research focuses predominantly on Twitter due to the relatively easy accessibility of Twitter data (Persily & Tucker, 2020) it is notable that similar patterns were found on both Twitter and Facebook.

Political ingroup and outgroup language also generated distinctly different forms of

engagement, reflecting clear patterns of ingroup favoritism and outgroup derogation. For instance, outgroup language strongly predicted "angry" reactions (as well as "haha" reactions, comments, and shares), and ingroup language strongly predicted love reactions. Though, notably, outgroup language was about twice as strong a predictor of "angry" reactions as ingroup words were of "love" reactions. Thus, posts about the outgroup may be so successful because they appeal to emotions such as anger, outrage, and mockery. Indeed, the "angry" reaction was the most popular reaction on Facebook in seven of the eight datasets analyzed.

This research is consistent with prior research showing that expressions of moral outrage (which involve emotions such as anger or disgust) are particularly likely to go viral (Brady, Crockett, et al., 2019, 2019), but it expands on that work by illuminating the role of outgroup animosity in eliciting outrage. The current research reveals the key role that outgroup identity language played in predicting sharing behavior above and beyond emotional words alone. In fact, in the supplementary analysis (*SI Appendix, Tables S6-S7, S17-18*), we found that emotional language was weakly associated with the various Facebook reactions, suggesting that out-of-context emotional language used in a post may not be the most precise way to measure actual emotions evoked by a social media post.

These results demonstrate how the predictions of Social Identity Theory play out in a modern social media context (Tajfel et al., 1971). As expected, posts that appeal to identity-based motives tend to receive more engagement in online social networks. Additionally, there is also a strong asymmetry such that outgroup negativity is stronger than ingroup positivity, reflecting the current state of negative partisanship in the United States (Abramowitz & Webster, 2018; Finkel et al., 2020). These results also expand on prior work on the motives behind social media sharing. Social media sharing often reflects a desire to maintain a positive self-presentation (Berger & Milkman, 2012; Kraft et al., 2020b). This can lead to different outcomes depending on the context, social norms, and design features of one's online network (Van Bavel, Harris, et al., 2021), since strategies to maintain a positive self-image may differ by context (Brady, Crockett, et al., 2019; Van Bavel, Harris, et al., 2021). While some studies find a negativity bias in online sharing (Fan et al., 2020; Soroka et al., 2019), there are other contexts where positive content is shared more often. For instance, the New York Times most emailed list tends to have more positive content (Berger & Milkman, 2012; Kraft et al., 2020b) as do viral articles about science (Milkman & Berger, 2014). However, in online contexts where political

identity is highly salient, and where political conflict is driven by negative partisanship, the best way to maintain an image of a good ingroup member and to distinguish oneself from the outgroup may be to share expressions of out-party animosity (Druckman et al., 2020). Additionally, other work has found that words that predict virality (such as moral or emotional words) are prioritized in early visual attention (Brady et al., 2020). Political identity-relevant content may also be similarly attention-grabbing, especially when political conflicts have become excessively negative and moralized (Finkel et al., 2020).

While much of the literature on social media and political polarization has focused on the formation of echo chambers, the finding that social media amplifies out-group animosity might be more concerning than the formation of echo chambers alone. Even if people are exposed to more cross-partisan content than expected (Bakshy et al., 2015; Barberá et al., 2015) our findings suggest that opposing views on social media may be excessively negative about one's *own* side. This may help explain why exposure to opposing views on Twitter can actually increase political polarization (Bakshy et al., 2015). Thus, the severity of online echo chambers appears to be a less important issue than the kind of content that tends to surface at the top of one's feed, since exposure to divisive in-party or out-party voices is unlikely to be productive. While future experimental work is needed to examine the consequences of these trends, the amplification of divisive posts on social media – from both in-party and out-party sources – may be playing a role in rising political polarization.

This big data approach comes with many benefits, such as allowing us to understand how political identity contributes to engagement with online content and thus has high ecological validity. However, this approach also comes with several limitations. While these results may be consistent with theoretical predictions, they are correlational, and further experimental work should be conducted to determine causation and help clarify why content about outgroup identities is engaging in online political conversations. Additionally, while we found this effect amongst contexts on two of the largest social media platforms, we were unable to follow up certain important questions, such as who is producing this engagement, due to data access limitations.

It is important to note that the data we observed are likely reflective of a specific time-frame, namely the years leading up to the 2020 election. Since the language of political elites can change depending on which party is in power (Wang & Inbar, 2020), and the United States is at

historically high levels of polarization (Lelkes, 2016), it is unclear whether these results would generalize to different time periods or nations. It is also unclear how algorithmic choices on the part of Facebook or Twitter might contribute to the amplification of out-group animosity, since social media companies are not transparent about how their algorithms work. Despite these drawbacks, this study reveals a consequential trend playing out within two of the most influential social networks, inspiring many questions for future research.

This research is also important on a practical level. Social media is encroaching on more aspects of our lives, becoming one of the main ways in which people consume news and interact with politicians (Boczkowski et al., 2018). Since the social media ecosystem operates as an attention economy (Williams, 2018) whereby users, politicians, and brands fight for attention and engagement, understanding what drives virality is crucial. Virality can contribute to the success of a social movement, business, or political campaign (Van Der Linden, 2017) so people have strong incentives to generate viral, engaging content. Virality is also essential for social media companies, as the business model of social media is grounded in generating engagement with the platform, which leads to advertising revenue. When the chief goal is virality, this may create negative externalities in the form of polarizing, hyper-partisan, false, or hostile content. This kind of content may be good at generating superficial engagement but ultimately harm individuals, political parties, or society in the long-term.

The design structure of social media platforms may be creating perverse incentives for polarizing content when users do not truly want this. For instance, people report that they do not want political leaders to express partisan animus (M. Costa, 2020), but our results suggest this content receives the most engagement. As further illustration of these perverse incentives, the New York Times reported on internal research from Facebook finding that posts that users rated as "bad for the world" received more engagement. When Facebook tested a feature to down-rank posts that were rated as "bad for the world," engagement decreased, and Facebook ultimately chose not to approve the feature (Roose et al., 2020). Thus, social media companies may be reluctant to implement features that could reduce polarization due to their strong financial incentives to maintain user engagement.

Conclusion

Understanding the factors that make social media posts go "viral" online can help create better social media environments. While social media platforms are not fully transparent about

how their algorithmic ranking system works, Facebook announced in a post titled "Bringing People Closer Together" that it was changing its algorithm ranking system to value "deeper" forms of engagement, such as reactions and comments (Mosseri, 2018). Ironically, posts about the political outgroup were particularly effective at generating comments and reactions (particularly the "angry" reaction, the most popular reaction across our studies). In other words, these algorithmic changes made under the guise of bringing people closer together may have helped prioritize posts including outgroup animosity. In addition to informing algorithmic changes (Rahwan et al., 2019), this research might inform other design changes, "nudges" (Lorenz-Spreen et al., 2020), or policy changes that can be implemented to improve social media conversations, as well as future research on the role of social identity in online engagement. Amid widespread discussion that social media may be contributing to discord and polarization, our work reveals how outgroup animosity predicts virality in two of the largest social networks.

Materials and Methods

All methods were approved by the University of Cambridge Research Ethics Committee. For Study 1, we collected tweets from several news media accounts across the political spectrum using the R package "rtweet" and the Twitter API. After collecting up to 3200 of the most recent tweets from each account (the total amount permitted by the Twitter API), we were left with a total of 227,229 total tweets for analysis. These news media accounts were chosen because they were classified by the All Sides Media Bias Chart (*SI Appendix Figure S3*), which aims to identify the political bias of various news sources. This allowed us to split the dataset into tweets from liberal (n = 143,702) and conservative (n = 83,527) media sources. In Study 2, we analyzed tweets from all members of congress (up to 3200 tweets per member). We split the dataset into tweets from Democratic (n = 747,675) and Republican (n = 611,292) US congressional members.

Facebook data was retrieved through a partnership with Crowdtangle, a tool owned by Facebook that aggregates data from public pages, and Social Science One, an organization that forms partnerships with industry and social science researchers. Using the Crowdtangle platform, we created lists of the same liberal and conservative media accounts from Allsides.com, and downloaded the 300,000 most recent posts from these lists of media accounts. We also used official lists assembled by the Crowdtangle staff of the current Democrat and Republican US

House of Representative and Senate Members. After combining the downloaded lists of the US House of Representatives and US Senate, we were left with 366,842 liberal congress Facebook posts and 458,582 conservative congress Facebook posts. All data were retrieved during 2020, and the majority of the observations for the media accounts range from 2018-2020, and range from 2016-2020 in the congressional accounts. More information about the exact timeline that the tweets and Facebook posts reflect is in *SI Appendix Figure S2*. Data and code are available on the OSF at: https://osf.io/py9u4/, though text of individual social media posts could not be shared for privacy reasons. We determined our sample size and exclusions in advance of analyzing the data, and, where possible, kept the analysis methods as close as possible for Twitter and Facebook.

We used the R package "quanteda" to analyze Twitter and Facebook text (Benoit et al., 2018). During text pre-processing, we removed punctuation, URLs, and numbers. To classify whether a specific post was referring to a liberal or a conservative, we adapted previously-used dictionaries that referred to words associated with liberals or conservatives (Osmundson et al., 2021; Rogers & Jones, 2021). Specifically, these dictionaries included 1) a list of the top 100 most famous Democratic and Republican politicians according to YouGov, along with their Twitter handles (or Facebook page names for the Facebook datasets) (e.g., "Trump," "Pete Buttigieg", "@realDonaldTrump"), 2) a list of the current Democratic and Republican (but not independent) US Congressional members (532 total) along with their Twitter and Facebook names (e.g., "Amy Klobuchar," "Tom Cotton," and 3) a list of about 10 terms associated with Democratic (e.g., "liberal," "democrat," "leftist") or Republican identity (e.g., "conservative," "republican," "ring-wing"). We then assigned each tweet a count for words that matched our Republican and Democrat dictionaries (for instance, if a tweet mentioned two words in the "Republican" dictionary, it would receive a score of "2" in that category). We also used previously validated dictionaries that counted the number of positive and negative affect words per post (Soroka et al., 2019) and the number of moral-emotional words per post (Brady et al., 2017). All dictionaries are available on the OSF (https://osf.io/py9u4/), except for the positive and negative affect dictionaries, which are proprietary and must be purchased through the program LIWC (Tausczik & Pennebaker, 2010).

In each dataset, adapting prior methods (Brady, Wills, et al., 2019), we fit OLS regression models to examine how language about the outgroup, language about the ingroup, as well as

language expressing various emotions (positive affect, negative affect, and moral-emotional language) predicted retweet rates. We controlled for whether a post contained a URL, media (i.e., photo or video), the number of followers each account had, and whether a tweet was a retweet. All variables were mean centered using the R package "jtools". Following prior work (Brady, Wills, et al., 2019), we log-transformed the retweet-count and share outcome variables to account for the fact that these variables are typically skewed. We applied the same models to each of the individual Facebook reactions to assess different forms of engagement. Afterward, we conducted several random-effects internal meta-analyses using the R package "meta." Analyses were performed using R version 4.0.1.

Chapter 3. Accuracy and Social Motivations Shape Judgements of (Mis)Information

A version of *Chapter 3* of the thesis has been revised and resubmitted to *Nature Human Behaviour* with co-authors Jon Roozenbeek, Jay Van Bavel, and Sander van der Linden. The following pre-print can be cited:

Rathje, S., Roozenbeek, R., Van Bavel, J. J., & van der Linden, S. (2022). Accuracy and Social Motivations Shape Judgements of (Mis)information. https://doi.org/10.31234/osf.io/hkqyv

Introduction

Misinformation – which can refer to fabricated news stories, false rumors, conspiracy theories, or disinformation campaigns – can have serious negative effects on society and democracy (Lewandowsky et al., 2017; Van Bavel et al., 2021). Numerous studies suggest that misinformation exposure[3] may reduce support for climate change (Biddlestone et al., 2022; Van der Linden et al., 2017b) and the COVID-19 vaccine (Loomba et al., 2021b; Pierri et al., 2022), and that the mere repetition of misinformation can increase belief in it (Dechêne et al., 2010; Pennycook et al., 2018). Anti-vaccination viewpoints are becoming increasingly popular online (Johnson et al., 2020), and there is widespread belief in misinformation and conspiracy theories about election fraud (Pennycook & Rand, 2021a) and COVID-19 (Roozenbeek et al., 2020). There has thus been a growing interest in understanding the psychology of belief in misinformation and how to mitigate its spread (Lewandowsky et al., 2017; Pennycook & Rand, 2021b; C. Robertson et al., 2022; Van Bavel et al., 2021; van der Linden, Roozenbeek, et al., 2021).

There are substantial partisan differences in how people judge information to be true or false. People are much more likely to believe news with politically-congruent content (Aslett et al., 2021; Batailler et al., 2021; Gawronski, 2021; Van Bavel & Pereira, 2018) or news that comes from politically-congruent sources (Traberg & van der Linden, 2022; van der Linden et al., 2020). However, there are multiple possible reasons that can explain why this partisan divide exists. One possible explanation is that people tend to engage in politically-motivated cognition (Kunda, 1990; Taber & Lodge, 2006): although people are often motivated to be accurate, they also have social goals (e.g., group belonging, status, etc.) for holding certain beliefs that can interfere with accuracy goals (Van Bavel & Pereira, 2018). Another potential explanation is that partisans have different pre-existing knowledge, or different prior beliefs, as a result of exposure to different partisan news outlets and social media feeds (Pennycook & Rand, 2021b). It is challenging to differentiate between these explanations unless accuracy or social motivations are

[3] It should be noted that there are some null results regarding the effects of misinformation, such as a small correlational study finding that belief in COVID-19 misinformation was not associated with vaccine hesitancy(Kreps et al., 2021). In contrast, however, a larger-scale correlational study found that belief in COVID-19 misinformation was a robust negative predictor of intentions to engage in preventative health behavior(Pavlović et al., 2022).

experimentally manipulated (Bayes et al., 2020b; Druckman & McGrath, 2019; Tappin et al., 2020b; van der Linden, 2022).

Several studies have also found that US conservatives tend to believe in and share more misinformation than US liberals (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Lawson & Kakkar, 2021; Pereira & Van Bavel, 2018; Roozenbeek et al., 2022a; van der Linden, Panagopoulos, et al., 2021). One interpretation behind this asymmetry is that US conservatives are exposed to more low-quality information and thus have less accurate political knowledge, perhaps due to US conservative politicians and news media sources sharing less accurate information (Evanega et al., 2020; Mosleh & Rand, 2021). Another interpretation again focuses on motivation, suggesting that US conservatives may, in some contexts, have greater motivations to believe ideologically or identity-consistent claims that could interfere with their motivation to be accurate (Baron & Jost, 2019; Jost et al., 2003). Yet, it is difficult to disentangle the causal role of motivation versus prior knowledge without experimentally manipulating motivations.

We examine the causal role of accuracy motives in shaping judgements of true and false political news via the provision of financial incentives for correctly identifying accurate headlines. Prior research about the effect of financial incentives for accuracy has yielded mixed results. For example, previous studies have found that financial incentives to be accurate can reduce partisan bias about politicized issues (Bullock & Lenz, 2019; Prior et al., 2015) and news headlines,(Jakesch et al., 2019) and improve accuracy about scientific information (Panizza et al., 2021). However, another study found that incentives for accuracy can backfire, increasing belief in false news stories (Aslett et al., 2021). Incentives also do not eliminate people's tendency to view familiar statements (Speckmann & Unkelbach, 2022) or positions for which they advocate (Melnikoff & Strohminger, 2020) as more accurate, raising questions as to whether incentives can override the heuristics people use to judge truth (Brashier & Marsh, 2020). These conflicting results motivate the need for a systematic investigation of when and for whom various motivations influence belief.

We also examine whether social identity-based motivations to identify posts that will be liked by one's political in-group interfere with accuracy motivations. On social media, content that appeals to social-identity motivations, such as expressions of out-group derogation, tends to receive higher engagement online (Rathje et al., 2021). False news stories may be good at

fulfilling these identity-based motivations, as false content is often negative about outgroup members (Garrett & Bond, 2021; Osmundsen et al., 2021). The incentive structure of the social media environment draws attention to social motivations (e.g., receiving social approval in the form of likes and shares), which may lead people to give less weight to accuracy motivations online (Brady et al., 2019; Ren et al., 2021).

Finally, we compare the effect of accuracy motivations to the effects of other factors that are regularly invoked to explain the belief and dissemination of misinformation, such as analytic thinking (Pennycook & Rand, 2018) political knowledge (Vegetti & Mancosu, 2020), media literacy skills (A. M. Guess, Lerner, et al., 2020), and affective polarization (Osmundsen et al., 2021). By including these variables in the same study, we are able to develop a more complete account of the factors that drive (mis)information belief and sharing (Van Bavel et al., 2021; van der Linden, Roozenbeek, et al., 2021).

Overview

Across four pre-registered experiments, including a replication with a nationally representative US sample, we test whether (A) incentives to be accurate improve people's ability to discern between true and false news and (B) reduce partisan bias (Experiment 1). Additionally, we test whether (C) increasing partisan identity motivations by paying people to correctly identify posts that appeal to one's in-group (mirroring the incentives of social media) reduces accuracy, even when paired with accuracy incentives (Experiment 2). Further, (D) we examine whether the effects of incentives are attenuated when partisan source cues are removed from posts (Experiment 3). Then, to test the generalizability of these results and help rule out alternate explanations, we test whether (E) increasing accuracy motivations through a non-financial accuracy motivation intervention also improves accuracy. Finally, in an integrative data analysis, we (F) examine whether motivation helps explain the gap in accuracy between conservatives and liberals, and (G) compare the effects of motivation to the effects of other variables known to predict misinformation susceptibility.

Results

Experiment 1: Incentives Improve Accuracy and Reduce Bias

In Experiment 1, we recruited a politically-balanced sample of 462 US adults via the survey platform Prolific Academic (Peer et al., 2021) Participants were shown 16 pre-tested news headlines with an accompanying picture and source (similar to how a news article preview would show up on someone's Facebook feed). In a pre-test, eight headlines (four false and four true) were rated as more accurate by Democrats than Republicans, and eight headlines (four false and four true) were rated as more accurate by Republicans than Democrats (Pennycook et al., 2020). An example of a Democrat-leaning true headline was "Facebook removes Trump ads with symbols once used by Nazis" from *apnews.com*, and an example of a Democrat-leaning false news headline was "White House Chef Quits because Trump Has Only Eaten Fast Food For 6 Months" from *halfwaypost.com*. After seeing each headline, participants were asked "To the best of your knowledge, is the claim in the above headline accurate?" and were then asked "If you were to see the above article on social media, how likely would you be to share it?" See ***Methods*** for more details.

Half of the participants were randomly assigned to the *accuracy incentives* condition. In this condition, participants were told they would receive a small bonus payment of up to one US dollar based on how many correct answers they could provide regarding the accuracy of the articles. The other half of participants were assigned to a *control* condition in which they were asked the same questions about accuracy and sharing without any incentive to be accurate.

We first examined whether accuracy incentives improved truth discernment, or the number of true headlines participants rated as true minus the number of false headlines participants rated as true (Batailler et al., 2021). As predicted, participants in the *accuracy incentives* condition ($M = 3.01$, 95% CI = [2.68, 3.34]) were better at discerning truth than those in the *control* condition ($M = 2.43$, 95% CI = [2.12, 2.73]), $t(457.64) = 2.58$, $p = 0.010$, $d = 0.24$. In other words, participants answered 11.01 (out of 16) questions correctly in the *accuracy incentives* condition, as opposed to 10.43 (out of 16) questions in the *control* condition.

We next examined whether incentives decreased partisan bias, or the number of politically-congruent headlines participants rated as true minus the number of politically-incongruent headlines participants rated as true. As predicted, partisan bias in accuracy

judgements was 31% smaller in the *accuracy incentives* condition (M = 1.31, 95% CI = [1.04, 1.58]) as compared to the *control* condition (M = 1.91, 95% CI = [1.62, 2.19]), $t(495.8) = 3.01$, $p = 0.001$, $d = 0.28$. Results from all four studies are plotted visually in **Fig 1.**

Additional analysis (See *Supplementary Appendix S1* for extended results) found that the *accuracy incentives* condition increased the percentage of politically-incongruent true headlines rated as true (M = 51.53, 95% CI = [47.36, 55.70]) as compared to the *control* condition (M = 38.25, 95% CI = [34.41, 42.08]), $p < 0.001$, $d = 0.43$. Incentives did *not* significantly impact judgements of politically-congruent true news, politically-incongruent false news, or politically-congruent false news when controlling for multiple comparisons with Tukey post-hoc tests ($ps >$ 0.444). Thus, the effects of incentives were mainly driven by an increased belief in true news from the opposing party.

Finally, we examined whether the incentives influenced sharing discernment, or the number of true headlines shared minus the number of false headlines people intended to share. Interestingly, even though sharing higher-quality articles was not explicitly incentivized, sharing discernment was slightly higher in the accuracy incentive condition (M = 0.38, 95% CI = [0.28, 0.48]) as compared to the control condition (M = 0.22, 95% CI = [0.15, 0.30]), $t(424.8) = 2.49$, $p = 0.037$, $d = 0.23$.

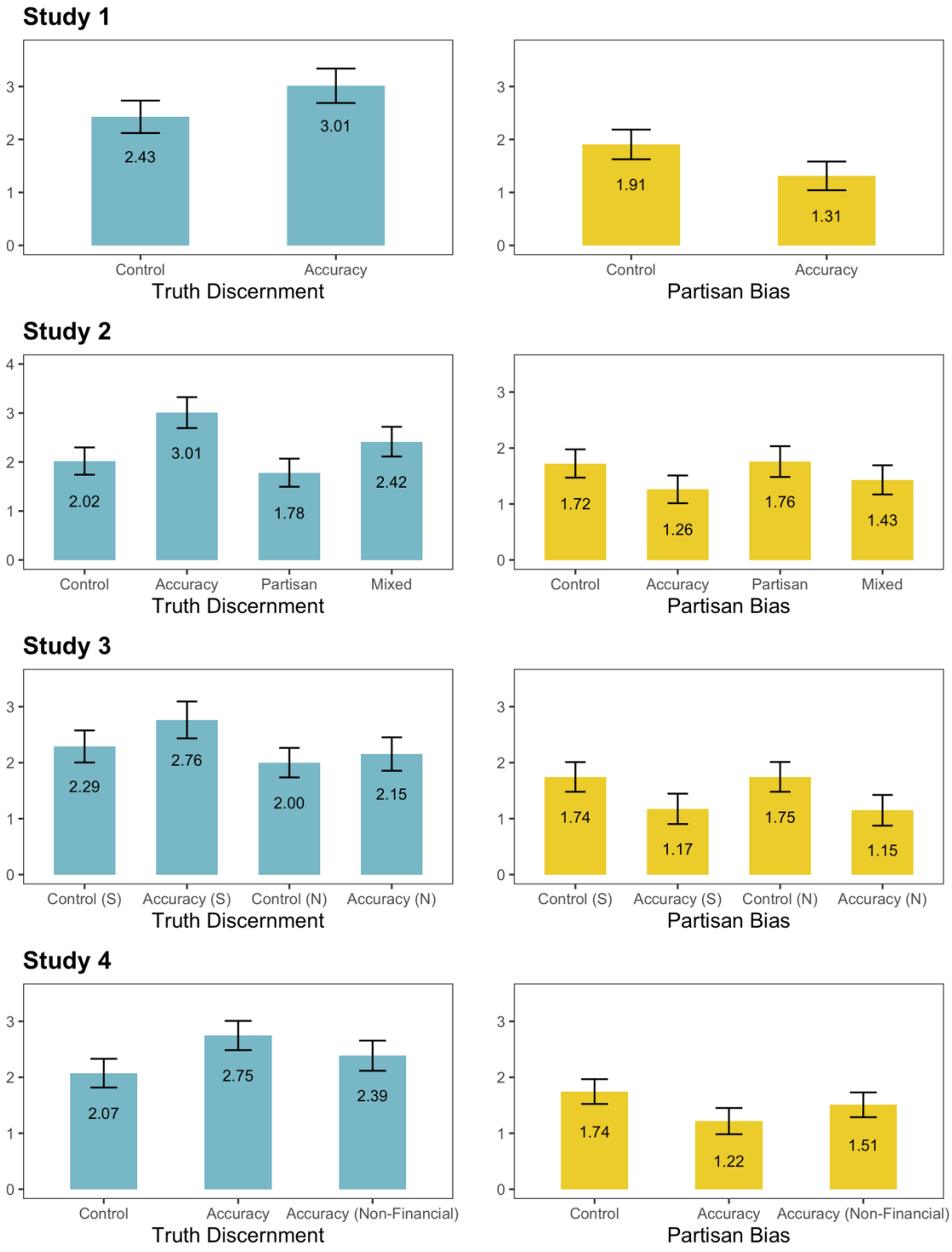


Figure 1. In **Study 1,** accuracy incentives improved truth discernment and decreased partisan bias in accuracy judgements, primarily by increasing belief in politically-incongruent true news. **Study 2** replicated these findings, but also found that incentives to identify articles that would be liked by one's political in-group decreased truth discernment – even when paired with the

accuracy incentive (the "mixed" condition). **Study 3** further replicated these findings and examined how effect sizes differed with and without source cues (S = source, N = no source). **Study 4** also replicated these findings and found that that a scalable, non-financial accuracy motivation intervention was also able to increase belief in politically-incongruent true news with a smaller effect size. Error bars represent 95% confidence intervals. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Experiment 2: Social Motivations Interfere with Accuracy Motivations

In Experiment 2, we aimed to replicate and extend on the results of Experiment 1 by examining whether social motivations to correctly identify articles that would be liked by one's political in-group might interfere with accuracy motives. This condition was meant to mirror the incentive structure of social media whereby people try to share content that will be liked by their friends and followers. We recruited another politically-balanced sample of 998 US adults (see ***Methods***). In addition to the *accuracy incentives* and *control condition*, we added a *partisan identity motivation* condition, whereby participants were given a financial incentive to correctly identify articles that would appeal to members of their own political party. Specifically, participants were told that they would receive a bonus payment of up to one dollar based on how accurately they identified articles that would be liked by members of their political party if they shared them on social media. Immediately after answering this question, participants were asked about the accuracy of the article and how likely they would be to share it. Building off of the predictions of the Identity-Based Model of Political Belief(Van Bavel & Pereira, 2018), we wanted to examine whether increasing partisan-identity related goals might interfere with accuracy goals. Thus, in a final condition, called the *mixed motivation condition*, participants received a financial incentive of up to one dollar to identify articles that would be liked by one's in-group, followed by an additional financial incentive to accurately identify true and false articles.

We first examined how these motivations influenced truth discernment. Replicating the results of Experiment 1, there was a significant main effect of the *accuracy incentives* condition on truth discernment, $F(1, 994) = 29.14$, $p < 0.001$, $\eta^2_G = 0.03$, a significant main effect of the *partisan identity* manipulation on truth discernment, $F(1, 994) = 7.53$, $p = 0.006$, $\eta^2_G = 0.01$, but no significant interaction between the accuracy and the partisan identity manipulation ($p =$

0.237). Tukey HSD post-hoc tests indicated that truth discernment was higher in the accuracy incentives condition (M = 3.01, 95% CI = [2.69, 3.32]) compared to the control condition (M = 2.02 95% = [1.74, 3.30]), $p < 0.001$, d = 0.41. Truth discernment was also higher in the accuracy incentives condition compared to the *partisan identity* condition (M = 1.78, 95% CI = [1.49, 2.07]), $p < 0.001$, d = 0.50, and the *mixed* condition (M = 2.42, 95% CI = [2.11, 2.71], p = 0.029, d = 0.27. However, the *mixed* condition did not differ from the control condition (p = 0.676), and the *partisan identity* condition also did not significantly differ from the control condition (p = 0.241). Taken together, these results suggest that accuracy motivations increase truth discernment, but partisan-identity motives can decrease truth discernment.

We then examined how these motives influenced partisan bias. Replicating the results from Experiment 1, there was a significant main effect of *accuracy incentives* on partisan bias, $F(1, 994) = 9.01$, p = 0.003, $\eta^2_G = 0.01$, but no effect of the *partisan identity* manipulation, $F(1, 994) = 0.60$, p = 0.441, $\eta^2_G = 0.00$, or the interaction between accuracy and the partisan identity manipulation, $F(1, 994) = 0.27$, p = 0.606, $\eta^2_G = 0.00$. Post-hoc tests indicated that there was a non-significant difference in partisan bias between the *accuracy incentives* condition (M = 1.26, 95% CI = [1.01, 1.51]) and the *control* condition (M = 1.72, 95% CI = [1.47, 1.98]), p = 0.062, d = 0.23 – a 27% decrease in partisan bias. There was a significant difference between the *accuracy incentives* condition and the *partisan identity* motives condition (M = 1.76, 95% CI = [1.48, 2.03]), p = 0.040, d = 0.24. No other post-hoc tests yielded significant differences (ps > 0.182).

Follow-up analysis (*Supplementary Appendix S1*) once again indicated that the incentives primarily impacted the percentage of *politically incongruent true* headlines rated as accurate (M = 55.61%, 95% CI = [51.68, 59.54]) when compared to the control condition (M = 37.65%, 95% CI = [33.83, 41.46]), $p < 0.001$, d = 0.58. The incentives again did not impact congruent true news, incongruent false news, or congruent false news (ps > 0.148).

There was no significant effect of accuracy incentives on sharing discernment (p = 0.996), diverging from the results of Study 1. However, follow-up analysis (*Supplementary Appendix S1*) indicated that those in the *partisan identity* motivation condition shared more politically-congruent news (either true or false) (M = 1.98, 95% CI = [1.90, 2.05]) as compared to the control condition (M = 1.80, 95% CI = [1.74, 1.87]), p = 0.015, d = 0.21. Additionally, those in the *mixed* condition (M = 2.02, 95% CI = [1.94, 2.10]) shared more politically-congruent

news (true or false) as compared to the control condition, $p < 0.001$, $d = 0.26$. Thus, prompting participants to identify whether an article will be liked by their political allies – whether or not they are also incentivized to be accurate – appears to indiscriminately increase intentions to share both true and false news that appeals to one's own political party.

Experiment 3: Accuracy Incentives and Source Cues in a Representative Sample

In Experiment 3, we sought to replicate our prior findings in a nationally representative sample in the United States. We recruited a sample of 921 US participants that was quota-matched to the national distribution on age, gender, ethnicity, and political party. We also tested a potential psychological process underlying the effects of accuracy incentives. Since prior work has found strong effects of source cues(Traberg & van der Linden, 2022) on judgements of news headlines, we suspected that people were responding to source cues when making judgements about news. Since true news often contains more recognizable sources with partisan connotations (e.g. "nytimes.com" as opposed to the fake news website "yournewswire.com")(Pennycook & Rand, 2019), this may explain why incentives only impacted judgements of true news in Experiments 1 and 2. To test this possibility, we examined the effect of incentives with and without source cues (e.g., a URL name such as "foxnews.com") present beside the headlines (see ***Methods*** for more details). Because we wanted to compare the effects of accuracy incentives with and without sources, this study had four conditions: accuracy incentives (with sources), control (with sources), accuracy incentives (without sources), and control (without sources).

Replicating the main results from Experiments 1 and 2, the *accuracy incentives* condition significantly improved truth discernment, $F(1, 917) = 4.44$, $p = 0.035$, $\eta^2_G = 0.01$, reduced partisan bias, $F(1, 917) = 18.21$, $p < 0.001$, $\eta^2_G = 0.02$, and increased the number of politically-incongruent true articles rated as accurate, $F(1, 917) = 20.94$, $p < 0.001$, $\eta^2_G = 0.02$. Thus, accuracy incentives appear to increase accuracy and reduce partisan bias in a large representative sample, suggesting that the results of these experiments likely generalize to the US population as a whole.

Although effect sizes appeared to be descriptively smaller when sources were removed from the headlines (see **Fig. 1** and *Supplementary Appendix S1* for detail), we did not find significant interactions between the main outcome variables and the presence or absence of source cues. However, this study design did not provide strong power to test whether this was

not due to chance, since interaction effects can require up to 16 times as much power as main effects(Blake & Gangestad, 2020; Gelman, 2018) (see ***Methods*** for power analysis). Additional analysis using Bayes factors(Wetzels et al., 2014) reported in *Supplementary Appendix S1* did not find strong evidence for the absence of interaction effects. Like in Experiment 2, there was once again no significant impact of accuracy incentives on sharing discernment ($p = 0.906$).

Experiment 4: The Effect of a Non-Financial Accuracy Motivation Intervention

In Experiment 4, we replicated the accuracy incentive and control condition in another politically-balanced sample of 983 US adults, but also added a *non-financial accuracy motivation condition*. This *non-financial accuracy motivation condition* was designed to rule out multiple interpretations behind our earlier findings. One mundane interpretation is that participants are merely saying what they believe fact-checkers think is true, rather than answering in accordance with their true beliefs. However, this non-financial intervention does not incentivize people to answer in ways that do not align with their actual beliefs. Additionally, because financial incentives are more difficult to scale to real-world contexts, the non-financial accuracy motivation condition speaks to the generalizability of these results to other, more scalable ways of motivating accuracy.

In the non-financial accuracy condition, people read a brief text about how most people value accuracy and how people think sharing inaccurate content hurts their reputation(Altay et al., 2019) (See intervention text in *Supplementary Appendix S2*.) People were also told to be as accurate as possible and that they would receive feedback on how accurate they were at the end of the study.

Our main pre-registered hypothesis was that this *non-financial accuracy motivation* condition would increase belief in politically-incongruent true news as relative to the *control* condition. An ANOVA found a main effect of the experimental conditions on the amount of politically-incongruent true news rated as true, $F(2, 980) = 17.53$, $p < 0.001$, $\eta^2_G = 0.04$. Supporting our main pre-registered hypothesis, the *non-financial accuracy motivation* condition increased the percentage of politically-incongruent true news stories rated as true ($M = 43.97$, 95% CI = [40.59, 47.34]) as compared to the *control* condition ($M = 35.19$, 95% CI = [31.93, 38.45], $p < 0.001$, $d = 0.29$. Replicating studies 1-3, the *accuracy incentive* condition also increased perceived accuracy of politically-incongruent true news ($M = 49.15$, 95% CI = [45.74,

52.55]), $p < 0.001$, $d = 0.45$. The *accuracy incentive* and *non-financial accuracy motivation* condition were not significantly different from one another ($p = 0.083$, $d = 0.17$), though this may be because we did not have enough power to detect a difference. In short, the *non-financial accuracy motivation* manipulation was also effective at increasing belief in politically-incongruent true news, with an effect about 63% as large as the effect of the financial incentive.

Since we expected the *non-financial accuracy motivation* condition to have a smaller effect than the *accuracy incentives* condition, we did not pre-register hypotheses for truth discernment and partisan bias, as we did not anticipate having enough power to detect effects for these outcome variables. Indeed, the *non-financial accuracy motivation* condition did not significantly increase truth discernment ($p = 0.221$) or partisan bias ($p = 0.309$). However, replicating studies 1-3, *accuracy incentives* once again improved truth discernment ($p = 0.001$, $d = 0.28$) and reduced partisan bias ($p = 0.003$, $d = 0.25$). The effect of the *non-financial accuracy motivation* condition was 47% as large as the effect of the *accuracy incentive* for truth discernment and 45% as large for partisan bias. There was also no overall effect of the experimental conditions on sharing discernment ($p = 0.689$). See *Supplementary Appendix S1* for extended results.

Together, these results suggest that a subtler (and also more scalable) accuracy motivation intervention that does not employ financial incentives may be effective at increasing the perceived accuracy of true news from the opposing party, but appears to have a smaller effect size than the stronger financial incentive intervention.

Integrative Data Analysis

To generate more precise estimates of our effects, we pooled data from all four studies[4] to conduct an integrative data analysis (IDA) (Curran & Hussong, 2009) For the IDA, we only used the 16 news headlines that were used in all four studies, and only included the *accuracy incentives* and *control* conditions that were used in all four studies.

Incentives had the largest positive effect on the perceived accuracy of politically-incongruent true news, $p < .001$, $d = 0.47$; and a smaller positive effect on the perceived accuracy of politically-congruent true news, $p = 0.001$, $d = 0.17$. Incentives did not significantly affect

[4] We did not have any studies in the file drawer on this topic, meaning that our estimate was not influenced by publication bias.

belief in politically-incongruent false news, $p = 0.163$, $d = 0.13$, or belief in politically-congruent false news, $p = 0.993$, $d = -0.04$ (See **Fig. 2**) after adjusting for multiple comparisons with Tukey post-hoc tests. Analysis for each individual item revealed that incentives significantly increased belief in all true items, but they did not significantly decrease belief in any false items (though they significantly increased belief in one false item). More details are reported in *Supplementary Appendix S1*, and a headline-level analysis is reported in *Supplementary Appendix S3*. Additional analysis using Bayes Factors reported in *Supplementary Appendix S4* found strong evidence that incentives impacted belief in both politically-congruent and politically-incongruent true news, but found inconsistent evidence that they affected belief in false news.

While effects on sharing discernment were inconsistent across studies, the IDA found that there was a small positive effect of the incentive on sharing discernment, $t(2020.20) = 2.19$, $p = 0.029$, $d = 0.10$. Finally, people spent slightly more time on each headline in the accuracy incentives condition, $t(818.53) = 2.34$, $p = 0.019$, $d = 0.16$, indicating that incentives may have led people to put more effort into their responses.

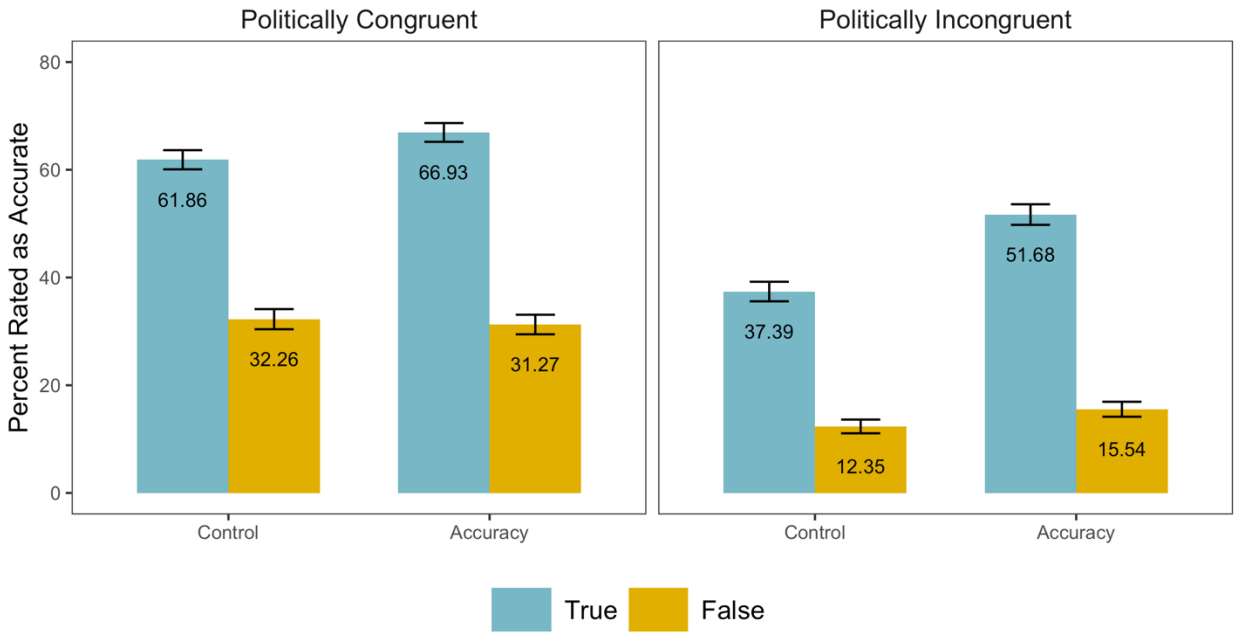


Figure 2. Integrative data analysis results (with data from all four studies, n = 2,092) broken up by headline type. Incentives had a large effect on belief in politically-incongruent true news, and also had an effect on politically-congruent true news. Incentives did not have a significant effect on politically-congruent or politically-incongruent false news when controlling for multiple

comparisons. Headline-level analysis revealed that incentives increased belief in all 8 true items, but did not decrease belief in a single false item (See *Supplementary Appendix S3* for item-level analysis*)*.

Incentives Reduce the Accuracy Gap Between Liberals and Conservatives

Replicating prior work (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Lawson & Kakkar, 2021; Pereira & Van Bavel, 2018; van der Linden, Panagopoulos, et al., 2021), conservatives were worse at discerning between true and false headlines than liberals. Conservatives answered about 9.26 (out of 16) questions correctly when not incentivized to be accurate and liberals answered 10.93 questions out of 16 correctly when unincentivized – a 1.67-point difference, 95% CI = [1.41, 1.94], $t(1035.69) = 12.53$, $p < .001$, $d = 0.77$. But, when conservatives were incentivized to be accurate, they answered 10.12 questions correctly—making the gap between incentivized conservatives and unincentivized liberals 0.81 points, 95% CI [0.53, 1.09], $t(951.91) = 5.65$, $p < .001$, $d = 0.35$. In other words, paying conservatives less than a dollar to correctly identify news headlines as true or false reduced the gap in performance between conservatives and (unincentivized) liberals by 51.50%. Incentives also considerably reduced the gap between conservatives and liberals in terms of partisan bias, sharing discernment, and belief in politically-incongruent true news. More detail is reported in *Supplementary Appendix S1* and plotted visually in **Fig. 3**. Altogether, these results suggest that a substantial portion of US conservatives' tendency to believe and share less accurate news reflects a lack of motivation to be accurate rather than lack of knowledge alone.

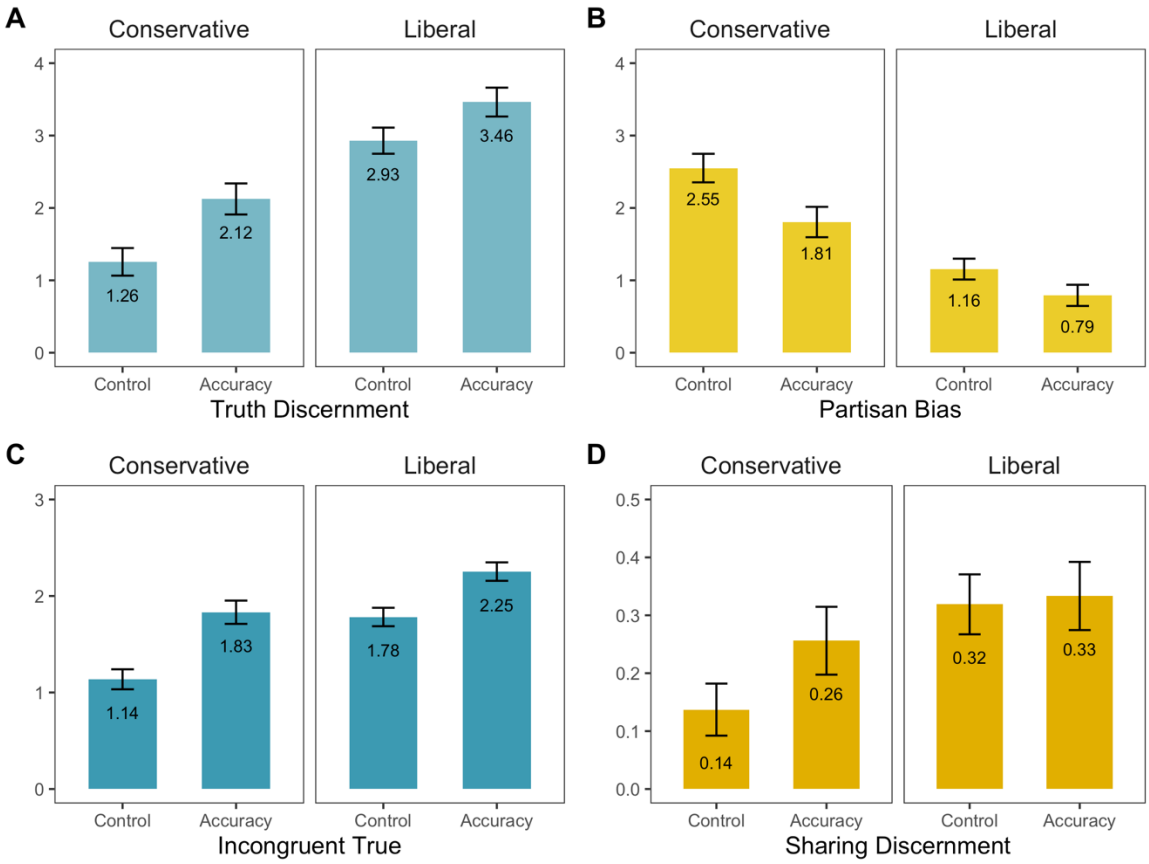


Figure 3. Conservatives were worse at truth discernment as compared to liberals (**Panel A**). They also showed more partisan bias (**Panel B**), less belief in politically-incongruent true news (**Panel C**), and worse sharing discernment (**Panel D**). However, incentives closed the gap between conservatives and liberals for all these outcome variables by more than half, suggesting that conservatives' greater tendency to believe in and share (mis)information may in part reflect a lack of motivation to be accurate (instead of lack of knowledge or ability alone).

Importantly, the incentives improved truth discernment for both liberals, $d = 0.23$, $p < 0.001$, and conservatives, $d = 0.40$, $p < 0.001$ (see *Supplementary Appendix S5* for table of effect sizes broken down by political affiliation). Descriptively, the effect sizes for our intervention were larger for conservatives than liberals, which diverges from other misinformation interventions that tend to show larger effect sizes for liberals (Pretus et al., 2021; Rathje, 2022). Furthermore, political ideology (liberal vs. conservative) was a significant moderator of belief in incongruent true news, $p = 0.033$, and partisan bias, $p = 0.029$, (though this moderation effects was not significant for truth discernment, $p = 0.095$, or sharing discernment, $p = 0.061$) such that

the effects of incentives appeared to be larger for conservatives than liberals. The effect of the incentives on truth discernment was not significantly moderated by cognitive reflection, political knowledge, or affective polarization (*ps* < 0.182). However, even though we had a large sample, we were still slightly underpowered to detect these interaction effects (see power analysis in ***Methods***), and supplemental Bayesian analyses also did not find strong evidence for the significant moderation effects (*Supplementary Appendix S11*), so these interaction effects should be interpreted with caution.

Relative Importance of Accuracy Incentives

In each experiment, we measured other individual difference variables known to be predictive of truth discernment, such as cognitive reflection, political knowledge, partisan animosity, as well as demographic variables, such as age, education, and gender. We ran a multiple regression analysis on our IDA with all of these variables included in the model (**Fig. 4**, **Panel A**). To compare the relative importance of each of these predictors, we also ran a relative importance analysis using the "lmg" method (Tonidandel & LeBreton, 2011), which calculates the relative contribution of each predictor to the R^2 (**Fig. 4, Panel B**). Full models and relative importance analyses are in *Supplementary Appendix S6* and *S7.*

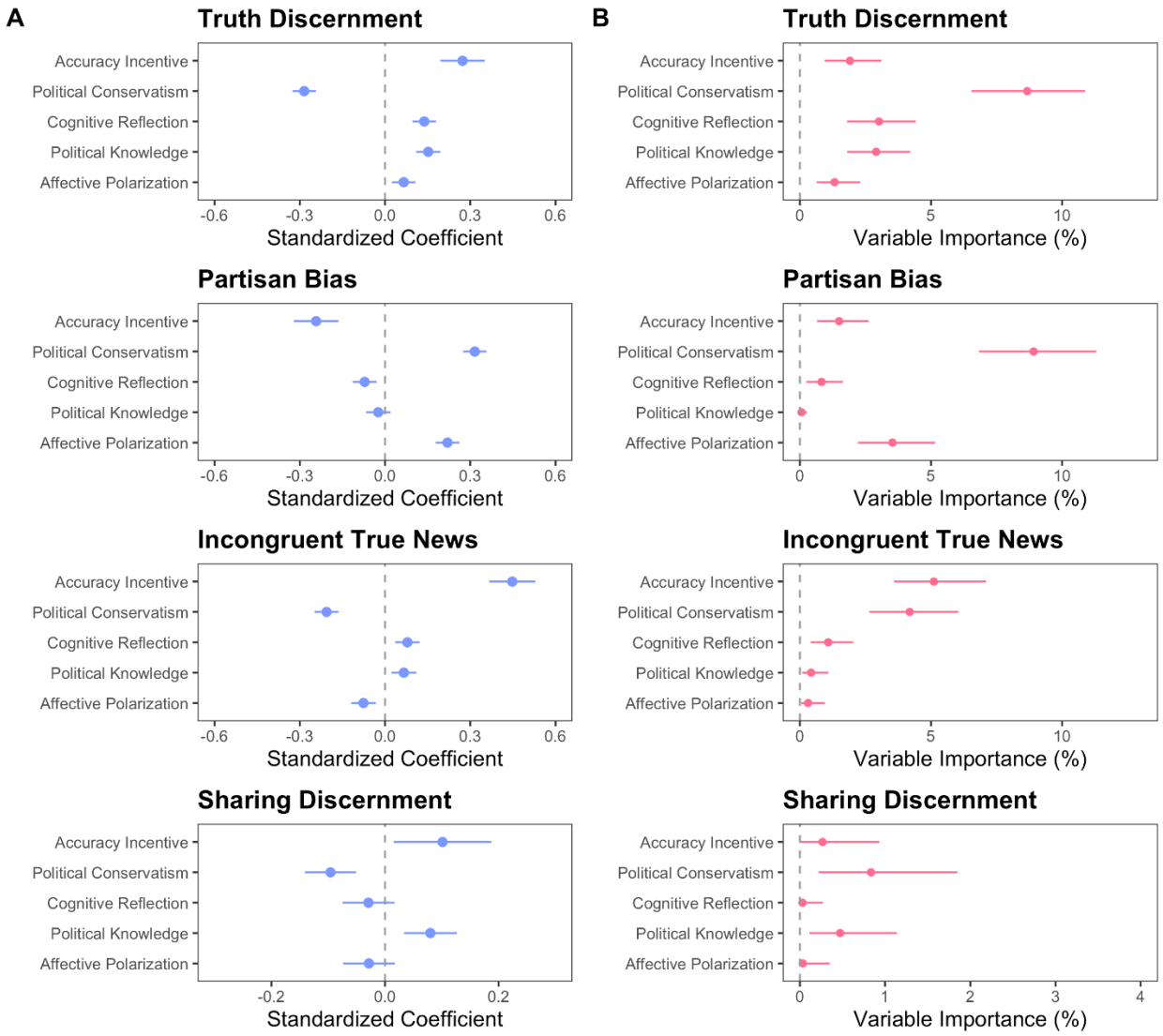


Figure 4. In **A**, multiple regression results for the main outcome variables: truth discernment, partisan bias, belief in incongruent true news, and sharing discernment. Standardized beta coefficients are plotted for ease of interpretation. In **B**, variable importance estimates (LMG values) with bootstrapped confidence intervals are shown to examine the estimated percentage contribution of each predictor to the R^2.

Political conservatism and accuracy incentives were among the most important predictors for many of the key outcome variables, although confidence intervals were large and overlapping for the relative importance analysis (See *Supplementary Appendix S4*). While prominent accounts of misinformation sharing claim that partisanship and politically motivated cognition play a limited role in the belief and sharing of misinformation as compared to other factors (such

as reflection or inattention) (Pennycook et al., 2021; Pennycook & Rand, 2021b), our results indicate that motivation and partisan identity or ideology are indeed very important factors.

Our data point to the importance of broad theoretical accounts of (mis)information belief and sharing that integrate motivation and partisan identity with other variables (Pennycook & Rand, 2021c; C. E. Robertson, Pretus, et al., 2022; Van Bavel, Harris, et al., 2021; van der Linden, 2022; van der Linden, Roozenbeek, et al., 2021). Indeed, an investigation using cognitive modeling found that a broad model of misinformation belief that included multiple factors (such as partisan identity, cognitive reflection, and more) performed better at predicting acceptance of misinformation than other models that included fewer variables (Borukhson et al., 2022).

Discussion

Across four experiments (n = 3,364), we find that increasing people's motivation to be accurate via a small financial incentive of up to one-dollar improved accuracy in discerning between true and false news and decreased the partisan divide in belief in news by about 30%. These effects were driven primarily by an increased belief in politically-incongruent true news (d = 0.47), and no significant effects were found for false news, which people encounter relatively infrequently online (A. M. Guess, Nyhan, et al., 2020). Furthermore, providing people with an incentive to identify articles that would be liked by their political in-group reduced accuracy and increased intentions to share politically-congruent true and false news. Thus, social or partisan identity goals appear to interfere with accuracy goals. Additionally, a non-financial accuracy motivation intervention that asked people to be accurate, provided people feedback about their accuracy, and emphasized the social norm and reputational benefits of being accurate, significantly increased the perceived accuracy of politically-incongruent true news (d = 0.29). This illustrates that accuracy motivation interventions can be applied at scale.

These results make two key theoretical contributions. First, they suggest that partisan differences in news judgements do not simply reflect differences in factual knowledge (Pennycook & Rand, 2021b). Instead, our data suggest that a substantial portion of this partisan divide can be attributed to a lack of motivation to be accurate. While there have been debates about whether partisan differences in belief reflect differing prior beliefs versus politically-motivated cognition (Druckman & McGrath, 2019; Tappin et al., 2020), our studies provide

causal evidence for the effect of motivation on belief. Along with other research (G. F. Bishop, 2004; Edwards, 1957; Prior et al., 2015), these findings suggest that survey data about belief in (mis)information should not be taken at face value, because people answer survey questions differently when they are highly motivated to be accurate. However, judgements of false headlines appeared to be unaffected by accuracy motivations, suggesting that other factors may play a more prominent role in people's assessment of false news as compared to true news.

Second, while a number of studies have observed that American conservatives tend to be more susceptible to misinformation than liberals (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Lawson & Kakkar, 2021; Pereira & Van Bavel, 2018; van der Linden, Panagopoulos, et al., 2021), our studies find that the gap in accuracy between liberals and (unincentivized) conservatives closes by more than half when conservatives are motivated to be accurate. Future work could examine whether this assymetry arises due to the dynamics of partisan identity, party leadership, and social norms in the United States during this specific political climate, or if it reflects broader differences between liberals and conservatives that can be observed across cultures (Imhoff et al., 2022; Jost et al., 2018).

These results also have practical implications for interventions (Bak-Coleman et al., 2022; Roozenbeek et al., 2022). Accuracy incentives improved the accuracy of people's judgements, and an integrative data analysis found that this effect may have spilled over into intentions to share more accurate articles. However, the effect on sharing intentions was small and inconsistent across studies. This may be in part because people were asked about accuracy before being asked about sharing intentions, and past research has found that merely asking people about accuracy can improve the accuracy of sharing intentions (Pennycook et al., 2021). Further, making partisan-identity motivations salient increased the sharing of both politically-congruent false (and true) news. Thus, interventions and social media design features should aim to both *increase* accuracy motivations and *decrease* motivations to share inaccurate content that receives high social reward. While effects were only found for false (and not true) headlines, people tend to encounter blatantly false news very infrequently (A. M. Guess, Nyhan, et al., 2020), leading some to suggest that increasing trust in reliable news is more important than reducing belief in falsehoods (Acerbi et al., 2022) and that researchers should employ a broad definition of misinformation (Traberg, 2022).

One limitation of this work is that survey experiments have unknown ecologically validity. To maximize ecological validity, we used real, pre-tested news headlines in the format in which they would be regularly encountered on social media websites such as Facebook. Additionally, self-reported sharing intentions are highly correlated with real online news sharing (Mosleh et al., 2020), and a field experiments suggests that priming accuracy can improve news sharing decisions on Twitter (Pennycook et al., 2021), illustrating that results from survey experiments on misinformation can translate to the field. Another potential limitation is that there are multiple ways to interpret the effects of financial incentives. For instance, people may be guessing what they think fact-checkers believe to earn money, rather than expressing their true beliefs. However, this interpretation is unlikely to explain the full effect, since a subtle non-financial accuracy motivation intervention had similar (albeit smaller) effects. Furthermore, supplementary analysis found that an extremely small percentage of participants reported answering in ways that did not accord with their true beliefs to receive money (See *Supplementary Appendix S1*).

Conclusions

There is a sizable partisan divide in the kind of news liberals and conservatives believe in, and conservatives tend to believe in and share more false news than liberals. Yet, these differences are not immutable. Motivating people to be accurate improves accuracy about the veracity of (true but not false) news headlines, reduces partisan bias, and closes a substantial portion of the gap in accuracy between liberals and conservatives. Theoretically, these results identify accuracy and social motivations as key factors in driving news belief and sharing. Practically, these results suggest that shifting motivations may be a useful strategy for improving the quality of the news content that people consume and share online.

Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the experiment. The research methods were approved by the University of Cambridge Psychology Ethics Committee (Protocol #PRE.2020.110). These studies were pre-registered. Stimuli, Qualtrics survey files, anonymized data, analysis code, and all pre-registrations are available on our OSF page: https://osf.io/75sqf.

Experiment 1

Participants. The experiment launched on November 30, 2020. We recruited 500 participants via the survey platform Prolific Academic (Peer et al., 2021). Specifically, we recruited 250 conservative participants and 250 liberal participants from the US via Prolific Academic's demographic pre-screening service to ensure the sample was politically balanced. Our a priori power analysis indicated that we would need 210 participants to detect a medium effect size of $d = 0.50$ at 95% power, though we doubled this sample size to account for partisan differences and oversampled to account for exclusions. 511 participants took our survey. Following our pre-registered exclusion criteria, we excluded 32 participants who failed our attention check (or did not get far enough in the experiment to reach our attention check), and an additional 17 participants who said they responded randomly at some time during the experiment. This left us with a total of 462 participants (194 M, 255 F, 12 Trans/Nonbinary; age: $M = 35.85$, $SD = 13.66$; Politics: 253 Democrats, 201 Republicans). The Experiment 1 pre-registration is available here: https://aspredicted.org/blind.php?x=gk9xg5.

Materials. The materials were 16 pre-tested true and false news headlines from a large pre-tested sample of 225 news headlines (Pennycook et al., 2020). In total, eight of these news headlines were false, and eight of the news headlines were true. Because we were interested in whether accuracy incentives would reduce partisan bias, we specifically selected headlines that had a sizable gap in perceived accuracy between Republicans and Democrats as reported in the pre-test, as well as headlines that were not outdated (the pre-test was conducted a few months before the first experiment). Specifically, we chose eight headlines (four false and four true) that Democrats rated as more accurate than Republicans in the pre-test, and eight headlines (four false and four true) that Republicans rated as more accurate than Democrats. See *Supplementary Appendix S8* for example stimuli and the OSF page for full materials.

News Evaluation Task. Participants were shown these 16 news headlines, along with an accompanying picture and source (similar to how a news article preview would show up on someone's Facebook feed), and asked "To the best of your knowledge, is the claim in the above headline accurate?" on a scale from 1 ("extremely inaccurate") to 6 ("extremely accurate"). Afterwards, they were asked "If you were to see the above article on social media, how likely would you be to share it?" on a scale from 1 ("extremely unlikely") to 6 ("extremely likely").

Accuracy Incentives Manipulation. Half of the participants were randomly assigned to a *control condition,* in which we explained the news evaluation task, but we did not provide any information about a bonus payment. The other half were assigned to an *accuracy incentives* condition. In this condition, we explained the news evaluation task, and then told participants they would receive a "bonus payment of up to $1.00 based on how many correct answers [they] provide regarding the accuracy of the articles. Correct answers are based on the expert evaluations of non-partisan fact-checkers." Specifically, they received one dollar for answering 15 out of 16 questions correctly, and fifty cents for answering 13 out of 16 questions correctly. Since we measured accuracy on a continuous scale, we told participants that "if the headline describes a true event, either 'slightly accurate,' 'moderately accurate,' or 'extremely accurate' constitute correct responses. Similarly, if the headline describes a false event, either 'extremely inaccurate,' 'moderately inaccurate,' or 'slightly inaccurate' constitute 'correct' responses." In other words, the continuous scale was measured dichotomously for the purposes of giving financial incentives. Participants were also notified that all other questions would not affect their bonus payment. See *Supplementary Materials S2* or the OSF for full manipulation text.

Other Measures. We gave participants a 3-item cognitive reflection task (Pennycook & Rand, 2018). We measured participants' political knowledge using a 5-item scale (Osmundsen et al., 2021) and in-group love/out-group hate with feeling thermometers (Druckman & Levendusky, 2019). See *Supplementary Appendix S9* and the OSF for question text. These measures were repeated across all studies.

Analysis. For truth discernment, partisan bias, and sharing discernment, independent samples t-tests were used. While we asked participants to rate the truth of headlines on a continuous scale, these variables were recoded as dichotomous for analysis because the financial incentive only rewarded participants based on whether they correctly identified a headline as true or false. Since we did not clearly specify this in the Experiment 1 pre-registration (but did for Experiments 2-4),

we show the results with a continuous coding in *Supplementary Appendix S10.* The continuous coding did not change the conclusions of our studies.

To test what types of headlines were affected by the incentives, we ran a 2 (accuracy incentive vs. no incentive) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA with the percent of articles rated as accurate as the dependent variable, and then followed up with Tukey HSD post-hoc tests. Extended analyses are in *Supplementary Appendix S1*.

Experiment 2

Participants. The experiment launched on January 22, 2021. We aimed to recruit 1000 total participants (250 per condition) via the survey platform Prolific Academic, though we over-sampled and recruited 1,100 to account for exclusion criteria. We chose this sample size because a power analysis revealed that we needed at least 216 participants per condition to detect the smallest effect size ($d = 0.24$) at 0.80% power using a one-tailed t-test (although two-tailed tests were used for all analysis). Once again, we used Prolific's pre-screening platform to recruit 550 liberals and 550 conservatives from the United States, and 1,113 participants took our survey. Following our pre-registered exclusion criteria, we excluded 76 participants who failed our attention check (or did not finish enough of the survey to reach the attention check) and an additional 39 participants who said they responded randomly at some point during the experiment. This left us with a total of 998 participants in total (463 M, 505 F, 30 transgender/non-binary/other; age: $M = 36.17$, $SD = 13.94$; politics: 568 liberals, 430 conservatives). This experiment was also pre-registered (pre-registration available here: https://aspredicted.org/blind.php?x=/FKF_15L).

Social Incentives & Mixed Incentives Manipulations. In the new *partisan identity* condition, participants were first asked before the experiment to report the political party with which they identify. Then, they were told that they would receive a bonus payment of up to $1.00 based on how accurately they identified information that would be liked by members of their political party if they shared it on social media. Bonuses were awarded based on how closely participants' answers matched partisan alignment scores from a pre-test[48]. Before each question about accuracy and sharing, participants were asked "If you shared this article on social media, how likely is it that it would receive a positive reaction from [your political party] (e.g., likes, shares,

and positive comments)?" In the *mixed* condition, participants were first given financial incentives for both correctly identifying whether the article would be liked by a member of their political party, and were then asked about accuracy and given incentives for identifying whether the article was accurate. See *Supplementary Appendix S2* for full intervention text.

Analysis. To understand the impact of accuracy and partisan identity motivations on truth discernment and partisan bias, we ran 2 (accuracy incentive vs. control) X 2 (partisan identity vs. control) ANOVAs and followed up on the results using Tukey HSD post-hoc tests. To test what types of headlines were affected by the incentives, we ran a 2 (accuracy vs. control) X 2 (partisan identity vs. control) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA with the percentage of articles rated as accurate as the dependent variable, and then followed up with Tukey HSD post-hoc tests.

Experiment 3

Participants. The experiment launched on June 13, 2021. We aimed to recruit a nationally representative sample (quota-matched to the US population distribution by age, ethnicity, and gender) of 1,000 participants via the survey platform Prolific. As in studies 1 and 2, we ensured that the nationally-representative sample was politically balanced, or half liberal and half conservative. 1,055 total participants took the survey. Then, we once again excluded 95 participants who failed our attention check (or did not make it to that point in the survey), as well as 39 participants who said they were responding randomly at some point in the survey. This left us with a total of 921 participants (439 M, 470 F, 12 transgender/non-binary/other; age: M = 40.07, SD = 14.67; politics: 542 liberals, 379 conservatives). This experiment was also pre-registered (pre-registration available at: https://aspredicted.org/7M2_9K9).

Materials. We once again used the same 16 pre-tested true and false news headlines in addition to eight extra true and false news items from the same pre-test. For consistency, we report the results of the 16 news items in the manuscript, but we also report the results for the full set of 24 items in the *Supplementary Appendix S3,* which did not change our conclusions.

Manipulations. In addition to the accuracy incentive and control condition, participants were assigned to identical accuracy incentive and control conditions *without source cues* present on the stimuli. In these conditions, the sources (e.g., "nytimes.com") were greyed out, so participants could only make assessments of the stimuli based on the photo and headline alone

(see *Supplementary Materials S8* for examples).

Analysis. To understand the impact of accuracy incentives and source cues on truth discernment and partisan bias, we ran 2 (accuracy vs. control) X 2 (source vs. no source) ANOVAs and followed up on the results using Tukey HSD post-hoc tests. To test what types of headlines were affected by the incentives, we ran a 2 (accuracy vs. control) X 2 (source vs. no source) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA with the percent of articles rated as accurate as the dependent variable, and then followed up with Tukey HSD post-hoc tests.

Power Analysis for Interaction Effects. Based on the effect sizes of Study 2 and the principle that 16 times the sample size is needed to detect an attenuated interaction effect (Blake & Gangestad, 2020; Gelman, 2018), a power analyses conducted after we ran the study found that we needed roughly 1536 participants to detect an interaction for the amount of politically-incongruent news rated as true, 2560 participants to detect an interaction effect for truth discernment, and 7488 participants to detect an interaction effect for partisan bias with 80% power. Thus, this particular design was underpowered to detect whether accuracy incentives interacted with source cues.

Experiment 4

Participants. This experiment launched on May 25, 2022. We aimed to recruit a total of 1000 participants (roughly 333 per condition) via the platform Prolific academic. We chose this sample size as a power analysis found that we would 312 per condition to detect the smallest effect size found in the previous study ($d = 0.26$) with 90% power. Additionally, we wanted relatively high power because we expected the effect of the non-financial accuracy motivation condition to be smaller than that of the financial incentive condition. We used Prolific's pre-screening platform to recruit a sample that was balanced by politics and gender. 1007 participants took our survey. Following our pre-registered exclusion criteria, we excluded 17 participants who failed our attention check (or did not finish enough of the survey to reach the attention check) and an additional 8 participants who said they responded randomly at some point during the experiment. This left us with a total of 993 participants in total (486 M, 483 F, 30 transgender/non-binary/other; age: $M = 41.46$, $SD = 15.06$; politics: 507 liberals, 476

conservatives). This experiment was also pre-registered (pre-registration available here: https://aspredicted.org/86W_BY4).

Materials. We once again used the same 16 pre-tested true and false news headlines extra "misleading" news headlines.

Analysis. Following our pre-registered analysis plan, we ran a 1-way (accuracy vs. control vs. non-financial accuracy motivation) ANOVA with the percent of incongruent-true articles rated as true as the dependent variable, followed up by Tukey post-hoc tests. We also ran 1-way ANOVAs with truth discernment and partisan bias and DVs and followed up with post-hoc tests.

Integrative Data Analysis

Analysis. We conducted moderation analysis on the pooled dataset by testing for an interaction between the condition and political ideology (liberal vs. conservative) in a linear regression. To test the relative importance of each predictor, we ran a relative importance analysis using the "reliampo" package in R. Bootstrapped confidence intervals were calculated for "lmg" variables using 1,000 bootstraps.

Power Analysis for Moderation Effects. Using effect sizes from the integrative data analysis and the principle that 16 times the sample size is needed to detect an attenuated interaction effect (Blake & Gangestad, 2020; Gelman, 2018), a post-hoc power analysis found that we needed 2336 participants to detect an interaction effect for the amount of politically-incongruent news rated as true, 5984 participants to detect an interaction effect for truth discernment, 7488 for partisan bias, and 50,336 to detect an interaction for sharing discernment. Thus, moderation effects should be interpreted with caution.

Signal Detection Analysis. As another robustness check, we also conducted supplemental analysis using signal detection modeling (Batailler et al., 2021). This analysis found that incentives increased participants' discrimination between true and false news (for both politically-congruent and politically incongruent headlines), and also increased the threshold by which people accepted politically-incongruent headlines as true (See *Supplementary Appendix S12*). In sum, analysis using signal detection modeling yielded highly similar results to our main analysis.

Chapter 4. Partisan Differences in the Effectiveness of Nudging Accuracy

A version of this chapter was published in *Psychological Science* with co-authors Jon Roozenbeek, Cecilie Steenbuch Traberg, Jay Van Bavel, and Sander van der Linden.

The following article can be cited:

Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the editors of psychological science: meta-analysis reveals that accuracy nudges have little to no effect for US conservatives: regarding Pennycook et al. (2020). *Psychological Science.*

Introduction

In response to the global spread of misinformation online, scholars have tried to develop scalable behavioral interventions to help counter misinformation belief and sharing (A. M. Guess, Lerner, et al., 2020b; Pennycook, Epstein, et al., 2021b; Roozenbeek et al., 2022; Roozenbeek & van der Linden, 2019b). One proposed misinformation-reduction intervention is called an "accuracy nudge," whereby people are subtly nudged or primed to consider accuracy before sharing a news article (Pennycook, Epstein, et al., 2021b; Pennycook, McPhetres, et al., 2020). The logic behind this intervention is that inattention to accuracy is a key factor behind the sharing of misinformation. Indeed, the authors of this intervention estimate that inattention to accuracy drives approximately 50% of the sharing of misinformation online (Pennycook, Epstein, et al., 2021b). Several lab experiments and a Twitter field experiment have shown that asking people about the accuracy of an unrelated headline reduces intentions to share inaccurate articles (Pennycook, Epstein, et al., 2021b; Pennycook, McPhetres, et al., 2020). The authors claim that their results "challenge the popular claim that people value partisanship over accuracy" (abstract) and note that the accuracy prime "significantly increased sharing discernment for both Democrats […] and Republicans" (Pennycook et. al, 2021, p.2). They further suggest that that "partisanship is not, apparently, the key factor distracting people from considering accuracy on social media" (Pennycook et. al, 2020, p. 777).

Building on several prior studies that find that conservatives are more likely to believe in and share information (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Lawson & Kakkar, 2021; Pereira & Van Bavel, 2018; van der Linden, Panagopoulos, et al., 2021), we examined whether conservatives were also less susceptible to the accuracy nudge intervention. While the authors argue that inattention, not partisanship, drives the sharing of misinformation online, we were interested in whether inattention and partisanship interact. This investigation has implications for forming integrated theoretical accounts of misinformation belief and sharing (Van Bavel, Harris, et al., 2021; van der Linden, Roozenbeek, et al., 2021), and practical implications for helping to reduce the spread of misinformation online among a population that tends to share misinformation more often.

Results

To examine the role of partisanship in accuracy nudges, we meta-analyzed data from "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention" (Pennycook, McPhetres, et al., 2020), a pre-registered replication of that paper (Roozenbeek et al., 2021) and three studies from a highly similar *Nature* paper called "Shifting Attention to Accuracy Can Reduce Misinformation Online" (Pennycook et al., 2021). In these studies, treatment group participants were asked about the accuracy of an unrelated headline before answering about their intentions to share several true and false news headlines. This subtle "accuracy nudge" reportedly improved sharing discernment, or sharing intentions of true headlines minus false headlines.

We first analyzed the data separately for Democrats and Republicans (excluding independents) for all five of the studies, see ***Table 1***. Independent samples t-tests found that the accuracy nudge significantly improved sharing discernment for Democrats in four of these studies, and the effect was marginally significant in one study (all $ps > 0.077$). However, the effect of the accuracy nudge was not significant for Republicans in any of the five samples (all $ps > 0.157$).

We then conducted a fixed effects internal meta-analysis using the R package "meta" and the DerSimonian-Laird estimator. The meta-analyzed effect size for Republicans across all five studies was negligible (*Mean ES* = 0.11, 95% CI = [0.00, 0.22], $p = 0.050$) in comparison to the meta-analyzed effect size for Democrats (*Mean ES* = 0.32, 95% CI = [0.23, 0.41], $p < 0.001$). In other words, the accuracy prime explains about 0.13% of the variance in sharing decisions for Republicans, as compared to about 2.5% for Democrats.

Table 1. Re-analysis of studies 3, 4 and 5 from Pennycook et al. (2020), Pennycook et al. (2021) and Roozenbeek, Freeman & van der Linden (2021), broken down by US political party.

	Democrats				Republicans			
Study	Cohen's *d*	95% CI	*p*	*n*	Cohen's *d*	95% CI	*p*	*n*
Study 3 (Pennycook et al., 2021)	0.38	[0.16; 0.60]	0.001	318	0.13	[-0.22; 0.48]	0.483	133
Study 4 (Pennycook et al., 2021)	0.42	[0.21; 0.63]	0.000	364	0.23	[-0.10; 0.55]	0.175	153
Study 5 (Pennycook et al., 2021)	0.23	[-0.02; 0.48]	0.077	277	0.07	[-0.21; 0.35]	0.624	211
Psych Science Paper (Pennycook, McPhetres et al., 2020)	0.32	[0.10; 0.54]	0.005	326	0.17	[-0.07; 0.41]	0.157	271
Psych Science Replication Study (Roozenbeek et al., under review)	0.27	[0.11; 0.42]	0.001	636	0.06	[-0.12; 0.24]	0.522	486
Mean Effect Size (All Studies)	**0.32**	**[0.23; 0.41]**	**0.001**	**1921**	**0.11**	**[-0.00; 0.22]**	**0.050**	**1254**

Note. Independent samples *t*-tests and Cohen's *d* effect size measurements for each study, run separately for Republicans (average $d = 0.11$) and Democrats ($d = 0.32$), excluding independents or unaffiliated participants. Fixed effects meta-analyses using the DerSimonian-Laird estimator were used to estimate mean effect sizes across all five studies.

We added another dataset at the request of the authors from the paper "Developing an accuracy prompt toolkit to reduce COVID-19 misinformation online" by Epstein et al. (2021), which added another 497 Republicans to the sample. This new dataset led to a similar effect size for Republicans (Mean *ES* = 0.09, 95% CI = [0.01, 0.19], $p = 0.028$). Analyses with and without data from this dataset are in the ***Supplementary Appendix***.

To test if the effectiveness of the accuracy prime was significantly moderated by party affiliation, we pooled the data from all five studies to test for an interaction effect between the accuracy prime treatment and political party. We found a significant interaction such that accuracy primes were less effective for Republicans than Democrats, $B = -0.15$, $SE = 0.06$, $p < 0.009$, see ***Figure 1***. This interaction effect remained significant when measured across six different measures of conservatism (See ***Supplementary Appendix Table S3*** and ***S4)***. It also remained significant when accounting for additional data supplied by the authors after our initial analysis. Moreover, the accuracy prime was the least effective for extreme conservatives (see ***Figure 1*** and ***Table S4***). Following Pennycook et al., (2020), we also conducted a series of linear regressions at the rating level, clustered on participants and headlines, which broadly support these findings (see ***Supplementary Appendix S1*** and ***Tables S5-S12)***. The code, data, and survey materials are freely available: https://osf.io/hgd3k/.

It is unclear exactly why accuracy primes are less effective for more conservative individuals. One possible explanation is that conservatism is negatively associated with trust in national media ($r = -0.31$, 95% CI [-0.34, -0.29], $t(4593) = -22.29$, $p < .001$) when analyzing the sample of five pooled studies. Thus, even when nudged to be accurate, conservatives may still see "accurate" sources, i.e., mainstream media outlets such as the New York Times, as biased[7,8]. Furthermore, the self-reported importance of sharing accurate content has a small negative correlation with conservatism across the five pooled studies ($r = -0.09$, 95% CI [-0.12, -0.07], $t(4591) = -6.43$, $p < .001$), suggesting that conservatives may have slightly lower accuracy motivations. Lastly, when re-analyzing the authors' data from Study 2 of Pennycook et. al, (2021) we find that Republicans say it is more important to share politically-congruent articles ($M = 3.62$, 95% CI = [3.17, 3.39]) than Democrats ($M = 3.05$, 95% CI = [2.88, 3.21]), $t(296.03) = 2.45$, $p = 0.015$, $d = 0.30$), suggesting that conservatives/Republicans might have slightly higher motivations to share politically-congruent content. Thus, partisan or ideological asymmetries in factors such as trust in media or motivations for accuracy and political

congruence may help explain our findings (Jost et al., 2018). Alternatively, if conservatives are worse at recognizing fake news (Pennycook, Epstein, et al., 2021b; Pennycook, McPhetres, et al., 2020), this may make the nudge less effective, since even when motivated to share accurate content, conservatives may have less accurate knowledge of what they should share.

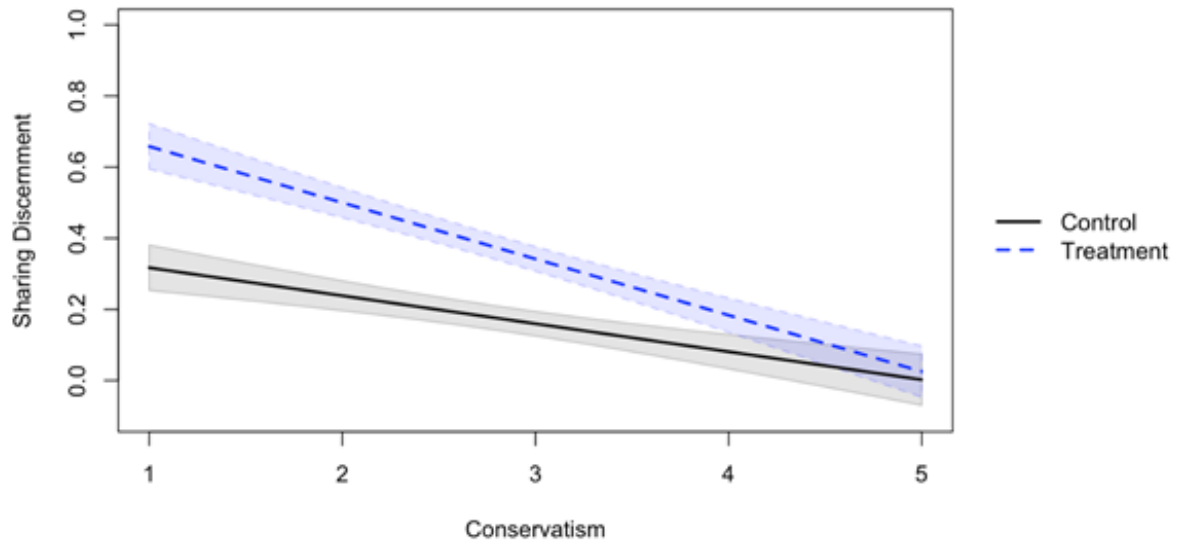


Figure 1. Predicted values of sharing discernment in the accuracy versus control conditions for different levels of political conservatism, measured on a scale of 1 (strongly liberal) to 5 (strongly conservative). As shown, the difference between the treatment and control condition decreases as participants become more conservative such that there is no effect of accuracy among strong conservatives. Shaded bands represent 95% confidence intervals. See ***Supplementary Appendix Table S6*** for more detail.

The original authors recently published a separate meta-analysis in which they claim that the accuracy nudge works for conservatives in some samples, such as in a Twitter field experiment (Pennycook & Rand, 2021). However, the data in their analysis was not publicly available at the time this paper was written, so we are not able to include it in this current analysis. We are in the process of currently testing our hypothesis in this larger dataset as part of an adversarial collaboration. Data is still not available from the authors' Twitter field experiment.

Discussion

While accuracy primes have been proposed as a scalable solution to the misinformation problem (Pennycook, McPhetres, et al., 2020; Pennycook, Epstein, et al., 2021b), our meta-analysis suggests they may have limited effectiveness for the population most likely to spread misinformation. Even if there is a small effect that can be detected for Republicans at very high statistical power (a power analysis found that 4,298 Republicans would be needed to detect the estimated effect size of 0.11 at 95% power), this effect appears to disappear at the highest levels of conservatism; see ***Figure 1***.

One limitation behind this work is that it only examines 5 datasets, and a number of datasets testing the accuracy nudge were not available for our analysis at the time this was written. Though, we are currently testing data from Pennycook & Rand (2022) to examine whether this effect replicates in a larger dataset. It is also unclear what the mechanism behind this effect is, or whether the effect depends in part on the stimuli chosen or the make-up of the samples. Future work should examine the replicability and generalizability of this asymmetry, and if it is robust, examine why conservatives are less susceptible to the accuracy nudge. However, some work has already replicated finding from this paper that accuracy nudges are less effective (or ineffective) for conservative participants (Pretus et al., 2021).

While the authors argue that their inattention-based account of misinformation sharing challenges accounts based on partisan identity (Pennycook et al., 2020, 2021), our results suggest that inattention to accuracy hinges on partisan identity or ideology. Thus, these data support broader theoretical accounts of misinformation sharing in which inattention, partisan identity, ideology, and other factors interact (Batailler et al., 2021; Gawronski, 2021; Van Bavel et al., 2021). While the authors show that accuracy primes may be useful for some groups, our findings suggest that approaches that are better able to speak to the motivations, norms, and beliefs of different groups are critical to understand and stop the spread of misinformation among those who are most likely to share it.

Chapter 5. Social Media Behavior is Associated with Vaccine Hesitancy

Introduction

Mitigating the COVID-19 pandemic and preventing future disease outbreaks requires understanding and overcoming vaccine hesitancy (Bavel et al., 2020; MacDonald, 2015). Many have expressed concern that a misinformation "infodemic" on social media platforms such as Facebook and Twitter may contribute to vaccine hesitancy (Cinelli et al., 2021; Johnson et al., 2020a; van der Linden, 2022; Zarocostas, 2020). Indeed, the US surgeon general has called vaccine misinformation on social media an "urgent threat to public health" (Stolberg & Alba, 2021) and US president Joe Biden has insinuated that platforms such as Facebook are "killing people" with vaccine misinformation (Kanno-Youngs & Kang, 2021). In the current work, we examine the potential link between social media behavior and vaccine hesitancy during a pandemic.

Past research has linked misinformation exposure to vaccine hesitancy. For instance, endorsement of COVID-19 misinformation is associated with reduced intentions to get vaccinated for COVID-19 (Romer & Jamieson, 2020; Roozenbeek et al., 2020), and exposure to COVID-19 misinformation can lead to a causal reduction in intentions to receive the vaccine (Loomba et al., 2021b). This is potentially deadly, as anti-vaccination viewpoints have been growing steadily on social media platforms such as Facebook (Johnson et al., 2020a). Indeed, one survey found that people who get their news primarily from Facebook were more vaccine-hesitant than a number of other groups – including those who get their news primarily from Fox News (Lazer et al., 2021). Altogether, these results indicate that exposure to misinformation on social media may have detrimental effects for vaccine uptake.

Other research has found that attitudes about the vaccine and COVID-19 have been strongly politicized, particularly in the United States. US conservatives report higher levels of vaccine hesitancy (Fridman et al., 2021; Roozenbeek et al., 2020), and the right-leaning media in the United States have disproportionately shared misinformation about COVID-19 (Motta et al., 2020). An analysis of mobility data for 15 million Americans found that voting for Trump and watching Fox News were two of the biggest predictors of not complying with social distancing regulations during the pandemic (Gollwitzer et al., 2020) Additionally, exposure to cues from

party elites (e.g., Trump or Biden promoting the vaccine) can causally influence vaccination intentions (S. Pink et al., 2021). However, vaccination attitudes and COVID-19 prevention behaviors have not been strongly related to conservatism in most other countries (Freeman et al., 2020; Pennycook, McPhetres, et al., 2021, p. 19; Roozenbeek et al., 2020; Van Bavel et al., 2022), indicating messages from political elites (Cohen, 2003), rather than conservative ideology on its own, may have played a unique role in politicizing attitudes about the vaccine.

Social media also tends to reflect "echo chambers" in which people are selectively exposed to like-minded opinions (Barberá et al., 2015; Cinelli et al., 2021) and form social ties with likeminded others (Mosleh, Martel, et al., 2021a). Though it should be noted that homophily – or seeking out likeminded others – is present on domains outside of social media (McPherson et al., 2001a); for instance, partisans also sort into neighborhoods with co-partisans (B. Bishop, 2009; Brown & Enos, 2021). In fact, there is debate about how strong "echo chambers" are on social media (A. Guess et al., 2018). Just as people might exist in political echo chambers online and offline, it is possible that people with vaccine hesitant attitudes also congregate in echo chambers, hearing views only from people with similar beliefs. If this is true, it could undermine public health efforts that try to encourage vaccine uptake, since people who are part of anti-vaccine "echo chambers" may not be exposed to accurate information about the vaccine or efforts to correct vaccine misinformation, for example via fact-checks (Zollo et al., 2017).

While it is important to understand the role of social media in shaping vaccine beliefs, most prior research has examined the predictors of vaccine hesitancy using either survey data or social media data on their own. To better understand how online behavior is related to vaccine attitudes beyond self-reported variables, we combined survey data with social media data. This allows us to have a more precise examination of how real-world social media behavior is associated with beliefs about vaccination.

Overview

To understand how social media behavior is related to vaccine hesitancy, we collected two samples of survey data about vaccine attitudes linked to Twitter data. Based on the prior literature, we tested four pre-registered hypotheses (see https://aspredicted.org/blind.php?x=c2jx6q for the preregistration):

H1: The number of conservative politicians one follows will be negatively associated with vaccine confidence.
H2: The number of hyper-partisan/low-quality news sites one follows will be negatively associated with vaccine confidence.
H3: People with high and low levels of vaccine confidence will cluster into online "echo chambers."
H4: People with lower vaccine confidence will share more hyper-partisan and low-quality news articles.

To test these hypotheses, we conducted two studies where we collected survey data linked to Twitter. In Study 1 (n = 464 Twitter handles), we collected a roughly politically-balanced sample of liberals and conservatives from the UK and the US along with a sample of participants who specifically reported being vaccine hesitant. In this sample, we used regression models to examine whether the number of conservative politicians and hyper-partisan websites one follows predicted vaccine hesitancy. Standardized beta coefficients are reported for all regression models for ease of interpretation, and regression models were all run with and without demographic control variables. Then, we conducted network analysis to examine whether vaccine hesitant and vaccine-confident participants clustered into "echo chambers" in the US and the UK. In Study 2, we recruited a convenience sample (n = 1600) of participants via a web app called "Have I Shared Fake News." Using this larger sample, we tested whether vaccine hesitancy predicted sharing and engaging with lower-quality information on social media in regression models.

Study 1

For Study 1, we collected a total sample of 1,246 participants via the survey platform Prolific Academic from May 11, 2021 to June 29, 2021. To recruit a large enough sample of vaccine-hesitant participants, as well as politically diverse participants, we used the survey platform's pre-screening options to oversample vaccine-hesitant and vaccine-neutral participants. We also aimed for a roughly equal number of participants from the US and the UK. See ***Materials and Methods*** for details about study recruitment.

Participants completed a two-item measure of COVID-19 vaccine confidence asking, on a scale of 1-7 (from 1 = "strongly disagree" to 7 = "strongly agree") whether "the currently

available COVID-19 vaccines are...” 1) effective and 2) safe (α = 0.97, *M* = 5.35, *SD* = 3.29). Participants completed a one-item measure asking if they intend to receive the COVID-19 vaccine, and a one-item measure indicating their political ideology on a scale of 1-7 (from 1 = “very liberal/very left-wing” to 7 = “very conservative/very right-wing”) (*M* = 3.97, *SD* = 1.95). Participants also completed a measure indicating whether they had or intended to receive the COVID-19 vaccine (896 yes, 349 no, 35 missing) and a number of demographic questions. See ***Supplementary Materials Section S1*** for full question wording.

587 participants voluntarily provided their Twitter handles, of which we were able to scrape 464 follower networks for analysis (175M, 207F, 6 Transgender/Non-Binary/Other, 73 Missing; M_{age} = 37.7; SD_{age} = 12.5). 157 handles were from participants who reported being from the United States, and 223 handles were from participants who reported being from the United Kingdom (the other 81 participants provided no answer or reported other countries). 118 participants reported that they did not intend to get the COVID-19 vaccine, while 342 reported that they intended to get the vaccine. See ***Supplementary Materials Table S1*** for details about demographics across all samples.

Following Behavior and Vaccine Confidence. We first tested whether following conservative politicians was negatively associated with vaccine confidence (**H1**). We found that the number of US Republican politicians an individual followed on Twitter (from a list of the Twitter handles of 331 US Republicans adapted from (Rathje, Van Bavel, & van der Linden, 2021b)) negatively predicted confidence in the COVID-19 vaccine, β = -0.12, 95% CI = [-0.21, -0.03], p = 0.011. Interestingly, this pattern still held in a multiple regression adjusting for self-reported political ideology, age, gender, education (e.g., having a Bachelor’s degree), number of Twitter followers, and number of accounts followed, β = -0.18, 95% CI = [-0.30, -0.05], p = 0.006 (See ***Supplementary Appendix Table S2*** for full models). However, the number of accounts followed by individuals from a list of UK Conservative Party politicians did not predict vaccine confidence, β = 0.06, 95% CI = [-0.04, 0.15], p = 0.230 (with control variables, β = 0.03, 95% CI = [-0.10, 0.16], p = 0.663. In sum, following US Republican politicians (but not UK conservative politicians) predicts vaccine hesitancy – even after adjusting for several covariates. These results are in support of **H1** in the US, but not the UK. We did not specifically pre-register predictions regarding differences in the US and the UK, but this observation is consistent with

other research on differences in polarization about vaccination and COVID-19 in the US and the UK (Freeman et al., 2020; Roozenbeek et al., 2020; Van Bavel et al., 2022).

We then tested whether following Twitter accounts associated with "hyper-partisan" websites negatively predicted vaccine confidence (**H2**). "Hyper-partisan" websites refer to websites that are rated as low-quality by independent fact-checkers (Pennycook & Rand, 2019) and often share highly partisan (though not always false) content (e.g., "Breitbart"). These websites tend to be much more common than "fake news" websites that share completely fabricated content (A. Guess et al., 2019). The number of hyper-partisan Twitter accounts a participant followed (out of a list of 32 hyper-partisan Twitter handles adapted from (Pennycook & Rand, 2019)) also negatively predicted vaccine confidence, β = -0.15, 95% CI = [-0.24, -0.06], p = 0.002. This result once again held even after adjusting for self-reported ideology, age, and gender, β = -0.20, 95% CI = [-0.32, -0.08], p = 0.002.

As a robustness check, we ran the same analysis using a larger list of 516 Twitter handles of news sites that were rated as untrustworthy by *NewsGuard* (Bhadani et al., 2022; Lapowski, 2018), which has a team of journalists rate the quality of news websites on a scale of 1-100 (low quality websites have a rating below 60). This broader list of news sites did not necessarily contain only hyper-partisan news, but also celebrity gossip sites (e.g., "TMZ"), alternative health sites, and more, as well as non-English and non-US-based sites. Once again, the number of Twitter handles of untrustworthy news sites one followed negatively predicted self-reported vaccine confidence, β = -0.19, 95% CI = [-0.28, -0.10], $p < 0.001$, including when adjusting for covariates, β = -0.19, 95% CI = [-0.31, -0.06], p = 0.003. Thus, following low-quality or hyper-partisan news sources predicts vaccine hesitancy over and above ideology alone, in support of **H2**. See ***Supplementary Appendix Table S3-S4*** for additional robustness checks.

Network Analysis. To test whether low and high vaccine-confident individuals would cluster into "echo chambers" (**H3**) we conducted social network analysis. Specifically, we investigated whether participants and the influencers they followed clustered into structurally separate communities (or "echo chambers") based on their beliefs about politics and the vaccine. In ***Figure 2***, we visualized the Twitter networks of left-wing and right-wing communities in the US and the UK (***Figure 2, Panels A and B***), vaccine-hesitant and vaccine-confident communities in the US and the UK (***Figure 2, Panels C and D***), and identified structural

communities using community detection analysis in the US and UK (***Figure 2, Panels E and F***). Detailed methods are in the ***Methods and Materials*** section.

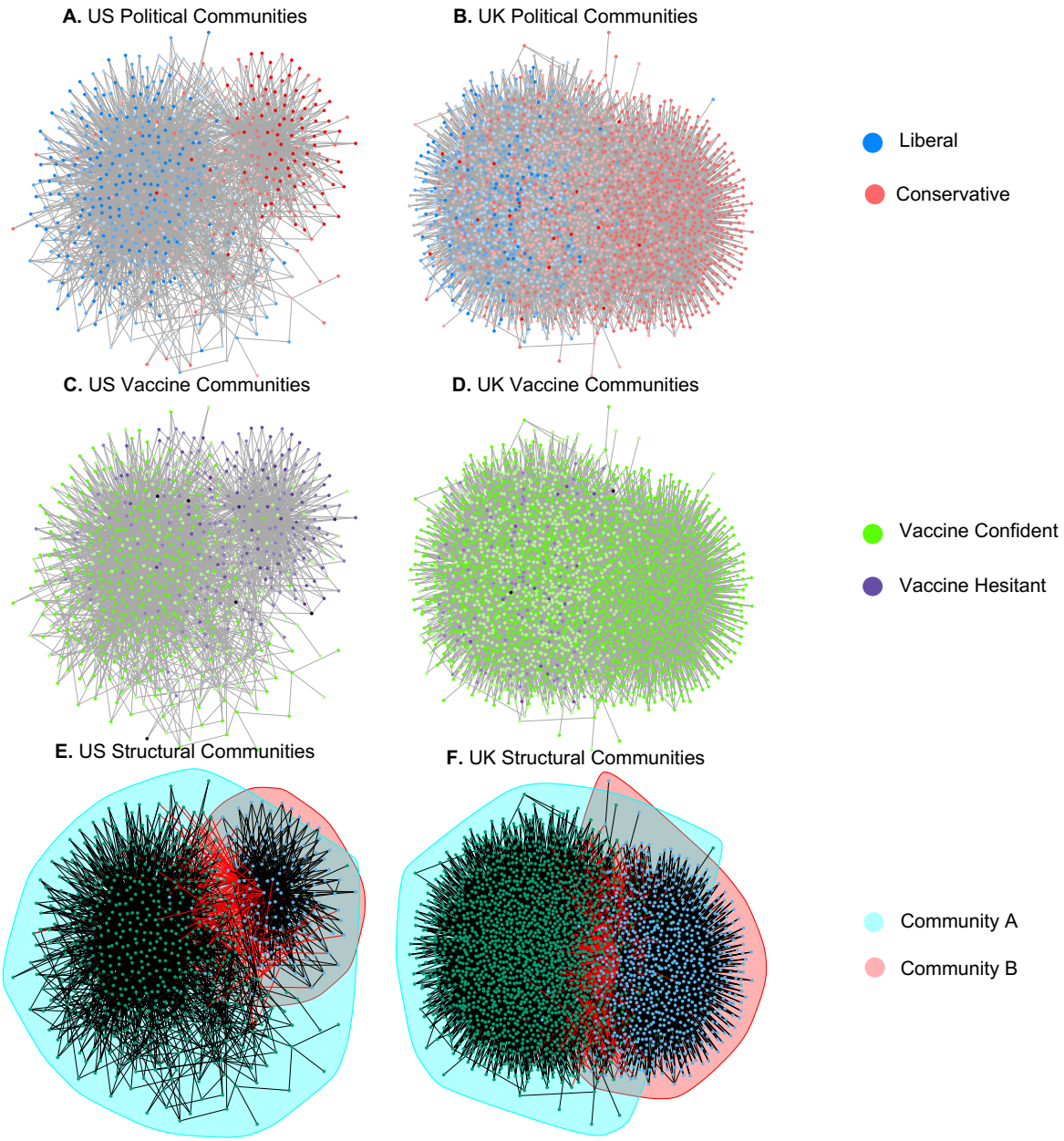


Figure 1. Visualizations of Twitter networks in the US (top) and UK (bottom). The first row shows the networks of those who are more liberal/left-wing (blue nodes) versus those who are more conservative/right-wing (red nodes) in the US **(A)** and UK **(B)**. The second row shows the networks of vaccine confident individuals (green nodes) and vaccine hesitant individuals (purple nodes) in the US **(C)** and UK **(D)**. In the third row, borders are drawn around structural

communities identified by a label-propagation graph partitioning algorithm in the US **(E)** and UK **(F)**. Each node represents either 1) an influencer that at least 3 of the participants were following, or 2) a participant that is following an influencer. Each edge between two nodes represents a following relationship (one person following another Twitter account). Layouts of the graphs were created using the large-graph-layout algorithm to visually highlight community structures. Absolute distances between nodes are not meaningful in these visualizations.

We then performed community detection analysis (Raghavan et al., 2007) which identified two distinct structural communities (Community A and Community B), see ***Figure 1, Panels E and F***. We calculated the average political conservatism and vaccine confidence of participants in each community. In the US, the average political conservatism of participants in community B (6.07, 95% CI = [5.80, 6.35]) was far higher than that of participants in community A (3.43, 95% CI = [3.00, 3.86]), $t(91.64) = 10.51$, $d = 2.20$, $p < .001$. Additionally, in the US, the average vaccine confidence of participants in community A (5.98, 95% CI = [5.72, 6.24]) was also much higher than that of participants in community B (4.46, 95% CI = [3.45, 5.47]), $t(15.05) = 3.12$, $d = 1.61$, $p = 0.007$. In the UK, the average political conservatism of participants in community B (5.09, 95% CI = [4.33, 5.85]) was marginally higher than the political conservatism of community A (4.10, 95% CI = [4.10, 4.72]), $t(14.62) = 1.80$, $d = 0.94$, $p = 0.092$. However, the average vaccine confidence of community B (6.50, 95% CI [6.24, 6.76]) was higher than the average vaccine confidence of community A (6.06, 95% CI [5.87, 6.25]), $t(26.52) = 2.92$, $d = 1.13$, $p = 0.007$). In other words, in the US, participants in community B were more conservative and less vaccine-confident than participants in community A. In the UK, participants in community B were marginally more conservative and significantly more vaccine-confident than participants in community A.

Next, we investigated whether centrality (Bolland, 1988) within the "liberal" community (community A) and the "conservative" community (community B) correlated with vaccine confidence. In the US, centrality within the "liberal" community was not significantly correlated with vaccine confidence ($r = 0.16$, 95% CI [-0.03, 0.33], $p = 0.402$), whereas centrality within the "conservative" community was negatively correlated with vaccine confidence ($r = -0.22$, 95% CI [-0.39, -0.03], $p = 0.024$). In the UK, however, centrality within the liberal community was not correlated with vaccine confidence ($r = 0.07$, 95% CI [-0.12, 0.24], $p = 0.480$), nor was

centrality within the conservative community (r = 0.15, 95% CI [-0.04, 0.32], p = 0.116). Thus, we find evidence that one's level of connectedness within the conservative community is negatively associated with lower vaccine confidence in the US, but the patterns are less clear in the UK.

We also examined whether structural polarization in the network was related to belief polarization about politics and the vaccine. Specifically, we examined whether the structural distance between two nodes (e.g., the length of the shortest path between two nodes) was correlated with the attitudinal distance between two nodes (e.g., the difference between two nodes' levels of political conservatism and vaccine confidence, see (Baldassarri & Gelman, 2008)). In the US, we found that that structural distance had a small correlation with attitudinal distance about politics (r = 0.08, 95% CI = [0.06, 0.11], p < .001) as well as vaccine confidence (r = 0.05, 95% CI [0.02, 0.07], p < .001). In the UK, however, structural distance was only marginally with attitudinal distance about politics (r = 0.02, 95% CI [-0.00, 0.05], p = 0.068), but had a small but significant positive correlation with attitudinal distance about vaccine confidence (r = 0.04, 95% CI = [0.01, 0.06], p = 0.002). In other words, nodes that were structurally separate from each other in the network also had separate attitudes about the vaccine and politics (in the US), and about the vaccine, but marginally about politics (in the UK). Overall, these results support **H3** in the US, but not in the UK.

We also reported a number of other network statistics in ***Supplementary Appendix Table S5***, including average pathlength (Albert & Barabási, 2002), modularity (Brandes et al., 2007), and assortativity (Newman, 2003). Broadly, these additional network statistics support the idea that the US network was less connected and more modularized (e.g., "polarized") than the UK network. We also performed robustness checks in ***Supplementary Appendix Section S2*** showing that different exclusion criteria do not substantially change the network topology for the scope of our analysis.

Specific Influencers Associated with Vaccine Confidence. To have a more granular picture of the kinds of Twitter "influencers" our participants followed, we explored some of the specific influencers in each community who had followers who were highest and lowest in vaccine hesitancy. For this analysis, we looked at Twitter influencers that were followed by at

least 10 people from sample 1, and calculated their followers' average vaccine confidence, as well as the proportion of followers in our sample who intended to or had received the vaccine.

The top 15 influencers associated with the highest and lowest vaccine confidence in the US and the UK, along with their membership in each community, are shown in ***Table 2*** and plotted visually in ***Figure 2***. In the US, right-leaning media personalities (e.g., Candace Owens, Ben Shapiro), Republican Party politicians (e.g., Senator Rand Paul), hyper-partisan news sources (e.g., Breaking 911), and a popular podcast host known for expressing vaccine hesitancy (Joe Rogan) (Thompson, 2021), were among the top accounts associated with low vaccine confidence. By contrast, liberal/Democratic Party politicians (e.g., former Secretary of State Hillary Clinton, Congresswoman Alexandria Ocasio-Cortez) and left-leaning media sources (The Washington Post) were associated with high vaccine confidence. In the UK, on the other hand, vaccine confidence did not appear to be as politicized, and no clear patterns emerged. In the US, most of the top influencers associated with low vaccine confidence were primarily in the "conservative" community (community B), whereas most of the top influencers associated with high vaccine confidence were in the "liberal" community (community A). See ***Supplementary Appendix Tables S9 and S10*** for robustness checks of this analysis using different thresholds.

Table 2. Twitter Influencers Associated with High and Low Vaccine Confidence Among Their Followers in the United States and United Kingdom.

United States				United Kingdom			
Twitter Handle	Vaccine Confidence	% Getting Vaccine	Community Membership	Twitter Handle	Vaccine Confidence	% Getting Vaccine	Community Membership
Influencers Associated with Low Vaccine Confidence							
RealCandaceO	3.15 (2.27)	30.00 (48.30)	B	Charlottegshore	4.10 (1.07)	40.00 (51.64)	B
joerogan	3.32 (2.11)	27.27 (46.71)	A	selenagomez	4.18 (1.37)	63.64 (50.45)	B
kayleighmcenany	3.55 (1.74)	30.00 (48.30)	B	example	4.25 (1.16)	80.00 (42.16)	A
TheBabylonBee	3.73 (2.34)	27.27 (46.71)	B	coldplay	4.36 (1.42)	72.73 (46.71)	A
dbongino	3.83 (1.91)	33.33 (49.24)	B	TheXFactor	4.40 (1.66)	40.00 (51.64)	A
benshapiro	3.92 (2.43)	50.00 (52.22)	B	LanaDelRey	4.45 (1.21)	72.73 (46.71)	B
RandPaul	3.95 (2.19)	40.00 (51.64)	B	RockstarGames	4.45 (1.47)	27.27 (46.71)	A
DonaldJTrumpJr	3.96 (1.80)	38.46 (50.64)	B	lilyallen	4.50 (1.60)	70.00 (48.30)	B
TuckerCarlson	4.00 (1.97)	42.86 (51.36)	B	JessieJ	4.57 (1.35)	60.00 (50.71)	A
seanhannity	4.04 (1.89)	38.46 (50.64)	B	NicoleScherzy	4.57 (1.40)	57.14 (51.36)	A
Jim_Jordan	4.05 (2.05)	45.45 (52.22)	B	Adele	4.57 (1.51)	73.33 (45.77)	A
JudgeJeanine	4.05 (2.05)	45.45 (52.22)	B	kourtneykardash	4.59 (1.20)	54.55 (52.22)	A
PressSec45	4.08 (2.22)	50.00 (52.22)	B	Drake	4.61 (1.33)	50.00 (51.89)	B
marklevinshow	4.15 (1.86)	40.00 (51.64)	B	katyperry	4.62 (1.40)	71.43 (46.29)	A
Breaking911	4.20 (2.15)	50.00 (52.70)	A	rihanna	4.63 (1.40)	69.57 (47.05)	A
Influencers Associated with High Vaccine Confidence							
VP	6.62 (0.48)	100.00 (0.00)	A	sarapascoe	6.42 (0.42)	100.00 (0.00)	A
HillaryClinton	6.57 (0.65)	92.86 (26.73)	A	StephenMangan	6.36 (0.60)	100.00 (0.00)	A
MichelleObama	6.46 (0.54)	91.67 (28.87)	A	BarristerSecret	6.36 (0.55)	100.00 (0.00)	A
WhiteHouse	6.42 (0.97)	83.33 (38.92)	A	Misskeeleyhawes	6.35 (0.71)	100.00 (0.00)	B
ewarren	6.36 (0.74)	90.91 (30.15)	A	mrjamesob	6.35 (0.58)	100.00 (0.00)	A
KamalaHarris	6.35 (0.97)	92.31 (27.74)	A	Number10cat	6.33 (0.78)	100.00 (0.00)	A
AOC	6.31 (0.98)	90.48 (30.08)	A	Dawn_French	6.29 (0.72)	100.00 (0.00)	A
dog_feelings	6.23 (1.17)	81.82 (40.45)	A	richardosman	6.24 (0.77)	96.00 (20.00)	A
TheOnion	6.23 (0.73)	84.62 (37.55)	A	BritishBakeOff	6.23 (0.68)	90.91 (30.15)	A
washingtonpost	6.20 (1.01)	80.00 (42.16)	A	joelycett	6.22 (0.66)	100.00 (0.00)	A
SenSanders	6.17 (1.64)	83.33 (38.92)	A	NASAPersevere	6.21 (0.86)	100.00 (0.00)	A
POTUS	6.08 (1.51)	90.00 (30.78)	A	neiltyson	6.21 (0.80)	100.00 (0.00)	A
dog_rates	6.05 (1.75)	81.82 (40.45)	A	JohnLewisRetail	6.19 (0.97)	92.31 (27.74)	A
BarackObama	5.96 (1.37)	85.71 (35.63)	A	BootstrapCook	6.19 (0.93)	100.00 (0.00)	A

ActuallyNPH	5.91 (1.71)	81.82 (40.45)	A	RobertDyas	6.18 (1.10)	81.82 (40.45)	B

Note. The Twitter accounts associated with the top 15 highest and top 15 lowest mean vaccine confidence scores among their followers (1 = low confidence, 7 = high confidence) in the US and the UK are shown above, along with the percentage of followers who are or intend to get vaccinated. Standard deviations are shown in parentheses. Additionally, each influencer's community membership (generated via community detection analysis) is shown. Individuals with high self-reported vaccine confidence among their followers tend to reside in the "liberal" community (Community A), whereas individuals low in self-reported vaccine-confidence among their followers tend to reside in the "conservative" community (Community B). Only influencers who were followed by at least 10 participants in our datasets are shown in the above analysis. 306 influencers were followed by at least 10 people in the US, and 492 influencers were followed by at least 10 people in the UK.

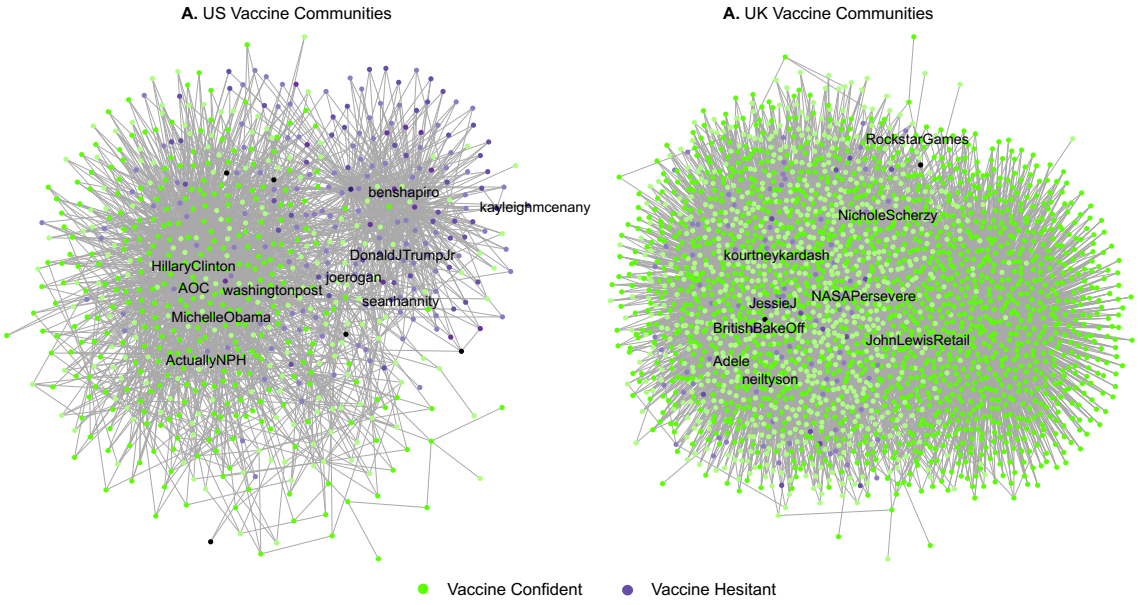


Figure 2. The same network plot from Figure 1 with example nodes (from Table 1) labeled. As shown above, more conservative influencers (e.g., Ben Shapiro, Sean Hannity) had followers who were lower in vaccine confidence (shaded purple), whereas more liberal influencers (e.g., Hillary Clinton, Alexandria Ocasio-Cortez) had followers who were higher in vaccine confidence (shaded green) in the US (Panel A). These patterns were less clear in the UK (Panel B).

Survey Data. The politicization of vaccine attitudes in the US, but not the UK, was also seen in our survey data alone. Looking just at the survey data (using the full sample without Twitter handles) we found that self-reported political conservatism was negatively associated with vaccine confidence, r = -0.33, 95% CI = [-0.39, -0.26], $p < 0.001$. This relationship was present in both the United States dataset, r = -0.43, 95% CI [-0.52, -0.35], $p < .001$, and, albeit weaker, in the United Kingdom dataset, r = -0.13, 95% CI [-0.24, -0.02], $p = .026$. The relationship between political conservatism and vaccine hesitancy was moderated by country (UK vs. US), β = -0.33, 95% CI = [-0.40, -0.25], $p < 0.001$, illustrating that vaccine confidence was more politically polarized in the US than the UK.

Study 2

The aim of Study 2 was to expand on the findings of Study 1 by examining how self-reported vaccine hesitancy was associated with sharing and interacting with low-quality information in a larger sample (**H4**). We recruited a convenience sample of participants who had used the web application "Have I Shared Fake News[5]." Data collection for this app started in April 2021, and ended October 2021 (for the purposes of this analysis), with most participants using the app in May and June of 2021 (see ***Supplementary Appendix Figure S1*** for a full timeline). At the time of analysis, 6,727 people used the app and 2,359 provided Twitter handles. After excluding participants who followed more than 50,000 people (as these Twitter accounts were likely people entering the handles of public figures) and people who did not answer the vaccine likelihood question, we were left with a total sample size of 1,600 participants (749M, 619F, 33 Non-Binary/Transgender/Other, 199 No Response, M_{age} = 38.4, SD_{age} = 12.6). Since we also invited Study 1 participants to use this app, 195 participants in this dataset were overlapping with the Study 1 participants. Location data was not collected via this app, meaning we could not explore differences between countries.

When using this app, participants who consented to take part in research were asked "How likely are you to get vaccinated for COVID-19 when it becomes available?" on a 1-100 scale (0 = very unlikely and 100 = very likely) (M = 93.36, SD = 21.55). 242 participants

[5] A link to the current version of the app is here: https://newsfeedback.shinyapps.io/HaveISharedFakeNews/.

reported a vaccine likelihood score of less than 100, and 93 participants reported a score of less than 50. People also entered demographic information in exchange for information about their news sharing behavior on Twitter. For example, people entered their political orientation on a 7-point scale (1 = "Extremely Liberal"; 7 = "Extremely Conservative") (M = 2.64, SD = 1.50). Since this was a convenience sample, it was more left-leaning and contained more vaccine-confident participants. Thus, while it was not ideal for visualizing network plots (because the "liberal" and "vaccine-confident" networks would be quite large, and many users of the app followed each other), it provided a larger sample of participants to examine associations between interacting with online misinformation and vaccine hesitancy.

Engagement with News on Social Media and Vaccine Confidence. We tested whether one's self-reported likelihood of receiving the vaccine predicted sharing or interacting with lower-quality information online. To do this, we first examined whether one's likelihood of receiving the COVID-19 vaccine was associated with the number of hyper-partisan URLs one shared on their Twitter timeline, based on a prior list of hyper-partisan URLs (Pennycook & Rand, 2019) and the "Iffy News" Index (Resnick et al., 2018). This variable was collected at the time participants used the app and shown to participants as part of their "fake news score." A total of 1030 hyper-partisan websites were shared in the full sample, and people shared about 0.77 (SD = 5.34) hyper-partisan news URLs on their Twitter timeline on average. One's likelihood of receiving the vaccine negatively predicted the sharing of hyper-partisan news sites, β = -0.07, 95% CI = [-0.12, -0.02], p = 0.007. This effect remained significant when controlling for a number of other factors measured in the app, such as political conservatism, affective polarization (favorability toward the ingroup minus favorability toward the outgroup), conspiracy mentality, mental health, age, gender, number of followers, and number of accounts followed, β = -0.14, 95% CI = [-0.20, -0.07], p < 0.001. Indeed, in this multiple regression, the only remaining significant predictor of hyper-partisan news sharing was age, β = 0.21, 95% CI = [0.14, 0.29], p < 0.001, replicating prior work about age and fake news sharing (A. Guess et al., 2019) and affective polarization was a marginally significant predictor, β = 0.08, 95% CI = [0.00, 0.16], p = 0.053 (all other ps > 0.184). Overall, these results support **H4**. See ***Supplementary Appendix S11-12*** for full regression models and robustness checks.

We then replicated the above analysis using a more sensitive measure of the quality of news shared. To do this, we used *NewsGuard*, which provided us with a dataset of over 4500

news URLs along with a trustworthiness rating of each URL (as an example, *breitbart.com* has a trustworthiness rating of 49.5 out of 100). Using the Twitter API, we scraped 1,831,308 tweets from timelines of 1600 participants using the Twitter handles participants provided when they used the app. Of these, 46,202 contained URLs that could be given a "trustworthiness" rating by *NewsGuard*. We calculated a variable indicating the average trustworthiness of URLs shared per user. The mean trustworthiness of the URLs participants shared was 92.88 (SD = 14.01), indicating that our sample tended to share trustworthy news.

Again, one's likelihood of receiving the vaccine predicted the quality of news URLs people shared online **H4**, $\beta = 0.23$, 95% CI = [0.17, 0.29], $p < 0.001$. This effect remained significant when including relevant control variables, such as political liberalism, age, and gender, $\beta = 0.19$, 95% CI = [0.09, 0.29], $p < 0.001$. The only other significant predictor of vaccine confidence in this model was having a Bachelor's degree, $\beta = 0.48$, 95% CI = [0.22, 0.74], $p < 0.001$ (all other $ps > 0.166$). The model is plotted in ***Figure 3*****,** see ***Supplementary Materials S11-S12*** for full regression models and robustness checks.

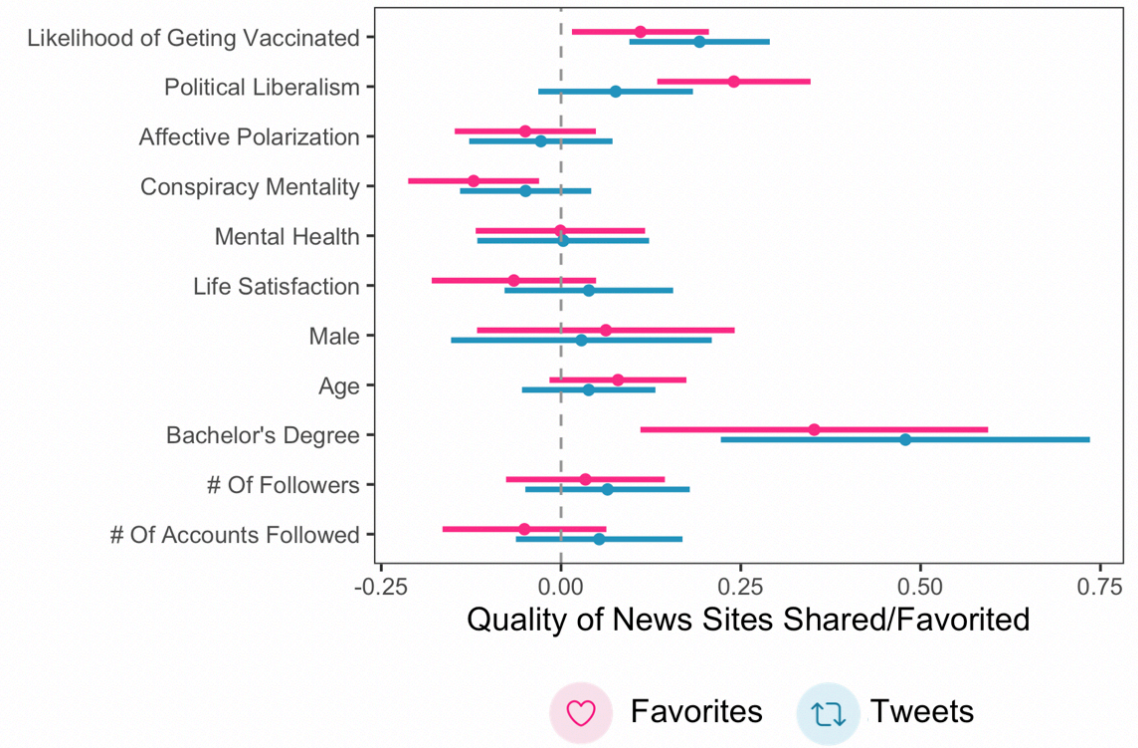


Figure 3. One's self-reported likelihood of getting the COVID-19 vaccine predicted the overall quality of news participants shared publicly (tweeted) or liked (favorited). These associations remained significant in multiple regression analyses accounting for political liberalism, affective polarization, conspiracy mentality, mental health, life satisfaction, age, gender, education, number of followers, and number of accounts followed. Error bars represent 95% confidence intervals, and the standardized beta coefficient is shown for ease of interpretation.

Finally, given that different types of content on social media tend to receive retweets as opposed to favorites (or likes) (28) – possibly because retweets are more public than favorites – we then examined whether one's likelihood of receiving the vaccine predicted favorites as well. We scraped 1,876,635 favorites from our sample, 61,140 of which contained URLs that could be given a NewsGuard "trustworthiness" rating ($M = 93.59$, $SD = 12.84$). One's likelihood of getting vaccinated once again predicted favoriting higher-quality news, with a similar effect size, $\beta = 0.23$, 95% CI = [0.06, 0.23], $p < 0.023$. Once again, this effect remained significant when controlling for political ideology, gender, and age, $\beta = 0.11$, 95% CI = [0.02, 0.21], $p = 0.023$. The other significant predictors of favoriting low-quality news in this model were political liberalism, $\beta = 0.24$, 95% CI = [0.13, 0.35], $p < 0.001$, conspiracy mentality, $\beta = -0.12$, 95% CI =

[-0.21, -0.03], p = 0.009, and having a Bachelor's degree, $\beta = 0.35$, 95% CI = [0.11, 0.59], p = 0.004. We also found that the quality of news URLs participants shared correlated strongly with the quality of news URLs that they favorited, $r = 0.53$, 95% CI = [0.47, 0.58], $p < 0.001$. Thus, an individual's more public sharing behavior online may not be strongly different than their more private favoriting behavior. Interestingly, across all 3 models, vaccine hesitancy was a robust predictor of interacting with online misinformation, whereas other variables such as age, political ideology, conspiracy mentality, and education were more inconsistent predictors that were not significant across all three multiple regression models.

Specific News Sites Associated with Vaccine Hesitancy. For a more granular examination of this data, we explored the specific news sites that tended to be shared by those who were less likely to receive the vaccine. To do this, we examined URLs that were shared by at least 10 people in the dataset and calculated the average likelihood of getting the vaccine among these news sharers. We also examined news sites favorited by at least 10 people in the dataset and calculated their average likelihood of getting the vaccine. Several news sites shared and favorited by those who reported being unlikely to get the vaccine (e.g., "zerohedge.com," "palmerreport.com," "breitbart.com," "rt.com," "thefederalist.com," "tmz.com") are rated as "untrustworthy" by *NewsGuard*, see ***Table 3.*** See ***Supplementary Appendix Tables S10 and S11*** for robustness checks of this analysis using different thresholds.

Table 3. Specific URLs shared or favorited associated with low self-reported likelihood of receiving the vaccine.

News Website Shared	Likelihood of Getting Vaccine	News Website Favorited	Likelihood of Getting Vaccine
billboard.com	69.08 (42.21)	billboard.com	70.07 (41.76)
zerohedge.com	72.62 (42.52)	deviantart.com	72.30 (41.00)
foxbusiness.com	73.50 (41.05)	zerohedge.com	73.94 (39.67)
washingtonexaminer.com	75.92 (35.16)	thepostmillennial.com	75.23 (38.98)
upworthy.com	79.15 (32.80)	thewrap.com	77.73 (38.93)
express.co.uk	80.69 (38.57)	breitbart.com	78.07 (37.21)
thefederalist.com	82.00 (38.24)	dailycaller.com	79.25 (38.59)
thinkprogress.org	83.46 (36.37)	rt.com	80.83 (33.67)
money.cnn.com	83.94 (32.36)	abc13.com	82.23 (37.41)
boston.com	84.10 (34.70)	webmd.com	83.18 (37.57)
tmz.com	85.38 (35.73)	tabletmag.com	84.47 (34.86)
gq.com	85.43 (32.15)	apod.nasa.gov	84.55 (33.28)
courier-journal.com	86.20 (32.56)	palmerreport.com	85.64 (26.18)
bizjournals.com	87.55 (28.10)	heraldscotland.com	86.06 (29.13)
spiegel.de	87.89 (29.13)	foxnews.com	88.21 (30.62)

Note. On the left are the URLs that are tweeted by at least 10 people along with the average likelihood of getting the vaccine (on a scale from 1-100) among those that tweeted each URL. On the right are the URLS that were favorited by at least 10 people along with the average likelihood of getting the vaccine among those that favorited each URL. Several of the news sites shown (e.g., “zerohedge.com,” “palmerreport.com,” “breitbart.com,” “rt.com,” “thefederalist.com,” “tmz.com”) received low trustworthiness ratings by NewsGuard.

Discussion

Across two studies with unique datasets connecting survey data about self-reported vaccine confidence to social media data, we found that social media behavior is associated with attitudes about the vaccine. Specifically, following US Republican Twitter influencers and hyper-partisan or low-quality news sites negatively predicted confidence in the COVID-19 vaccine, though following politicians from the UK's Conservative party did not predict confidence in the vaccine. These results held even when controlling for a number of relevant variables, such as self-reported ideology, age, and gender, meaning that social media behavior explains unique variance in predicting vaccine attitudes beyond ideology/partisan affiliation alone.

Community detection analysis revealed that Twitter networks in the US and the UK divided into communities (or "echo chambers") broadly reflecting liberal and conservative attitudes. Centrality in the more "conservative" community in the US negatively predicted self-reported vaccine confidence; however, this was not true in the UK. We also found that structural polarization in the network modestly correlated with belief polarization about the vaccine both in the US and the UK. A specific examination of the influencers in each cluster found that prominent US influencers associated with the Republican party (e.g., Tucker Carlson, Candace Owens), as well as influencers who have caused controversy about spreading misinformation about the vaccine online (e.g. Joe Rogan) (Thompson, 2021) tended to have followers with low levels of vaccine confidence.

Finally, in Study 2, we found that one's likelihood of receiving the vaccine was associated with the quality of news articles shared (tweeted) and liked (favorited) on Twitter, even when controlling for demographic variables. This suggests that vaccine-hesitant individuals are not only consuming lower-quality news, but are spreading lower-quality news to their networks. These results were similar when looking at both more private forms of engagement (favorites) and more public forms of social media sharing (retweeting), which were highly correlated with each other.

One limitation of this work is that it captures a specific time-point in history. Most of the data were collected during the summer of 2021, and dynamics around these issues online and offline may have evolved. Furthermore, there are limitations with our samples. Neither study was nationally representative, though Study 1 was roughly politically-balanced and included a large

portion of vaccine-hesitant and vaccine-neutral respondents. Study 2, while larger and better-powered to analyze the amount of misinformation shared by individuals, was a convenience sample recruited online via an app. It is possible that some of our findings were influenced by idiosyncrasies of these two samples. That said, it's important to note that the main conclusions were consistent across both samples.

An important limitation of this work is that it is correlational. While our results are consistent with the theory that exposure to misinformation and partisan cues in one's online social network influences vaccine attitudes, they are also consistent with other interpretations, such as vaccine-hesitant individuals selectively following and engaging with content that confirms their beliefs (Aral et al., 2009b; Mosleh, Martel, et al., 2021a). Twitter following behavior could also be a proxy for other kinds of media exposure (for instance, people who follow Republican politicians may also frequently watch Fox News). Other research should follow up on this study by testing the causal effects of exposure to certain information sources on vaccine attitudes, through lab and field experiments, network interventions that manipulate the structure of one's network (Valente, 2012), or network modeling approaches (L. da F. Costa et al., 2011).

Many of the effect sizes we found were small-to-medium (e.g., $r = 0.23$ for the correlation between quality of news shared and likelihood of getting the vaccine) (Funder & Ozer, 2019; Lovakov & Agadullina, n.d.), though other effect sizes were large, such as the difference in vaccine confidence between participants in the liberal community (community A) and the conservative community (community B) in the network. Given that almost four billion people use social media worldwide (Statista, 2022b), even small associations between exposure to certain types of online content and vaccine beliefs are practically significant.

There are also multiple possible reasons for differences between the UK and US samples. For instance, they may reflect differences in conservatism between the US and the UK. It has been noted that the UK conservatives are generally less conservative than the US, or that UK conservatism may reflect different priorities and values, such as traditionalism (Gest et al., 2018). Though, another interpretation behind the differences we found in the US and the UK is that partisan elite cues early in the pandemic guided polarization around the vaccine, and certain political figures, such as Donald Trump or Conservative Prime Minister Boris Johnson, played an important role in driving opinions about COVID-19 early on. Indeed, Conservative Prime

Minister Boris Johnson called anti-vaxxers "nuts" in 2020 (Walker & correspondent, 2020). By contrast, one study from 2020 estimated that Donald Trump was the largest source of COVID-19 misinformation at the time (Evanega et al., 2020). Experiments support the idea that partisan elite cues play a causal role in sharing opinions about the virus (Flores et al., 2022; S. Pink et al., 2021).

Our results demonstrate potential challenges of promoting vaccine confidence in a polarized social media environment (Van Bavel, Harris, et al., 2021; Van Bavel, Rathje, et al., 2021a; van der Linden, Roozenbeek, et al., 2021), since accurate messages about the vaccine may not be seen by those who need it most unless they come from trusted influencers in their networks, such as influencers associated with the Republican party. Hopefully, these results will help researchers and policymakers understand and help create solutions for vaccine hesitancy. For example, targeted messages from figures trusted by people in communities associated with low vaccine confidence (Chu et al., 2021; S. Pink et al., 2021), interventions that protect against susceptibility to misinformation (A. M. Guess, Lerner, et al., 2020b; Rathje, 2022; van Der Linden et al., 2020), or algorithmic solutions that improve the overall quality of news presented to people on social media (Bhadani et al., 2022) may be useful for improving vaccine confidence. Amid frequent discussion about an "infodemic" of misinformation on social media contributing to vaccine hesitancy (Zarocostas, 2020) and controversy over prominent influencers such as Joe Rogan spreading vaccine misinformation online (Thompson, 2021), our work demonstrates the crucial link between online behavior and vaccine attitudes.

Materials and Methods

Code, surveys, materials, dictionaries, lists of URLs used, and de-identified data are available at: https://osf.io/shjdb/?view_only=60b22cb131404190856b1e68df9a0f57. We could not share all Twitter data due to privacy concerns (e.g., Twitter handles, raw Twitter texts, or raw URLs shared), though we attempted to share limited, anonymized data and code for replicating the main models and network analysis. Furthermore, lists of URLs and Twitter handles along with their "trustworthiness" ratings cannot be accessed without a license agreement from NewsGuard. NewsGuard data was retrieved on February 29, 2022 and reflects ratings as of that particular date. Data was analyzed using R version 4.0.1. The study was pre-registered at:

https://aspredicted.org/blind.php?x=c2jx6q**.** This study was approved by the University of Cambridge Psychology Research Ethics Committee (PRE.2020.144).

We deviated from our pre-registered hypotheses in a few ways. First, we said that we would examine associations between the misinformation susceptibility test (Maertens et al., 2021), life satisfaction, mental health, and Twitter behavior. We are now examining these associations in a separate publication, since we believe they are less relevant to the current examination. Second, we said that we would examine influencers who are followed by at least 25 participants, following (Mosleh, Pennycook, et al., 2021), and calculate the average vaccine attitudes of their followers. Because we had a smaller sample of vaccine-hesitant participants than anticipated, we instead used a threshold of 10. However, we show the results from this same analysis using different thresholds in ***Supplementary Appendix S8 and S9***, finding qualitatively similar results (e.g., following conservative influencers in the US seems to be associated with vaccine hesitancy across multiple thresholds).

Participants. For Study 1, we collected a total sample of 1,246 participants (465M, 556F, 15 Non-Binary/Transgender /Other, M_{Age} = 44.33) via the survey platform Prolific Academic from May 11, 2021 to June 29, 2021. To recruit a large enough sample of vaccine-hesitant participants, as well as politically diverse participants, we used Prolific pre-screening criteria to recruit a target sample size of 400 participants who reported being either hesitant or neutral about the COVID-19 vaccine. In addition to this, we recruited 200 US liberals, 200 US conservatives, 200 UK liberals, and 200 UK conservatives. This is a slight deviation from the pre-registration, where we said we would sample 300 conservative politicians and 300 liberal politicians, but did not mention anything about the country. Because we were interested in the dynamics of vaccine hesitancy in multiple countries, we decided to collect a slightly larger sample of liberals and conservatives from the US and the UK. 587 participants voluntarily provided their Twitter handles, of which we were able to scrape 464 follower networks for analysis (175M, 210F, 6 Transgender/Non-Binary/Other, 73 Missing; M_{age} = 37.7; SD = 12.5). In addition to our key measures, we asked a number of other measures as well, such as a measure of Misinformation Susceptibility (Maertens et al., 2021), mental health, life satisfaction, country, and education. We report other demographic data in ***Supplementary Appendix Table S1***.

For Study 2, we recruited a convenience sample of participants who had used the web application “Have I Shared Fake News.” We shared the web application on Twitter in May 2021,

and recruited participants up until October 2021 via snowball sampling. We also gave Study 1 participants the opportunity to use the app. While some of this dataset was collected before the pre-registration, much of it was collected afterwards as well, and it was not analyzed until after the pre-registration. See ***Supplementary Materials Figure S2*** for more information about when the dataset was collected.

Network Analysis. To investigate whether political and vaccine opinion communities are structurally separated into "echo chambers," we constructed community network graphs for the US and the UK participants and the "influencers" they follow (that are followed by at least 3 participants). Before filtering out small influencers, the dataset contained in total 50,276 following relationships from 124 participants in the US and 77,160 following relationships from 123 participants in the UK. The smaller number of participants in both countries is the result of filtering out participants who did not report political conservatism or vaccine confidence values, both of which are crucial to this network analysis. After filtering out influencers who were not followed by at least 3 participants, we constructed network graphs based on the 2,588 following relationships from 109 participants in the US, and the 11,055 following relationships from 118 participants in the UK. See ***Supplementary Appendix Section S2*** for further explanation about the different number of following relationships in the US and the UK and a robustness check of the filtering criteria for influencers. After constructing the network graphs, we calculated several descriptive statistics, including average path lengths, modularity coefficients, and assortativity coefficients based on political and vaccine opinions, which are reported in ***Supplementary Table S4***.

Then, we used a label-propagation algorithm for graph partitioning (Raghavan et al., 2007) to identify two structural communities in the US and the UK. We chose the label-propagation algorithm because it was designed for large-scale complex networks and can be performed at near linear-time, which is suitable for our network dataset and limited computational power. We calculated and compared the average political and vaccine attitudes among participants within each community in the US and the UK. We did not include influencers in this comparison since their political and vaccine attitudes are calculated from the participants' values and we did not want to double-count participants. To examine the relationship between structural properties of the network and attitude differences, we correlated the degree centrality of a node in a community with a node's political or vaccine opinion, and

also correlated each pair of nodes' path length with their belief difference about politics and the vaccine. Influencers were also excluded in these correlational analyses to avoid double-counting participants. See ***Supplementary Appendix Section S2*** for further explanation about the network analysis.

Chapter 6. How Social Media (Unfollowing) Behavior Influences Affective Polarization and Well-Being: Results from a Social Media Field Experiment

Introduction

About 4 billion people use social media, an innovation that was in its infancy a mere 15 years ago (Clemente, 2020). In a time of historic levels of polarization in the United States (Finkel et al., 2020), it is important to understand how this new technology might be shaping affective polarization. Affective polarization is characterized as disdain for out-party members and is independent from ideological polarization, which reflects different attitudes about policies (Finkel et al., 2020; Iyengar et al., 2019). Affective polarization may be linked to negative downstream consequences, such as political violence (Mernyk et al., 2021) and misinformation belief and sharing (Osmundsen et al., 2020).

Yet, researchers are divided on whether social media plays a causal role in polarization (Van Bavel, Rathje, et al., 2021b). For instance, some research suggests that the oldest individuals (who are the least likely to be on social media) tend to be the most polarized (Boxell et al., 2017). Additionally, other research shows that many countries are not showing the same increase in polarization as the United States (Boxell et al., 2017), despite similar levels of social media usage. Though, other research suggests that social media may have a causal effect on polarization (Kubin & von Sikorski, 2021). For instance, one randomized control trial found that deleting Facebook for one month decreased issue polarization and marginally decreased affective polarization (Allcott et al., 2019).

Though studies investigating the causal effect of social media on polarization have yielded conflicting results. For instance, a replication of Alcott et. al (2019) among participants in Bosnia found that those who deleted Facebook during genocide remembrance week actually showed increases ethnic polarization (Asimovic et al., 2021). A field experiment found that exposure to messages from the opposing party on Twitter increased ideological polarization (Bail et al., 2018), but another field experiment found that exposure to messages from the opposing party decreased affective polarization (Levy, 2021b).

These conflicting findings could be attributed to the fact that people's online and offline social networks differ greatly. For example, many social media and internet users appear to be

politically disengaged. Contrary to the common idea that users that are trapped in online "echo chambers," some scholars argue that Facebook users have a relatively politically diverse intake of information (Bakshy et al., 2015) and that internet users do not appear to live in political bubbles online (Eady et al., 2019; A. M. Guess, 2021). More than a third of Twitter users follow any media sources (Eady et al., 2019), and another analysis finds that around 60% of Twitter users do not follow any political elites (Wojcieszak et al., 2021). However, those who are politically engaged online appear to be very polarized and engage in echo-chamber-like behavior. For example, politically-engaged Twitter users are 14 times more likely to retweet in-group as opposed to out-group politicians, and when they do retweet about out-group politicians, these retweets are usually paired with a negative comment (Wojcieszak et al., 2021). Additionally, partisans are much more likely to follow-back other in-group Twitter accounts (Mosleh, Martel, et al., 2021b). Given the large differences in how people use social media, instead of exploring whether social media, as a whole, causes polarization, it is necessary to explore how certain specific ways of using social media causally contribute to polarization.

While politically-engaged social media users appear to be polarized, it is difficult to discern whether being a part of highly partisan online networks *causally* contributes to polarization. People organize into homophilous networks (McPherson et al., 2001b), choosing offline and online and communities of people who are similar to themselves (Brown & Enos, 2021; Mosleh et al., 2020). A recent analysis found that polarization on Reddit appeared to be driven by an influx of new conservative users joining the platform in 2016, as opposed to Reddit itself polarizing users (Waller & Anderson, 2021). Other work has found that exposure to partisan media has minimal effects (A. M. Guess et al., 2021). Thus, it is likely that people both self-select into polarizing online networks and that people are influenced by social norms and partisan cues in their own network (Aral et al., 2009b), but it is difficult to discern the causal impact of one's online social network without experimentally manipulating that network (Valente, 2012).

Overview

This study aims to examine how being embedded within "polarizing" online social networks (e.g., following highly partisan politicians and news sources) is both associated with and causally contributes to affective polarization. In the first study, which is a correlational

analysis linking Twitter data to survey data (n = 1447, after exclusions), we examined whether people within more "polarizing" online networks would show higher degrees of affective polarization, and would also engage in more polarizing online behavior (sharing more misinformation and using more toxic language). In the second study (n = 643, after exclusions), we explored whether incentivizing people leave these polarizing online networks by making them unfollow highly partisan Twitter accounts (and follow non-partisan accounts about space, science, and nature) causally decreased polarization.

Results

Study 1

The purpose of Study 1 was to examine whether following certain Twitter accounts was correlated with affective polarization, or negative feelings toward the opposing party. This study helped us select accounts for participants to follow and unfollow in the Study 2 intervention. For this study, we used data collected from an app I created called "Have I Shared Fake News."

Methods

Study 1 data came from a convenience sample of participants who used an app we created called "Have I Shared Fake News." Data collection for this app began in April 2021 and ended during December 2021 for the purpose of this analysis. This app was shared widely on Twitter, and data was primarily collected via snowball sampling from Twitter users. When using this app, participants entered their Twitter handles and received feedback about how many fake and hyper-partisan websites they shared on Twitter (https://newsfeedback.shinyapps.io/HaveISharedFakeNews/). Participants also answered a number of free-response questions with this app, such as demographic questions about age and gender and questions about their favorability toward Republicans and Democrats on a scale of 1-100. After excluding participants who did not enter their Twitter handles, did not answer questions about attitudes toward Democrats or Republicans, or had more than 50,000 followers, we were left with a total of 1447 participants (M_{age} = 39.34, SD_{age} = 12.53, 768M, 458F, 30 non-binary/transgender/other, 1011 liberals, 254 conservatives). Note, due to different exclusion criteria and a different time of Twitter retrieval, this sample is slightly different to the sample presented in *Chapter 4*, even though the data source is the same.

Analysis

Choosing Accounts for Participants to Unfollow. Using the Twitter API, we scrapped all participants' follower networks (e.g., the Twitter accounts participants were following). Then, we examined the Twitter "influencers" (users with at least 100,000 followers) that participants in our sample followed, and examined the average favorability toward Democrats and Republicans among their followers. For each "influencer" we calculated the average favorability toward Democrats and Republicans among their followers. We used this data to help select Twitter accounts associated with high levels of affective polarization that we would ask people to unfollow in Study 2. We generated a number of candidate accounts that were relevant for our intervention by looking at the influencers associated with the lowest favorability toward Democrats and Republicans among their followers in our sample. See ***Supplementary Appendix S1*** for the top accounts associated with favorability toward Democrats and Republicans followed by at least 25 participants in our sample. In addition to this data-driven approach of selecting potentially relevant accounts for our intervention, we also used manual selection to pick accounts that 1) tweeted frequently, 2) had many followers, and, if they were a media source, 3) were rated as being politically-slanted (rather than centrist or non-partisan) by sources such as "AllSides" or "MediaBiasFactCheck." This left us with a total of 30 Twitter accounts associated with low favorability toward Democrats among their followers, and 30 Twitter accounts associated with low favorability toward Republicans among their followers. These Twitter accounts, along with the average favorability toward Democrats and Republicans among their followers, are shown in ***Table 1***. As shown in ***Figure 2***, followers of these two sets of liberal and conservative-leaning influencers formed into two main clusters.

Table 1. Accounts People Were Asked To Unfollow in the Experimental Condition

Accounts Associated with Animosity Toward Democrats				Accounts Associated with Animosity Toward Republicans			
Account	**Republican Favoribility**	**Democrat Favorability**	n	**Account**	**Republican Favoribility**	**Democrat Favorability**	n
realDailyWire	61.33 (30.88)	20.22 (24.08)	9	OccupyDemocrats	4.13 (8.05)	91.20 (10.51)	15
DailyCaller	60.75 (30.11)	20.50 (28.02)	8	MaddowBlog	8.75 (11.88)	81.19 (15.05)	99
dbongino	62.13 (25.01)	25.27 (21.24)	15	RepAdamSchiff	9.59 (12.06)	79.79 (16.47)	175
kayleighmcenany	61.23 (30.75)	25.54 (24.13)	13	IlhanMN	9.60 (11.56)	72.32 (22.10)	169
OANN	49.75 (35.07)	26.33 (20.16)	12	PalmerReport	9.67 (12.42)	85.41 (14.22)	46
theblaze	50.33 (33.67)	27.78 (28.15)	9	maddow	9.85 (11.15)	79.37 (14.88)	213
DineshDSouza	58.00 (25.94)	28.75 (24.91)	12	Lawrence	10.08 (13.25)	81.47 (15.41)	95
RepMattGaetz	44.07 (31.06)	30.71 (26.94)	14	chrislhayes	10.16 (11.44)	79.31 (16.04)	184
TomiLahren	58.57 (35.67)	32.00 (28.11)	14	ewarren	10.18 (11.81)	76.78 (17.86)	346
newsmax	57.29 (28.56)	33.14 (26.02)	7	TheDemocrats	10.49 (14.12)	80.65 (13.19)	81
TheBabylonBee	42.43 (29.78)	35.06 (30.10)	35	dailykos	10.77 (14.97)	75.29 (21.39)	35
BreitbartNews	44.62 (32.98)	35.44 (28.79)	16	SenWarren	10.84 (11.41)	76.79 (18.36)	300
MrAndyNgo	36.90 (29.48)	35.81 (27.93)	31	BuzzFeedNews	10.94 (13.18)	73.60 (23.97)	50
Jim_Jordan	46.33 (33.68)	36.44 (30.36)	18	RawStory	10.95 (18.09)	80.79 (18.51)	19
DonaldJTrumpJr	43.79 (29.86)	37.91 (28.62)	33	AOC	11.28 (14.31)	73.41 (20.82)	513
DanCrenshawTX	43.63 (29.13)	39.41 (28.09)	27	MotherJones	11.43 (13.19)	76.88 (19.05)	98
RealCandaceO	44.71 (30.51)	39.89 (29.56)	28	chucktodd	12.48 (13.45)	79.86 (13.60)	50
jordanbpeterson	36.91 (26.53)	41.98 (29.09)	55	voxdotcom	12.64 (12.82)	76.00 (17.05)	129
benshapiro	37.98 (30.81)	42.38 (29.03)	53	TheDailyShow	12.71 (13.74)	77.36 (15.40)	223
TuckerCarlson	39.24 (31.36)	44.45 (29.58)	33	BernieSanders	12.83 (14.16)	72.16 (22.43)	255
megynkelly	33.14 (26.72)	46.29 (28.80)	28	MSNBC	13.00 (16.55)	81.36 (16.30)	69
RandPaul	38.43 (25.68)	46.29 (29.99)	35	jaketapper	13.54 (15.37)	77.60 (16.90)	167
tedcruz	36.12 (29.13)	46.58 (31.73)	33	thedailybeast	13.55 (15.35)	75.14 (17.79)	51
joerogan	28.36 (25.88)	47.04 (30.79)	56	thenation	13.91 (14.14)	75.62 (15.12)	58
seanhannity	40.75 (32.20)	47.57 (30.20)	28	CNN	13.94 (16.86)	77.43 (19.17)	164
GOPLeader	46.64 (34.66)	47.93 (30.76)	14	washingtonpost	14.25 (15.15)	75.03 (20.15)	245
marcorubio	39.03 (29.93)	53.45 (29.81)	29	Slate	14.58 (15.55)	76.58 (17.21)	72
GOP	30.14 (27.76)	57.21 (28.37)	28	HuffPost	14.85 (16.01)	77.75 (15.00)	99
FoxNews	30.89 (28.69)	59.45 (28.84)	47	cnnbrk	15.38 (19.39)	73.56 (22.54)	179
Mike_Pence	27.57 (25.67)	59.63 (26.71)	30	democracynow	15.91 (20.77)	68.35 (25.32)	43

Note. Above are the 30 right-leaning and 30 left-leaning accounts, as well as the mean favorability toward Republicans and Democrats among their followers. Favorability was measured on a scale of 1-100 with higher numbers indicating greater favorability. Standard deviations are shown in parentheses, and "n" indicates the number of participants in our sample who followed each account.

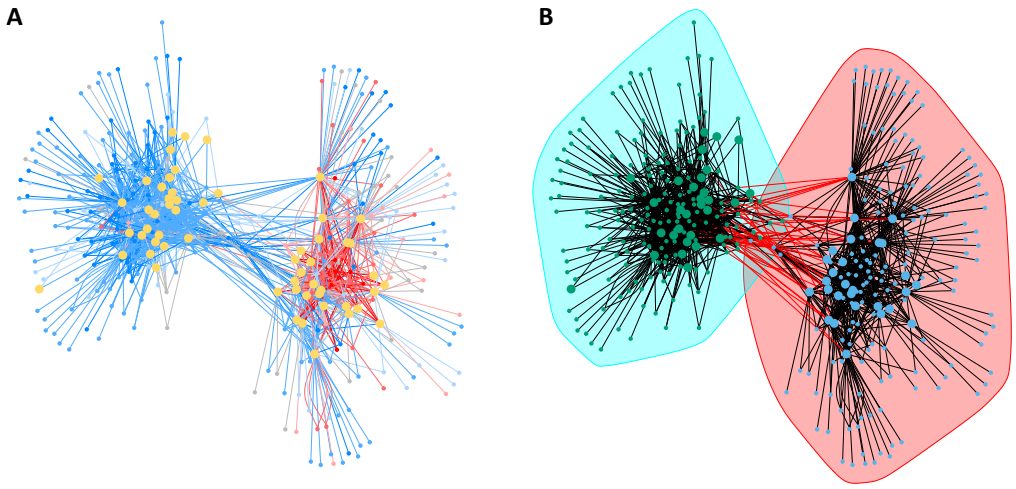


Figure 1. In **A**, a network visualization showing influencers (in yellow) and followers (blue and red nodes). Blue nodes indicate high favorability toward Republicans, and red nodes indicate high favorability toward Democrats. Since our original sample had more Democrats than Republicans, we selected a random sample of Democrats that was the same size as our sample of Republicans for this network visualization. In **B**, we present cluster analysis, which shows that liberal and conservative-leaning influencers in their followers clustered into two separate networks, or "echo chambers."

Since we in part used manual selection to select these accounts as opposed to a purely data-driven approach, we aimed to validate whether following them was indeed associated with affective polarization. We tested whether the number of accounts a participant followed from the list of 30 Democrat-leaning accounts we selected was associated with low favorability toward Republicans, and tested whether the number of accounts a participant followed from the list of 30 Republican-leaning accounts we selected was associated with low favorability toward Democrats. We found that the number of accounts a participant followed from the list of Democrat-leaning accounts was negatively correlated with favorability toward the Republican party, r = -0.18, 95% CI = [-0.23, -0.13], $p < 0.001$ and was positively correlated with favorability toward the Democratic party, r = 0.20, 95% CI = [0.16, 0.26], $p < 0.001$. The number of followers a participant followed from the list of 30 Republican-leaning accounts was negatively correlated with favorability toward the Democratic party, r = -0.25, 95% CI = [-0.29, -0.20], $p < 0.001$, and was positively correlated with favorability toward the Republican party, r = 0.24, 95% CI = [0.20, 0.29], $p < 0.001$.

These results were similar when we also looked at a list of Twitter handles of Republican and Democrat politicians used in prior research (Rathje, Van Bavel, & van der Linden, 2021c). For instance, the number of accounts followed by a user from a list of Democrat-leaning politicians negatively correlated with favorability toward the Republican party, r = -0.15, 95% CI [-0.20, -0.09], $t(1445) = -5.59$, $p < .001$, and positively correlated with favorability toward the Democratic party, r = 0.19, 95% CI [0.14, 0.24], $p < .001$. Similarly, the number of accounts followed from a list of Republican politicians negatively correlated with favorability toward the Democratic party, r = -0.07, 95% CI [-0.12, -0.02], $p = 0.006$, and positively correlated with favorability toward the Republican party, r = 0.08, 95% CI [0.03, 0.13], $p = 0.001$. Thus, these associations are robust to the lists of Twitter accounts we use.

Because we are specifically interested in affective polarization, we created an affective polarization variable by subtracting warmth toward one's out-party from warmth toward one's in-party. The number of partisan accounts one followed correlated with this affective polarization variable, r = 0.13, 95% CI [0.07, 0.19], $p < .001$.

We examined how affective polarization differed among non-political, slightly political, and very political users on Twitter. 555 participants followed 0 partisan accounts, and among these non-political users, the average affective polarization score was 50.82 (95% CI = [47.66,

53.98]). 713 participants followed 1-9 partisan accounts, and among these somewhat political users, the average affective polarization score was 58.91 (95% CI = [56.62, 61.20]). 236 participants followed 10 or more partisan accounts, and among this sample, the average affective polarization score was 67.08 (95% CI = [62.81, 71.36]). In other words, participants who followed many partisan Twitter accounts scored 16.26 points higher on a measure of affective polarization than those who followed 0 partisan accounts, $t(293.61) = 6.04$, $p < 0.001$, $d = 0.55$, and participants who followed a few partisan Twitter accounts scored 8.09 points higher on a measure of affective polarization than those who followed 0 partisan accounts, $t(841.20) = 4.07$, $p < 0.001$, $d = 0.26$. Thus, while many social media users are non-political (Wojcieszak et al., 2021), those who are politically engaged, following many partisan accounts, have higher levels of affective polarization.

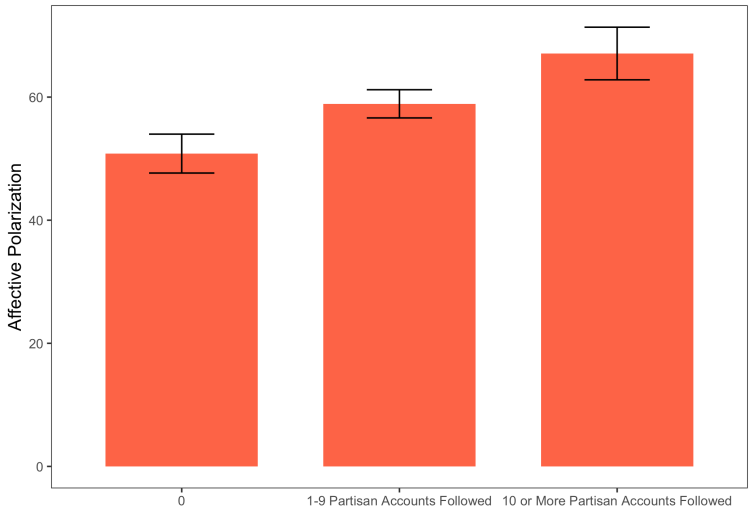


Figure 2. Affective polarization among Twitter users who either 1) followed 0 partisan accounts, 2) followed 1-9 partisan accounts, or 3) followed 10+ accounts from the list of partisan accounts we examined.

Twitter Behavioral Measures. We also merged this dataset with a dataset of the average quality of news URLs people shared as rated by NewsGuard, which rates the trustworthiness of more than 4500 News Sites on a scale of 0-100 (Lapowski, 2018). We found that the number of liberal partisan accounts an individual followed was not significantly associated with the average quality of news URLs that they shared, $r = 0.06$, 95% CI = [0, 0.12], $p = 0.096$, and the number of conservative partisan accounts an individual followed was negatively associated with the

average quality of news URLs that they shared, r = -0.17, 95% CI [-0.23, -0.10], p < .001. Thus, following the conservative partisan accounts we selected was associated with sharing low-quality news URLs.

We measured the average toxicity of tweets in this dataset using the Perspective API, which is a machine learning classifier developed by Google that rates the average "toxicity" level of Tweets[6]. "Toxic" tweets are defined as tweets that might be considered harmful, abusive, and non-constructive (e.g., "shut up you are stupid!" is 96.18% likely to be toxic according to the Google Perspective API). Because of the time it takes to run this classifier on all tweets, we only measured a subset of 486,896 of primary tweets (excluding retweets) from participants who had answered all demographic and questionnaire questions offered through the app. This left us with a total of 527 participants. The number of accounts from the list of 60 partisan accounts one followed correlated with the average toxicity of a Twitter user's tweets, r = 0.10, 95% CI [0.01, 0.18], p = 0.027.

Then, we ran a multiple regression including all predictors measured within the app, including self-reported dislike of Democrats, self-reported dislike of Republicans, the number of partisan accounts followed, conspiracy mentality, mental health, life satisfaction, likelihood of getting the COVID-19 vaccine, political conservatism, age, gender, and education. As shown in ***Figure 3***, the only significant positive predictors of toxicity were dislike of Democrats, b = 0.15, 95% CI = [0.04, 0.25], p = 0.006, dislike of Republicans, b = 0.14, 95% CI = [0.03, 0.25], p = 0.013, and the number of partisan accounts one followed, b = 0.09, 95% CI = [0.00, 0.17], p = 0.044 Having a bachelor's degree, however, was a strong negative predictor of the use of toxic language online, b = -0.41, 95% CI = [-0.64, -0.18], p < 0.001. The other variables in the multiple regression were non-significant (all ps > 0.111). Full regression results are in ***Supplementary Appendix S1***.

[6] More information about this machine-learning classifier is here: https://perspectiveapi.com/.

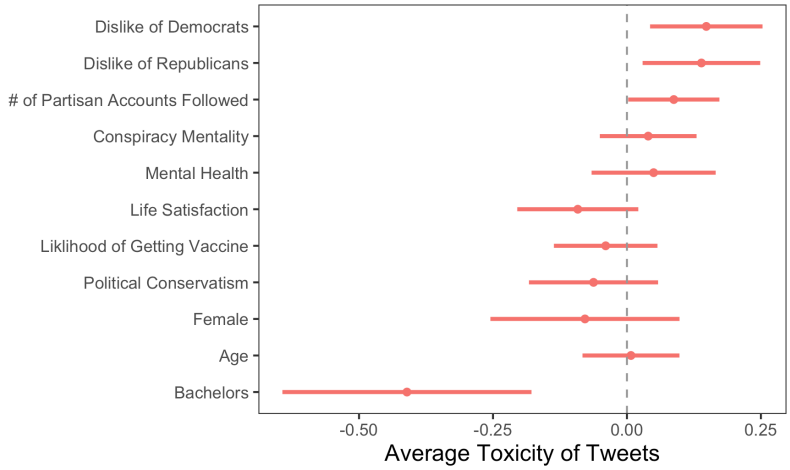


Figure 3. Dislike of democrats, dislike of Republicans, and the number of partisan accounts followed all predicted the average "toxicity" of people's tweets, as measured by the Google Perspective API machine learning classifier. Having a bachelors' degree, however, was a negative predictor of the average toxicity of one's tweets. Standardized regression results are presented for ease of interpretation, and full regression tables are shown in the ***Supplementary Appendix S1***.

Choosing Accounts for Participants to Follow. We aimed to select accounts that were not associated with polarization (and perhaps might decrease polarization) for participants to follow for the intervention. Past research has shown that content that evokes awe (Stancato & Keltner, 2021; Yaden et al., 2016), through, for instance, videos of nature, the earth from afar, or space-flight, can reduce political polarization and increase a sense of inter-connectedness. A study also demonstrated that exposure to Facebook memes that evoked a sense of common humanity decreased negative attitudes toward one's political out-group (Masullo, 2022). Furthermore, other work has shown that scientific curiosity is associated with greater receptiveness to politically-congruent ideas (Kahan et al., 2017). Thus, we thought it would be good to select Twitter accounts about science, nature, space-flight, or photography that we expected would not share political content, and would also be associated with emotions that we hoped to evoke in participants (awe, scientific curiosity, or a science of common humanity). Another purpose of having people follow these accounts was to block out partisan content from people's online social networks with relatively non-partisan content.

We selected 18 accounts for participants to follow that matched these criteria and that did not appear to have high levels of animosity toward Democrats/Republicans among their followers, see ***Table 2***. To validate that these accounts were less strongly associated with polarization than the accounts we selected for participants to unfollow, we examined the correlation between the number of accounts followed in this list and attitudes toward Democrats and Republicans. We found that the number of accounts on followed from the list of 18 accounts we asked people to follow was not significantly associated with affective polarization, r = 0.00, 95% CI = [-0.05, 0.06], p = 0.899. Following the non-partisan science accounts we selected did not correlate with the average quality of URLs shared, r = 0.00, 95% CI [-0.06, 0.07], p = 0.894. Following these non-partisan accounts was also not significantly associated with the average toxicity of tweets people shared, r = -0.05, 95% CI [-0.13, 0.04], p = 0.261.

Table 2. List of Accounts for Participants to Follow

Account	Republican Favoribility	Democrat Favorability	n
500px	43.14 (27.03)	59.71 (32.23)	14
NASAKennedy	25.61 (23.19)	73.22 (18.95)	36
bigthink	22.42 (22.69)	68.50 (25.65)	12
NASAHubble	20.00 (23.21)	66.80 (26.33)	40
PopSci	19.32 (18.44)	73.00 (18.60)	31
NASAGoddard	18.94 (19.07)	69.06 (22.66)	17
newscientist	18.84 (18.94)	75.57 (17.68)	102
NewsfromScience	16.94 (15.13)	74.39 (20.45)	31
NatGeoPhotos	16.23 (21.51)	69.77 (28.07)	30
NASA	16.07 (18.09)	71.10 (22.23)	277
wonderofscience	15.07 (20.47)	72.31 (20.82)	29
WIREDScience	14.95 (15.08)	74.85 (18.95)	55
sciam	14.93 (16.11)	76.68 (15.88)	135
Space_Station	14.89 (14.52)	70.35 (22.04)	80
NatGeo	14.30 (15.95)	75.10 (20.72)	134
NatGeoTravel	13.20 (13.97)	72.45 (22.66)	20
DiscoverMag	12.11 (11.38)	74.37 (16.09)	19
HUBBLE_space	11.33 (11.91)	66.07 (26.62)	15

Note. Above are the 18 accounts that we selected for participants to follow, as well as the mean favorability toward Republicans and Democrats among their followers. Favorability was measured on a scale of 1-100 with higher numbers indicating greater favorability. Standard deviations are shown in parentheses, and "n" indicates the number of participants in our sample who followed each account.

Choosing Placebo Accounts for Participants to Unfollow. To avoid demand effects, we also wanted to ask control group participants to unfollow accounts. Here, we asked participants to unfollow the accounts of 8 stores (such as @BestBuy or @Walmart) that we expected would not

be related to our key outcome variables, see ***Table 3***. To ensure similarly between the experimental and control group, and to provide some "distractor" accounts for participants to follow in the experimental group, we also made sure that experimental group participants unfollowed these accounts as well. Following these accounts was, indeed, not significantly associated with affective polarization, $r = 0.03$, 95% CI [-0.03, 0.09], $p = 0.349$.

Table 3. List of "Placebo" Accounts for Participants to Unfollow

Account	Republican Favoribility	Democrat Favorability	n
BestBuy	26.33 (32.72)	55.67 (20.65)	3
McDonalds	23.73 (26.15)	52.36 (27.85)	11
amazon	18.73 (24.96)	74.00 (26.86)	30
Walmart	17.29 (17.10)	77.21 (23.56)	14
Walgreens	13.50 (6.35)	70.17 (22.34)	6
Target	9.13 (9.24)	71.20 (26.36)	15
Albertsons	1.25 (2.50)	97.75 (4.50)	4

Note. Above are the 8 "distractor" accounts we asked both experimental and control condition participants to unfollow, as well as the mean favorability toward Republicans and Democrats among their followers and the number of participants in the app dataset who were following the accounts.

Study 2

Study 1 found that following certain accounts on Twitter is associated with affective polarization, and that political-engaged Twitter users who followed highly partisan accounts were much more polarized than non-politically-engaged Twitter users. Following highly partisan accounts was also associated with sharing more "toxic" tweets, and following highly partisan conservative accounts was associated with sharing lower-quality news URLs. However, it is not clear whether these effects are correlational, or whether social media following behavior has a causal influence on people's beliefs. Building off of the results of *Study 1*, *Study 2* aimed to examine the causal impact of changing one's social media following behavior on affective polarization, as well as other variables, such as well-being and perceptions of one's Twitter feed.

Procedure. This study was conducted as part of a research visit at New York University (NYU), and was approved by the NYU Institutional Review Board (Protocol #: IRB-FY2022-5783). All participants completed a Time 1 survey, in which they were asked to change their social media

following behavior for four weeks and answer a number of questions. Participants were then automatically re-contacted four weeks later to answer these questions again. All participants were told they would receive $4 for completing Study 1, $8.5 for following and unfollowing the requested Twitter accounts, $6 for completing Study 2, and $2.5 for correctly identifying tweets shown in their feed during the intervention. Thus, participants could earn up to $21 for participating in all parts of the study.

Manipulation. 60% participants were assigned to an *experimental condition*, in which they were instructed to follow and unfollow Twitter accounts that we identified in *Study 1*. Specifically, they were asked to unfollow accounts (shown in ***Table 1***) that were associated with affective polarization, and follow accounts (shown in ***Table 2***) that were not strongly associated with affective polarization. 40% of participants were assigned to a *control condition*, in which they were instructed to follow and unfollow accounts that we believed would be unrelated to these outcome variables (e.g., @Walmart, see ***Table 3***). Additionally, half of the participants in the experimental condition and half of the participants in the control condition were assigned to an *algorithm condition* in which they were asked to turn off their algorithmic Twitter feed and switch to the "latest tweets" setting. This setting shows people the tweets that were most recently posted from the accounts people follow, as opposed to the tweets the Twitter algorithm chooses to present to users. We added this condition to our study to test whether changing the Twitter algorithm settings would interact with the main outcome variables. We also pre-registered this study (pre-registration: https://aspredicted.org/blind.php?x=B9G_T2W). However, since we considered this a pilot study that we aim to replicate, much of the analysis we report here is exploratory (See ***Supplementary Appendix*** for analysis with various different exclusion criteria and robustness checks). Data, code, and Qualtrics files for this intervention are available on the Open Science Framework (OSF): https://osf.io/uaz3y/?view_only=e508b94a32944deab6530d32b820a53f.

Participant Recruitment. Using Twitter's targeted ad settings, we targeted ads toward US Twitter users between the ages of 18 and 24. We targeted this age range because this study was funded by a grant that was primarily interested in examining effects among college students. We also used Twitter's "follower look-alike" setting to recruit Twitter users who followed accounts that were similar to the accounts that we asked people to unfollow. In other words, Twitter would target ads toward people who followed accounts that were similar to the accounts we asked people to unfollow during the intervention. This maximized our chance of the intervention instructions being

relevant to the participants that we recruited. We also contacted a sample of self-reported college students we recruited on Twitter via snowball sampling.

Participant Characteristics. After removing duplicate respondents, we found that 1005 participants filled out the Time 1 survey and 795 filled out the Time 2 survey, providing us with an attrition rate of 79%. As a data quality check, we examined if people entered a valid Twitter handle using the Twitter API. Those who did not enter a valid Twitter handle were excluded from analysis, because if they did not use Twitter (or if we could not measure their Twitter usage), they were not relevant to our analysis. This left us with 643 participants to analyze for our intent-to-treat (ITT) effects (M_{age} = 28.35 (SD = 9.48), 394M, 217F, 31 Trans/Non-Binary/Other, 515 Democrats, 127 Republicans), 335 participants in our final sample were college students, whereas 307 were not.

The median daily Twitter usage in our sample on Twitter in our sample was 120 minutes (the mean daily Twitter usage was 424.58, though this appeared to be driven by extreme outliers). Participants also reported using Twitter, on average, more than other social media platforms (M = 4.01, SD = 1.10 on a 5-point likert scale from "Much Less" = 1 to "Much More" = 5), and more than other media in general (M = 3.91, SD = 1.09 on the same likert scale).

An additional 147 failed an attention check in our survey[7]. Because so many participants failed this attention check, we decided to include them in this main analysis (since the attention check may have been too difficult and we wanted to maintain high statistical power), though excluding participants who failed this check did not lead to substantially different conclusions. To prevent bot respondents, we prevented participants from continuing the survey if they incorrectly answered a bot check question "dog is to puppy as kitten is to…."

Outcome Variables. Each of the following outcome variables was asked at Time 1 and at Time 2, four weeks later:

Affective Polarization. We measured affective polarization by subtracting feelings toward one's out-party (measured on a scale of 1-100) from feelings toward one's in-party (measured on a scale

[7] The attention check was as follows: "We would like to get a sense of your general preferences. Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options. What is your favorite color?" The options were White, Pink, Green, Red, Blue, and the correct answer was selecting both Red and Green.

of 1-100) (Druckman & Levendusky, 2019). For Democrats, affective polarization was measured by subtracting warmth toward Republicans from warmth toward Democrats, and for Republicans, affective polarization was measured by subtracting warmth toward Democrats from warmth toward Republicans.

Perceptions of Twitter Feed. We created a 10-item composite score measuring feelings toward one's Twitter feed. 5 items asked whether one's Twitter feed was positive over the past four weeks, using adjectives such as "educational," "informative," and "inspiring" on a five-point likert scale from "very often" to "not often at all." Another 5 reverse-scored items asked whether one's Twitter feed was negative over the past four weeks, using adjectives such as "polarizing," "divisive," and "angry." We also asked people how often they saw fake tweets, accurate tweets, tweets about science, and tweets about politics.

Well-Being. We adapted a subjective well-being measure from Asimovic et al., (2021). We asked people how much they felt 5 positive feelings over the past four weeks (e.g., "satisfaction with life," "joy") as well as 5 negative feelings (e.g., "depression," "anxiety") on a 5-point likert scale.

Additional Outcome Variables. Since this was a pilot study, we also tested a number of exploratory outcome variables. See ***Supplementary Appendix S2-S5*** for exploratory analysis involving these outcome variables, ***Supplementary Appendix S6*** for the wording of our main outcome variables, and our OSF for full Qualtrics surveys and analysis code: https://osf.io/uaz3y/?view_only=e508b94a32944deab6530d32b820a53f.

Compliance. Because we expected the treatment only to work for participants who complied with our intervention instructions, we carefully measured compliance with the intervention. First, as a data quality check, we examined if people entered a valid Twitter handle using the Twitter API. Those who did not enter a valid Twitter handle were excluded from analysis, because if they did not use Twitter (or if we could not measure their Twitter usage), they were not relevant to our analysis. This left us with 643 participants to analyze for our intent-to-treat (ITT) effects.

We used the Twitter API to download the accounts followed by all participants who provided their Twitter handles in the four weeks during the intervention, as well as during the two weeks after the intervention. In ***Table 3***, we show the accounts from each list we asked participants to follow and unfollow in the weeks during and after the intervention. We also asked participants in the experimental and control condition to follow two accounts that were controlled by the research team. One of these accounts tweeted out pictures of animals Monday-Friday each day

during the intervention. In the Time 2 survey, we asked participants which animals were tweeted to get a sense of how much participants payed attention to Twitter accounts we asked them to follow. ***Table 4*** reveals that in each week, more participants in the experimental condition, as compared to the control condition, followed the accounts we asked them to follow and unfollowed the accounts we asked them to unfollow. Additionally, a roughly equal number of participants in the experimental and control condition followed the two accounts controlled by the research team. However, since we did not achieve full compliance with the intervention, we also report complier average causal effects (CACE), as well as effects for people who followed minimal instructions with the intervention (e.g., followed at least one of the two accounts controlled by the research team during at least one point of the intervention). In ***Figure 4***, we plot the follower networks visually and show how they changed over the course of the experiment.

Table 4. Mean Levels of Compliance Throughout and After the Intervention

Mean # of Accounts Followed (From List of 18 Accounts to Follow)			
Week	Control	Experimental	p-value
Week 1	0.90	9.77	< 0.001
Week 2	1.96	10.84	< 0.001
Week 3	2.02	10.61	< 0.001
Week 4	1.94	9.52	< 0.001
Week 5	1.83	8.92	< 0.001
Week 6	1.79	8.63	< 0.001
Mean # of Accounts Followed (From List of 60 Accounts to Unfollow & 8 Placebo Accounts to Unfollow)			
Week	Control	Experimental	p-value
Week 1	1.87	1.10	0.012
Week 2	1.96	1.12	0.007
Week 3	1.92	1.12	0.010
Week 4	1.87	1.22	0.037
Week 5	1.95	1.25	0.027
Week 6	1.91	1.25	0.038
Mean # of Accounts Followed (From 2 Accounts Controlled by Researchers)			
Week	Control	Experimental	p-value
Week 1	1.11	1.06	0.565
Week 2	1.11	1.06	0.565
Week 3	1.10	1.04	0.425
Week 4	1.04	0.92	0.157
Week 5	0.97	0.85	0.122
Week 6	0.94	0.82	0.116

Note. Mean levels of compliance in the experimental and control conditions among the 643 participants with valid Twitter handles.

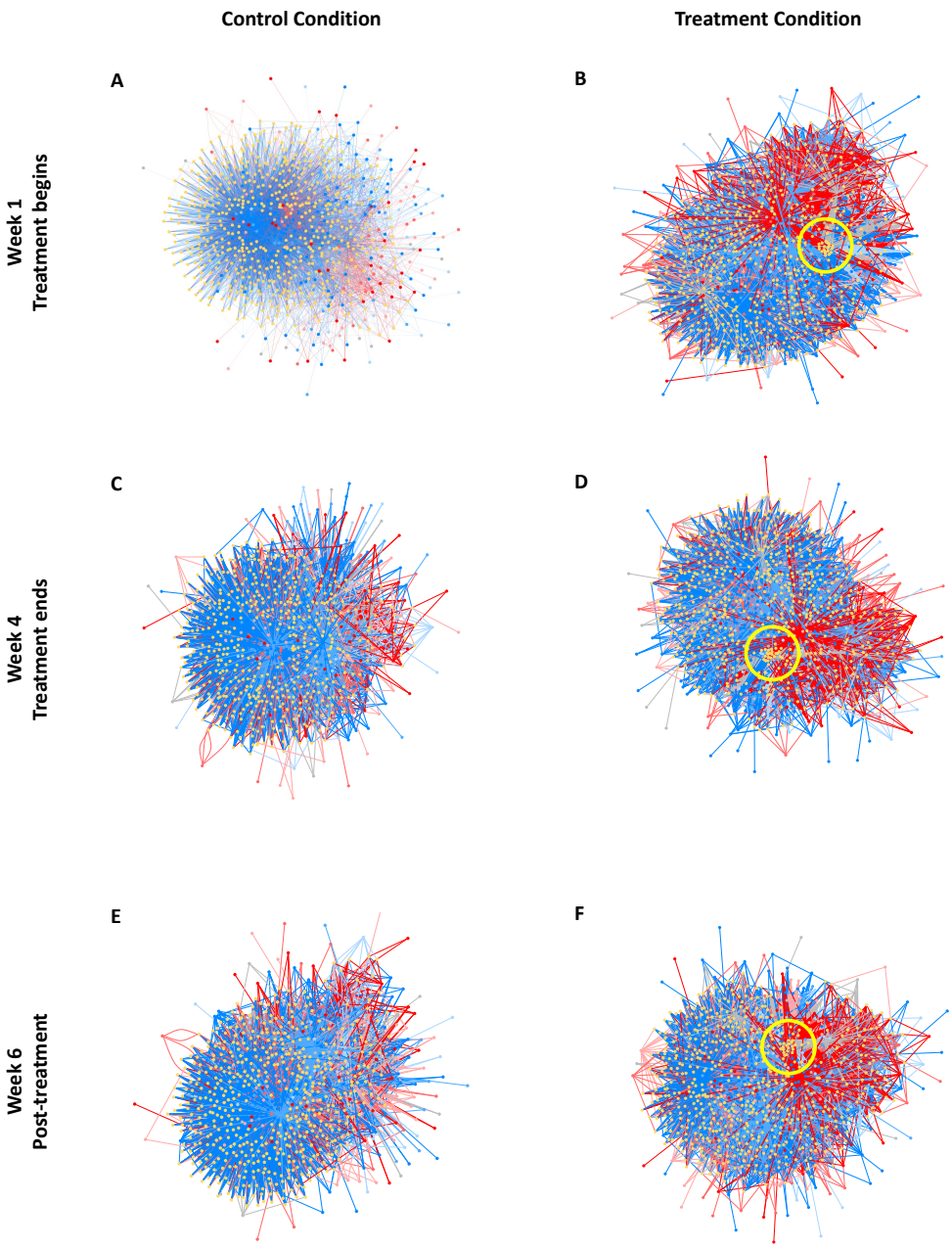


Figure 4. Follower networks of the treatment and control groups at beginning of the experiment (Week 1), end of the experiment (Week 4), and two weeks after the experiment (Week 6). Each node denotes a twitter account and each edge connecting two nodes denote that one account follows the other. Nodes colored in yellow are the accounts followed by more than 10 participants ("influencers"), other nodes are colored based on their favorability towards Republicans, with more red meaning greater favorability. Yellow rings in the treatment condition network graphs highlight the accounts that the participants were asked to follow at the start of the experiment. All

network graphs are generated using the large-graph layout algorithm in the R igraph package, where nodes are arranged to best reflect connections within the network.

Analysis

Intent-To-Treat Analysis. For our intent-to-treat analysis, we report results for all participants with all 643 participants with a valid Twitter account who filled out both surveys, whether or not they complied with the treatment. For this analysis, we ran linear regressions predicting time 2 scores as a function of the condition variable, controlling for time 1 scores. Results are plotted visually in ***Figure 5***. Afterwards, we report a number of robustness checks and exploratory analyses. Additional robustness checks are in the ***Supplementary Appendix S2-S5***.

Affective Polarization. Affective polarization at Time 2 was lower in the experimental condition as compared to the control condition, controlling for Time 1 scores, $b = -0.12$, 95% CI = [-0.23, -0.00], $p = 0.043$.

Perceptions of Twitter Feed. The intervention led people to believe their Twitter feed was more positive over the four-week intervention, $b = 0.17$, 95% CI = [0.05, 2.78], $p = 0.006$.

Well-Being. The intervention led to increased scores on the well-being index, $b = 0.14$, 95% CI = [0.03, 0.25], $p = 0.013$.

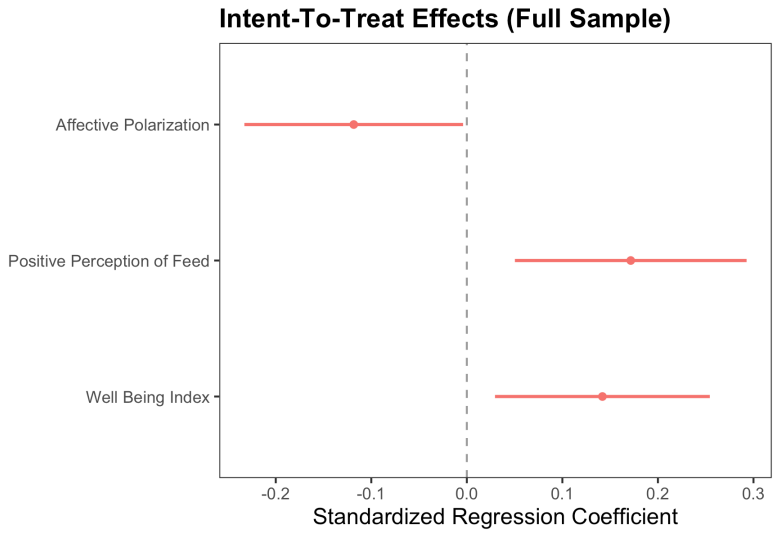


Figure 5. Intent-To-Treat Effects for the Full Sample (regardless of compliance) that provided a valid Twitter handle.

Specific Aspects of Twitter Feed That Changed. We then further investigated what aspects of people's Twitter feed changed in the experimental, as compared to the control, condition, as shown in ***Figure 6***. In addition to the composite measure of positive perceptions of one's Twitter feed increasing, people reported their Twitter feeds to be significantly less polarizing in the experimental condition, b = -0.035, 95% CI = [-0.32, -0.01], p = 0.035, and also thought they saw significantly less fake news, b = -0.18, 95% CI = [-0.18, 0.08], p = 0.018, and significantly more tweets about science, b = 0.16, 95% CI = [0.02, 0.29], p = 0.024.

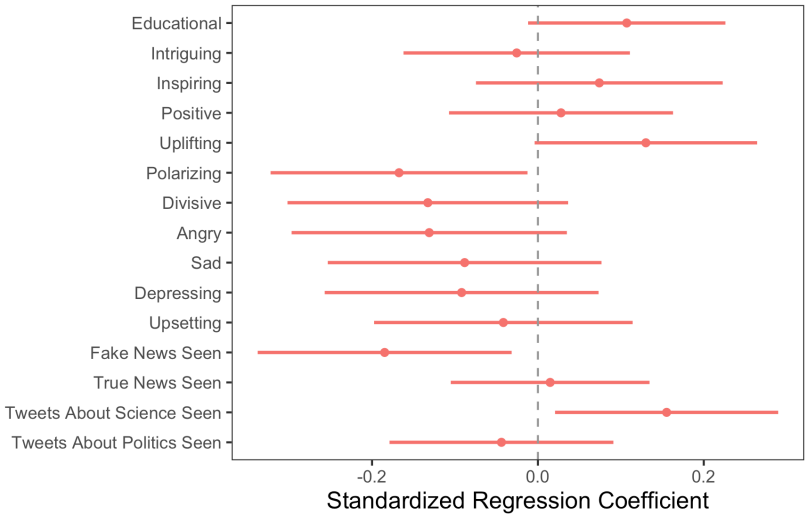


Figure 6. How the month-long intervention affected perceptions of one's Twitter feed. The intervention significantly decreased perceptions of one's Twitter feed as polarizing, decreased the perceived number of tweets containing fake news seen, and increased the perceived number of tweets about science seen.

Following Versus Unfollowing. We expected that effects would be larger for those who unfollowed participants, since not all participants were following any accounts from the list of partisan accounts we assembled to begin with. As an initial exploration of this, we split the dataset into two separate samples: one sample included participants who reported unfollowing at least one person during the intervention (233 in the experimental condition), and another sample that included participants who did not need to unfollow anyone during the intervention (157 in the experimental condition). All participants in the control condition were kept in both samples. See analysis conducted separately in ***Figure 7.***

From an examination of the effect sizes, it appeared as though the effect sizes were larger for those who unfollowed at least one person as compared to those who did not unfollow at least one person. Though it should be noted that the effect sizes are wide and overlapping, and none of the effects were moderated by whether or not a participant unfollowed at least one individual (p = 0.185). Further research should examine whether following or unfollowing behavior drive these effects, or whether these effects are additive.

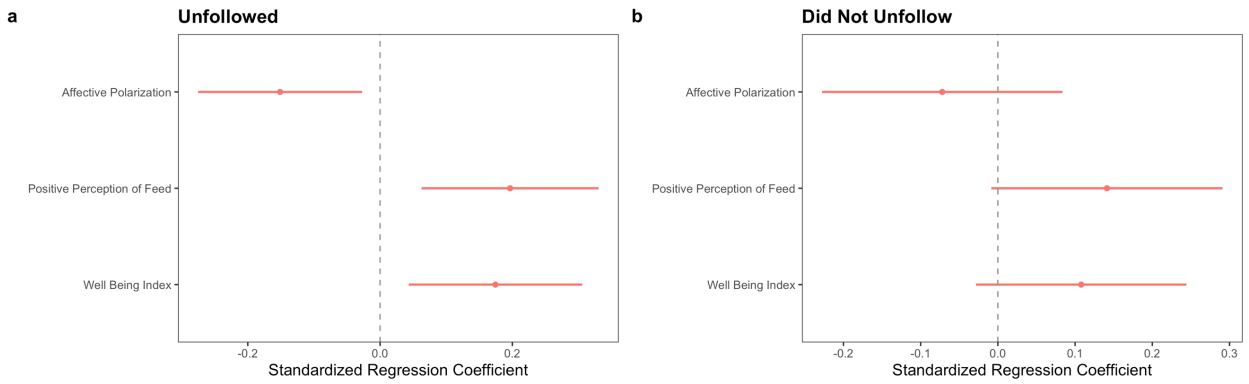


Figure 7. Estimated effect sizes for **A** participants who unfollowed at least one account and **B** participants who did not unfollow at least one account.

Complier Average Causal Effects. To account for differential rates of compliance among participants, we measured complier-average causal effects (CACE) using the two-staged least squares regression approach (Bail et al., 2018; Imbens & Rubin, 2015). We estimated effects for people who fully complied with the treatment. Specifically, we defined compliance as 1) following on average 14 of the accounts from the list of the "accounts to follow" over the four-week intervention (this criteria allowed people who completed the survey early or started late to count as compliers), 2) following 0 of the accounts from the list of "accounts to unfollow" for 3 out of the 4 weeks of the intervention, and 3) correctly recognizing one of the pictures tweeted from our research account over the course of one month, without saying they saw two "distractor" pictures. 186 participants matched these criteria for compliance.

This analysis found significant effects for all the same variables, yet with larger effect sizes. Specifically, we found a decrease in polarization, $CACE$ = -0.26, 95% CI = [-0.51, -0.00], p = 0.047, an increase in well-being, $CACE$ = 0.32, 95% CI = [0.06, 0.57], p = 0.016, and an increase in positive perceptions toward one's Twitter feed, $CACE$ = 0.38, 95% CI = [0.10, 0.66], p = 0.007. Complier average causal effects are shown visually in ***Figure 7***.

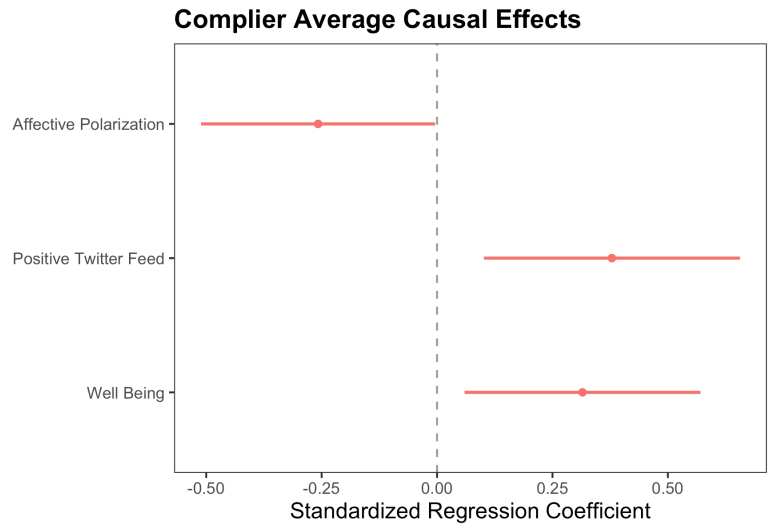


Figure 7. Complier-Average Causal Effects.

Effects for Participants Who Followed Minimal Instructions. We also simply measured Intent-To-Treat effects for the subset of participants who followed one of the two accounts controlled by the researchers at least one point during the intervention, which we defined as the subsample of participants who followed minimal intervention instructions. Only 384 participants actually followed at least one of these accounts, and as shown in ***Table 4***, there was no significant difference in the number of people who followed these accounts in the experimental vs. control condition. Thus, this is the sample who followed minimal intervention instructions, as opposed to answering both surveys without engaging in the intervention in any way. Among this subsample, we found a decrease in polarization, B = -0.17, 95% CI = [-0.33, -0.02], p = 0.024, a significant increase in positive perceptions of ones' Twitter feed, B = 0.26, 95% CI = [0.09, 0.42], p = 0.002, and a significant increase in well-being, B = 0.16, 95% CI = [0.03, 0.29], p = 0.016. For the sample of 281 participants who did not follow minimal intervention instructions, there was no significant effect on any of these outcome variables (ps > 0.300). Thus, it appears as though effects were only present among people who followed the intervention instructions. Results are plotted visually in ***Figure 8***.

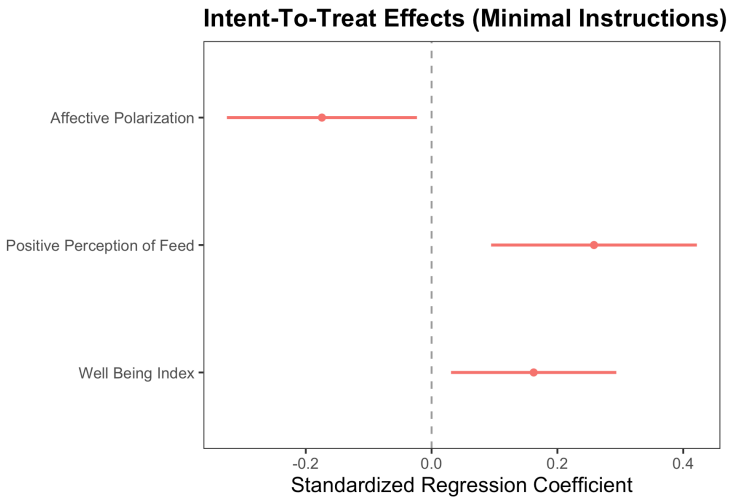


Figure 8. Results from the 384 participants who followed minimal intervention instructions, or followed at least one of the two accounts controlled by the research team at some point during the four-week intervention. 281 participants did not follow these minimal intervention instructions.

Effects of Disabling the Twitter Algorithm. Being asked to disable one's Twitter algorithm had no significant effects on affective polarization, well-being, perceptions of one's Twitter feed (all *ps* > 0.398). Furthermore, the algorithm condition did not significantly moderate any of the other effects (all *ps* > 0.790). Despite the lack of causal effects of the algorithm condition, there was a correlational effect such that those who tended to use the algorithmic Twitter feed more often had a more positive view of their Twitter feed ($M = 2.95$, $SD = 0.52$) than those who did not ($M = 2.82$, $SD = 0.51$), $t(377.23) = -2.61$, $p = 0.009$, $d = 0.27$.

Social Media Behavior. We also examined whether the intervention influenced social media behavior. Following our pre-registered analysis, we used all data from participants who reported a valid Twitter handle that data could be retrieved from via the Twitter API (whether or not these participants completed all survey responses), and excluded participants with more than 50,000 followers. This left a sample size of 663 participants for Twitter analysis.

Quality of News Sources Shared. We used a dataset from NewsGuard, which is a company that rates the trustworthiness of more than 4500 News Sites on a scale of 0-100 (Lapowski, 2018). We scraped all the news URLs tweeted by participants during the intervention and after the intervention and assigned them a quality score based on NewsGuard's rating system. We then calculated an average quality score of each participant in the month before and the month during the intervention. We found that the intervention marginally (but non-significantly) increased the

quality of tweets shared over the course of the intervention when controlling for the quality of tweets sent in the month for the intervention, $b = 0.41$, 95% CI = [-0.03, 0.84], $p = 0.065$. However, only 88 people shared a news article that could be rated by NewsGuard in the month before and the month after the intervention, meaning this analysis was likely underpowered.

Toxicity of Tweets Shared. We measured the average toxicity of tweets using Perspective API, which is a machine learning classifier developed by Google that rates the average "toxicity" level of Tweets. We rated the toxicity of all tweets sent by participants in the month before and the month after the intervention. We then created an average "toxicity" score for all the tweets users in the month during and the month after the intervention. We found that the intervention did not significantly affect the toxicity of tweets sent in the month during the intervention when controlling for the toxicity of tweets sent in the month before the intervention ($p = 0.769$). These results are plotted visually in ***Figure 9***.

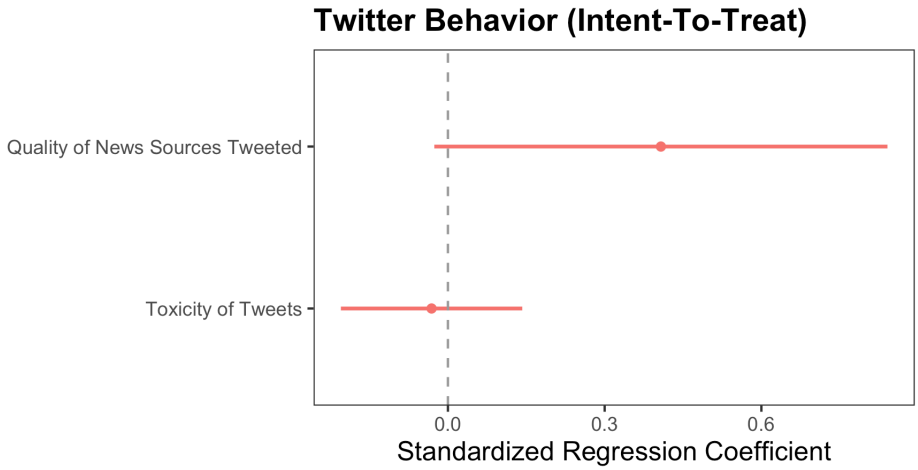


Figure 9. Twitter Behavioral Measures. The intervention marginally increased the quality of news people shared on Twitter in the month during (as opposed to the month before) the intervention. However, the intervention did not have significant effects of the toxicity of tweets people shared in the month during (as opposed to the month before) this intervention.

Free Response Questions. One question that arises from this work is whether participants noticed a change in their news feed over the course of the intervention. When asked what changed in their social media feeds in a free response question, many participants mentioned that their feeds became more positive, less polarizing, and contained more content about science. For example,

one participant said “There was a lot more scientific news that cut the political noise as I found many of the headlines compelling,” and another said "less political quote tweet dunks, and a slight uptick in scientific/nature tweets on my TL.”

Discussion

We found that social media (un)following behavior is both correlationally related to and causally influences beliefs. Specifically, in *Study 1*, we found that following highly partisan political elites and news sources on both the left and the right was associated with affective polarization. People who were very politically engaged on Twitter (or followed more than 10 partisan accounts) showed much higher levels of affective polarization than people who followed 0 partisan accounts. Following conservative (but not liberal) partisan accounts was also associated with sharing lower-quality information online. While previous work has suggested that only a small portion of social media users are politically active, this study suggests that those who are politically active show the highest levels of polarization. Following highly partisan accounts on Twitter was also associated with sharing tweets with high levels of toxicity, and following conservative partisan accounts was associated with sharing lower-quality news URLs.

Yet, it is unclear from prior research whether following highly partisan accounts plays a causal role in increasing polarization. To address this question, we conducted a large-scale randomized control trial where we asked Twitter users to unfollow a list of partisan accounts that and follow another list of accounts that tweeted less partisan content (such as accounts about science, space, or the natural world). We found that following and unfollowing these accounts for a month significantly reduced affective polarization, improved perceptions of one’s Twitter feed, and improved well-being. The intervention also marginally improved the quality of articles shared on Twitter. Thus, following highly partisan accounts on social media has a causal impact on one’s level of affective polarization, suggesting that online polarization is not just driven by self-selection.

While these results are promising, there are limitations behind this work. First, the effects were small, which may have been due in part to low levels of compliance. For instance, the intervention led to a 0.12 standard deviation change in polarization, and many of the key effects were of similar magnitude for our main intent-to-treat analysis. However, a similar study found

that deactivating Facebook altogether for a month led to a 0.16 standard deviation decrease in polarization (Allcott et al., 2020), meaning that deleting social media in its entirety has comparably small effects. Additionally, effect size estimates were higher for people who minimally followed intervention instructions (0.17 standard deviations), and even higher for people who complied with all instructions (0.26 standard deviations). Though, these effect size estimates had large confidence intervals and should not necessarily be considered a precise estimation of the true effect without further replication. Our behavioral data also found a marginal but non-significant increase in the quality of news shared as a result of the intervention. However, since only 88 participants in our sample shared news at all in the month before and month after the intervention, we likely did not have enough power to detect an effect for this outcome variable. As a next step, we plan to replicate and extend on this this work with a larger and more representative sample, and have received grant funding for a large-scale replication and extension of this initial work. We also plan to encourage higher levels of compliance in our replication study, make the intervention stronger, and measure participants' attitudes at more time points.

While previous work has shown that deleting social media as a whole can have causal effects on polarization (Allcott et al., 2020) and well-being (Asimovic et al., 2021), our analysis suggests that social media may have very different impacts based on how it is used. Many people are not politically-engaged online, but those who are show very high levels of polarization. However, moving people out of highly partisan networks online causally decreases polarization. This work makes an important theoretical contribution, showing that polarization is not entirely driven by self-selection and that social influence online matters. This work can also inform interventions, policy, and social media design changes, since this study suggests that there may be benefits to moving people out of highly partisan networks.

Chapter 7. Discussion

These studies presented in this thesis represent a multi-method investigation of the psychology of (mis)information and political polarization online. They aim to address two important questions: 1) why do people believe in and share (mis)information online? And 2) What are the consequences of (mis)information exposure? They also aim to contribute to a number of debates in the current psychological literature. For instance, while some think fake news sharing is due to motivated reasoning and partisanship (Van Bavel, Harris, et al., 2021; Van Bavel & Pereira, 2018), others argue that reflection and lack of attention play a larger role (Lawson & Kakkar, 2021; Pennycook, Epstein, et al., 2021b). Furthermore, while some argue that social media plays a causal role in polarization (Haidt, 2022; Van Bavel, Rathje, et al., 2021a), others suggest that polarization has been increasing overtime and that social media may play a more limited role in polarization than often thought (Boxell et al., 2017). Resolving some of these conflicts, this thesis integrates multiple perspectives, demonstrating the role that accuracy and social motivations, as well as other factors, such as individual difference variables and information exposure, play in shaping the belief in and sharing of misinformation. Further, rather than making large-scale claims about social media usage as a whole, this thesis demonstrates how specific ways of using social media (e.g., following highly partisan accounts) contribute to polarization. These results make contributions to psychological theory while also having practical implications for improving social media platforms.

Summary of Findings

Chapter 2 builds on prior work on social media virality, which primarily focuses on the role of emotion. Taking a different approach from prior work on emotion, this chapter examined whether social identity motivations (such as in-group favoritism and out-group derogation) were successful at generating engagement on two of the largest social media platforms: Facebook and Twitter. Analyzing posts from news media accounts and US congressional members (n = 2,730,215), this chapter found that posts about the political out-group were shared or retweeted about twice as often as posts about the in-group. Each individual term referring to the political out-group increased the average amount of shares or retweets of a post by 67%. The effect size

of out-group language was about 4.8 times as strong as that of negative affect language and 6.7 times as strong as that of moral-emotional language. Language about the out-group was a strong predictor of "angry" reactions, and language about the in-group was a strong predictor of "love" reactions, which may reflect in-group favoritism and out-group derogation (Tajfel et al., 1971). These effects were not moderated by political orientation or social media platform, but the effects were stronger among political leaders than among news media accounts. In sum, out-group derogation was the strongest predictor of social media engagement among all predictors measured, suggesting that social identity motivations should be considered in future research examining social media virality. Further, these results suggest that social media may be creating perverse incentives for people to post divisive content about out-group members, which could be a potential mechanism by which social media increases polarization.

Chapter 3 then experimentally manipulated participants' accuracy and social motivations across four online experiments (n = 3,364). Providing people with financial incentives to accurately identify true versus false headlines both improved accuracy and reduced partisan bias in belief in headlines by about 30%. This effect was primarily driven by incentives leading people to report true news from the opposing party as more accurate. Incentives did not significantly influence accuracy judgements of false news, however. Turning to social motivations, incentivizing people to correctly identify news that would be liked by political allies decreased accuracy at discerning between true and false headlines. Replicating prior work, US conservatives were less accurate than US liberals (Garrett & Bond, 2021; A. Guess et al., 2019; Roozenbeek et al., 2022b). However, incentives closed the gap in accuracy between conservatives and liberals by more than half, suggesting that conservatives' worse truth discernment may in part reflect a lack of motivation to be accurate instead of lack of knowledge or ability alone. A non-financial accuracy motivation intervention was also effective, which helps rule out alternative explanations and also suggests that motivation-based interventions can be applied at scale. These results contradict the interpretation that partisan bias in belief in (mis)information – as well as conservatives' worse ability to discern between true and false news – reflect differences in prior beliefs alone (Pennycook & Rand, 2021c). Instead, they suggest that accuracy and social motivations can play a causal role in shaping judgements of (mis)information. However, other factors besides motivation alone mattered as well: political ideology, political knowledge, and cognitive reflection were also predictive of the accurate

identification of true versus false headlines. These results help resolve a debate in the literature about motivation versus prior beliefs by looking at the causal role of various motivations (Druckman & McGrath, 2019; Pennycook & Rand, 2021c), and also help contribute to an integrated account of (mis)information belief and sharing (Van Bavel, Harris, et al., 2021).

Chapter 4 turns to misinformation-reduction interventions, presenting a meta-analysis of data from a popular misinformation-reduction intervention called the accuracy nudge (Pennycook, Epstein, et al., 2021b; Pennycook, McPhetres, et al., 2020). I found that the effectiveness of the accuracy nudge intervention depends upon partisanship such that the nudge has little impact for US conservatives and Republicans. Since US conservatives and Republicans are far more likely to share misinformation than US liberals and Democrats (Guess et al., 2019; Lawson & Kakkar, 2021; Osmundson, 2021), this intervention may be ineffective for those most likely to spread fake news. Building on some of the findings of *Chapter 3*, it is possible that different motivations among the two political parties in the United States help explain this asymmetry (Jost et al., 2018), though this hypothesis needs to be tested further in future work. This chapter suggests that an integrated account of (mis)information belief and sharing that incorporates a number of interacting factors such as inattention, partisanship, and motivation may be needed to help understand how to effectively mitigate the spread of misinformation (Van Bavel, Harris, et al., 2021).

Chapter 5 examines how real-world social media usage is associated with belief in (mis)information. This chapter combines survey data measuring attitudes toward the COVID-19 vaccine with Twitter data in two studies (N_1 = 464 Twitter users, N_2 = 1,600 Twitter users) to examine how real-world social media behavior is associated with vaccine hesitancy in the United States (US) and United Kingdom (UK). In Study 1, we found that following the accounts of US Republican politicians or hyper-partisan/low-quality news sites was associated with lower confidence in the COVID-19 vaccine – even when controlling for key demographics such as self-reported political ideology. US right-wing influencers (e.g., Candace Owens, Tucker Carlson) had followers with the lowest confidence in the vaccine. Network analysis revealed that low and high vaccine-confident participants separated into two distinct communities (or "echo chambers"), and centrality in the more right-wing community was associated with vaccine hesitancy in the US, but not in the UK. In Study 2, we found that one's likelihood of not getting the vaccine was associated with retweeting and favoriting low-quality news websites on Twitter.

In sum, vaccine hesitancy is associated with following, sharing, and interacting with low-quality information online, as well as centrality within a conservative-leaning online community the US. Building off of prior work that analyzes survey data (Loomba et al., 2021b) and social media data (Johnson et al., 2020b) separately, our work examined how different ways of interacting with (mis)information on social media (e.g., favoriting, tweeting, and following) predicted vaccine hesitancy.

While *Chapter 5* was a correlational study, *Chapter 6* examined how certain ways of using social media causally impact beliefs and attitudes. This chapter begins with correlational research demonstrating that those who are politically active (or follow partisan accounts) on social media show much higher degrees of affective polarization than those who are not politically active (or do not follow any partisan accounts). These politically-active Twitter users also share create more "toxic" tweets on average, and, among conservatives, share lower-quality news URLs. Then, in a randomized controlled trial (n = 795), participants were incentivized to unfollow highly partisan accounts on Twitter and follow accounts about non-partisan topics (science, space, etc.) for one month. Changing social media following behavior reduced affective polarization, improved well-being, and led people to report more positive perceptions of their Twitter feed. These results address debates in the literature about whether social media causally increases polarization. While randomized-controlled trials have yielded contradictory results about the causal effects of deleting social media, this randomized controlled trial suggests that a specific way of using social media (e.g., following highly partisan accounts) has a causal impact on polarization and well-being. This study also has practical implications – rather than encouraging people to reduce their overall social media usage, this study suggests that changing the ways in which one interacts with social media may have benefits.

Theoretical Contributions

This thesis makes a number of theoretical contributions to the social psychological literature, as well as literature in related fields, such as political science. *Chapter 2* takes predictions from Social Identity Theory (Tajfel et al., 1979) and Self Categorization Theory (Turner et al., 1987) and analyzes them within the domain of social media. The paper proposes that the way we interact with content online reflects our identity-based motivations, which might

lead certain types of content to go viral – especially when certain identities, such as our political identities, are highly salient (Brady, Crockett, et al., 2019; Turner et al., 1979). The correlational data largely finds support for this prediction that identity-related content achieves greater virality online. However, mirroring findings about negative partisanship (Abramowitz & Webster, 2018) and political sectarianism (Finkel et al., 2020) in the United States, out-group animosity appeared to achieve more virality than in-group favoritism. While much of the support from these theories come from experiments and survey data, this chapter finds support for these theories using online, observational data. Furthermore, building on theories of online virality that mostly look at emotion (Berger & Milkman, 2012), this study provides new evidence for the role that social identity plays in predicting online virality. While there is some debate in the social identity literature (Brewer, 2017) and political science literature (Finkel et al., 2020) about whether out-group hate is stronger than in-group love, this study showed a context in which out-group hate attracted more attention than in-group love. This could possibly be because of the current state of political sectarianism in the United States (Finkel et al., 2020), or because negativity and moral outrage in general captures more attention on social media (Brady et al., 2020; C. E. Robertson, Pröllochs, et al., 2022).

Chapter 3 addresses key theoretical debates within the motivated reasoning literature. Recently, it been suggested that future research needs to "carefully manipulate people's motivations in the processing of (mis)information that is politically (dis)concordant" to help differentiate between multiple explanations for partisan bias in fake news belief (van der Linden, 2022). While some claim that partisan bias in belief in fake news may reflect partisan differences in prior beliefs and information exposure (Pennycook & Rand, 2021c), others suggest that partisan differences may in part reflect motivated reasoning (Van Bavel & Pereira, 2018) or expressive responding (Schaffner & Luks, 2018). Helping to reconcile these competing explanations, our results present evidence that motivation causally affects belief in misinformation. However, motivation does not completely explain judgements of news headlines: incentives only influenced perceptions of true, but not false, headlines, and incentives did not fully eliminate partisan bias. Thus, motivation and prior knowledge likely both play a role in judgements of news headlines, and motivation may play different roles depending on the context. Following the suggestion that the misinformation literature should offer a "more integrated theoretical account of susceptibility to misinformation," (van der Linden, 2022) these

results demonstrate how multiple disparate factors, such as motivation, ideology, cognitive reflection, political knowledge, affective polarization, and prior knowledge all contribute to the belief and sharing of misinformation. This chapter also points to the danger with presenting false dichotomies between competing theories or focusing too strongly on one variable in predicting misinformation belief, as multiple factors appear to better explain misinformation belief than one factor on its own.

Chapter 4 of the thesis questions a key assumption of the inattention-based account of misinformation sharing. While it has been suggested that the inattention-based account of fake news sharing challenges accounts based on partisanship (Pennycook, Epstein, et al., 2021b; Pennycook & Rand, 2021c), the results from *Chapter 3* suggest that inattention and partisanship may interact, since an inattention-based intervention appears to work better for liberals/Democrats than for conservatives/Republicans. These results once again help contribute to a broad theoretical account of fake news sharing (Van Bavel, Harris, et al., 2021; van der Linden, 2022), showing that factors such as partisanship and inattention both help explain why people share fake news.

Chapter 5 helps contribute to recent theoretical debates about the existence of "echo chambers," (Barberá et al., 2015; Cinelli et al., 2021; A. Guess et al., 2018; Wojcieszak et al., 2021) suggesting that there not only appear to be echo chambers about politics, but also about other factors, such as vaccine hesitancy. Using real-world behavioral data, it also contributes to theories about partisan cues (Bisgaard & Slothuus, 2018b; S. Pink et al., 2021) and misinformation exposure (Loomba et al., 2021c; Van Bavel, Harris, et al., 2021; van der Linden, 2022), demonstrating that interacting with politicians and low-quality news sources online is associated with vaccine hesitancy. The cross-cultural perspective of the chapter also shows how an issue can quickly become politicized in one cultural context but not another.

Chapter 6 helps resolve debates about the causal effects of social media. There has been much debate about how social media causally impacts polarization (Boxell et al., 2017; Van Bavel, Rathje, et al., 2021a), and randomized-controlled trials about the effects of deleting social media have yielded contradictory results (Allcott et al., 2020; Asimovic et al., 2021). These conflicting results may reflect the fact that people use social media in very different ways and have different offline and online social networks (Asimovic et al., 2021). This chapter examines how a specific way of using social media (e.g., following highly partisan accounts) causally

increases polarization and reduces well-being. In doing so, it suggests that social media use on its own may not be detrimental to well-being and intergroup relations; instead, the specific ways in which we interact with social media may be more important to analyze. Since everyone uses social media differently and is embedded within unique social networks, this study helps us move away from the broad question of "does social media cause polarization," and instead helps answer the more meaningful question of "what way of using social media contribute to polarization?" It should be noted that this study found effects for affective polarization (or distain for the opposing party), as opposed to ideological polarization (or polarized opinions on policy positions) (Druckman et al., 2020), and future studies should be precise about which type of polarization is measured clearly to avoid conceptually confusion.

Methodological Contributions

This thesis also provides a number of methodological contributions. *Chapter 2* was one of the early papers that took advantage of big data from Facebook's *Crowdtangle* data source for collecting public Facebook data, which Facebook will allegedly be shutting after the 2022 midterm elections (Lawler, 2022). This unique data source provides the opportunity to analyze Facebook reactions (e.g., angry, haha, wow, love, like, and sad reactions), which allows for a more detailed analysis of the sentiment evoked by social media posts that goes beyond traditional sentiment dictionary-based analysis methods. This big data approach also has a number of benefits beyond traditional survey experiments in social psychology and political science, such as a very large sample size and ecological validity. Finally, analysis of both Facebook and Twitter data allowed us to make comparisons across social media platforms.

Chapter 5 uniquely connected survey data to Twitter data to examine how different ways of interacting with social media are associated with self-reported attitudes. This unique method provided the chance to examine at a detailed level how factors such as real following behavior, tweeting behavior, favoriting behavior, and centrality within certain online networks all were associated with attitudes about the vaccine. This methodology has substantial benefits over studies that examine self-reported social media use, since self-reported social media use only has limited correlations with actual social media use (Hodes & Thomas, 2021). This study also has benefits over analysis of social media data alone, which does not provide insight into the beliefs

and characteristics of those who post on social media. A large portion of the data collection for this chapter was made possible because of a web app I developed called "Have I Shared Fake News" (link: https://newsfeedback.shinyapps.io/HaveISharedFakeNews/) which has been used by thousands of people since it was launched in 2021. This web app presents a case study for how researchers can collect massive datasets of social media data linked to self-report data from participants around the globe in a gamified way.

Finally, *Chapter 6* presented a new type of social media field experiment. While previous social media field experiments have asked people to delete their social media accounts altogether (Allcott et al., 2020; Asimovic et al., 2021) or to follow different accounts on social media (Bail et al., 2018), this study looked at the effect of social media *unfollowing* behavior, which provides a new way of testing how leaving highly partisan online networks affects attitudes and behaviors. As social media usage grows around the world, and as social media apps diversify, it will be necessary to create more field experiments like this that analyze different ways of using social media. More field experiments like this one will also be helpful to address causality and move beyond problems with self-reported social media behavior (Hodes & Thomas, 2021).

This thesis also employed open sciences practices. All four experimental studies in *Chapter 3* were pre-registered (Van't Veer & Giner-Sorolla, 2016), data collection and analysis in *Chapter 5* was pre-registered, and the randomized control trial *Chapter 6* was also pre-registered (analysis in *Chapter 2* and *Chapter 4* were not pre-registered as these were analyses of secondary data). Code and data were shared for each study so that people can build on these methods, though it should be noted that only limited social media data could be shared due to privacy concerns.

Practical Contributions

These chapters also have a number of practical implications that could be of interest to those concerned about social media, polarization, and misinformation, such as social media companies, policy-makers, or those trying to design and deploy interventions. For instance, *Chapter 2* demonstrates how divisive content may be amplified by social media platforms, and how algorithmic decisions (such as giving more weight to the "angry" reaction) can contribute to the amplification of divisive content, which is relevant for those who are trying to build healthier

and less polarizing social media platforms. The paper suggests that certain algorithmic changes could be made to discourage the sharing of divisive content, such as giving less weight to the "angry" reaction and more weight to the "love" and "like" reaction in the news feed algorithm. More broadly, it highlights that social media may be generating perverse incentives for the creation of divisive content, and suggests that the incentive structure on social media platforms can be shifted to help create more constructive online conservations.

Chapter 3 demonstrates that making accuracy motivations more prominent and decreasing partisan identity-based motivations may be useful for misinformation-reduction interventions. Social media companies can consider design changes that help motivate people to share more accurate content (e.g., through social norms, prompting reputational concerns, etc.), and design changes that do not reward people for viral misinformation. Policy-makers can also consider regulations that help shift incentives on social media toward accuracy.

Chapter 5 shows that a popular misinformation-reduction intervention might have limited effects for US conservatives (the population that shares the most misinformation), suggesting that new strategies may be necessary to curb the spread of fake news online. Thus, people can consider alternative strategies that help motivate conservatives to share accurate content. For example, identity-based interventions that show participants the social norms of one's ingroup have been shown to help motivate conservatives to share more accurate content (Pretus et al., 2022), and inoculation interventions have been found not to be moderated by political affiliation (Roozenbeek et al., 2022). Future efforts should be made to help design effective interventions among populations that are at high risk of sharing misinformation.

Chapter 6 identifies online communities, influencers, and media sources that are associated with vaccine-hesitancy, which has implications for those who are trying to communicate accurate information about public health. For instance, the White House has at times used social media to try to communicate important public health information, such as information about the monkeypox outbreak (Garcia, 2022), and this study could also inform future efforts like this one. This study also identifies the limitations of communicating public health information in a polarized social media environment, and demonstrates the risks associated with following and interacting with certain influencers and news sources online.

Chapter 6 illustrates the causal effects of being embedded within highly partisan online networks on social media. Social media companies may want to consider design changes that

lead users out of highly partisan networks, given the detrimental effects on polarization and well-being. Policy-makers can also consider these negative effects when designing social media regulations. Additionally, social media users may want to re-evaluate the people they choose to follow on social media, and this intervention can help inform more constructive social media following behavior.

Limitations

These chapters are not without their limitations. For instance, *Chapter 2* found that polarizing content about one's out-group was highly likely to go viral in large-scale datasets on Facebook and Twitter. However, this is a correlational study that cannot address the factors that *cause* social media sharing behavior. Furthermore, this study did not have any data behind the social media users interacting with these posts, so it is unclear whether this behavior was driven by certain demographic subgroups (e.g., highly partisan social media users). Nonetheless, the effects were robust across social media platforms and context, and have been conceptually replicated by other research teams using machine learning methods (Yu et al., 2021). Replications of this study are also being conducted by separate research teams in the context of Turkey and Russia/Ukraine, and pilot data from these studies show similar results to our study.

Chapter 3 addresses the limitations of *Chapter 2* regarding causality by experimentally manipulating motivations for engaging with news headlines, but this chapter lacks the ecological validity of *Chapter 2*, since it is a survey experiment that does not look at real social media behavior. In other words, it is still unclear how the results of this study might translate to the field. However, survey results about fake news have generalized to the field in the past (Mosleh et al., 2020; Pennycook, Epstein, et al., 2021b). Future studies should build on this work in the field, aiming to change accuracy and social motivations in a real-world context.

Chapter 4 presented a meta-analysis showing that accuracy nudges were less effective for conservatives than liberals. While this meta-analysis included a very large overall sample, it still did not include a number of other datasets that tested the accuracy nudge in other studies due to inavailability of these datasets at the time of analysis (Pennycook & Rand, 2022). Testing these additional datasets would have provided greater power and generalizability to different stimuli and samples to ensure the robustness of our results. Furthermore, *Chapter 4* gives no insight into

why accuracy nudges appear to be less effective for conservatives; it merely points out an interesting pattern that has implications for misinformation interventions in the future. Future work should examine the mechanism behind this partisan asymmetry, examining whether it is due to conservatives having less baseline knowledge, different motivations, or other factors.

Chapter 5 demonstrates that social media behavior is correlated with self-reported vaccine hesitancy. However, this chapter presents a correlational study and does not address whether social media usage causally impacts vaccine attitudes. Thus, it is subject to multiple interpretations: vaccine-hesitant individuals could be seeking out like-minded individuals and information that confirms their beliefs, or people's opinions about the vaccine could be influenced by the people they are exposed to on social media. Despite this limitation regarding causality, the study finds a robust association between attitudes and real-world social media behavior, which provides advantages over studies that only look at self-reported social media behavior.

Chapter 6 addresses the limitation regarding causality in *Chapter 5* by presenting the results of a social media field experiment. While this study is ecologically valid, provides a combination of survey results and behavioral data, and allows for causal interpretation, this study was underpowered to test some variables (such as the quality of news articles shared on Twitter). It also had a majority college student sample, and had somewhat high levels of non-compliance. Furthermore, the intervention only manipulated one aspect of people's media diet (Twitter), and we do not know the extent to which people were consuming partisan news from other sources. However, because we recruited our sample via Twitter advertisements, most of our sample reported being very active Twitter users. Future work should replicate these results in a larger, more representative sample.

While each study has its own set of unique limitations, conducting many studies that utilize multiple methods (e.g., large-scale text analysis, online experiments, survey data linked to Twitter data, and a social media field experiment) helps address the limitations of each individual study. Though, this thesis faces a larger limitation: all studies in this thesis focus primarily on United States or United Kingdom participants, even though social media is used globally by about half the world's population. A psychological account of the effects of social media needs to take a global perspective and acknowledge cross-cultural differences. As noted in the thesis, the effects of social media differ in different contexts (Asimovic et al., 2021), and

other researchers have proposed that social media might have more beneficial effects in developing, as opposed to established, democracies (Lorenz-Spreen et al., 2021), facilitating access to important information instead of fostering intergroup conflict. While there are benefits to zooming into the United States cultural context (as it has been studied in depth by social psychologists, economists, and political scientists, making it easier to build on existing theory and data sources) it is difficult to generalize conclusions from these studies to much of the rest of the world. Future research on social media should follow in the footsteps of other global studies (Awad et al., 2018; Jackson et al., 2019; Mehr et al., 2019; Ruggeri et al., 2021; Van Bavel et al., 2022) that aim to examine human universals and cross-cultural differences. It is possible that the kind of content that goes viral, the predictors of (mis)information sharing, and the causal effects of social media use are very different across diverse cultures.

Future Directions

These chapters present only the beginning of body of work that aims to better understand social media's impact on society. Future work is already underway to build on specific projects in this thesis. For example, we are currently conducting an adversarial collaboration with the original authors of the accuracy nudge studies (Pennycook, Epstein, et al., 2021b; Pennycook, McPhetres, et al., 2020) to examine whether political party affiliation and ideology moderate the effect of the accuracy nudge in a much larger sample of studies we were not yet able to analyze (Pennycook & Rand, 2022), and are conducting additional analysis to figure out potential mechanisms behind this partisan asymmetry. This collaboration addresses key limitations from *Chapter 4* and presents a model for how scientists with competing theories and hypotheses can try to reach consensus (Clark & Tetlock, 2021). We are also replicating the field experiment reported in *Chapter 6* in a larger and more representative sample to build on the effects we found in this initial study, which will help ensure the robustness of the results we found. In this second study, we will also create a slightly stronger manipulation, aim to achieve greater compliance, and measure people's Twitter networks before the intervention to have a within-subjects measurement of how people's networks changed before and after the intervention.

We are also currently planning "global studies" to examine the phenomena explored in the thesis around the world, building on the efforts of previous global studies (Van Bavel et al.,

2022). For instance, in one planned study, we plan to translate the dictionaries we used for analysis in *Chapter 2* into many languages (with the help of cross-cultural collaborators) and examine the factors that predict virality around the globe. In another planned study, we are aiming to conduct a field experiment in which we ask people to deactivate their social media accounts (Facebook, Instagram, etc.) across several countries to examine the causal effects of social media on inter-group attitudes and well-being cross-culturally. In these global studies, we aim to examine whether the effects of social media deactivation are moderated by country-level variables, such as the strength of a country's democracy (Lorenz-Spreen et al., 2021). These ambitious studies will help examine cross-cultural universals and differences in the phenomena presented in this thesis.

Conclusions

When *Chapter 2* of the thesis was published in June of 2021, it warned that Facebook's algorithm change in 2018 – which gave more weight to comments, reactions, and shares (as opposed to likes) in the News Feed algorithm – may have amplified posts reflecting out-group animosity. Shortly afterward, the *Wall Street Journal* reported a series of articles called the "Facebook Files," which shared information about leaked internal documents from Facebook Whistleblower Francis Haugen. One of these documents said that Facebook was aware that its algorithm shift in 2018 led to the promotion of divisive content (Hagey & Horwitz, 2021). In other words, our analysis allowed us to make a data-driven prediction about Facebook's algorithm shift that was largely supported by subsequent *Wall Street Journal* reporting about Facebook's leaked internal research.

As this anecdote illustrates, one of the unique things a social science researcher can do in the current age is reveal truths that social media companies would be unlikely to reveal themselves. While social media companies have many advantages over social scientists because they can collect massive amounts of data and conduct experiments on their platforms, social scientists have the freedom to research topics that could risk painting these companies in a negative light, such as the incentive structures that lead people to create and share false and divisive content online. Social scientists are (or should be) motivated by accuracy, and do not have the same profit-driven motives social media companies have.

In the coming years, as social media continues to grow, expand, and perhaps become unrecognizable, social scientists who study social media should continue to critically examine the incentive structures that undergird behavior on online platforms and explore ways to improve those structures. Psychology and related disciplines should aim to create an enduring science of social media, with theory-driven studies that are sensitive to the ever-changing nature of social media platforms while also being able to speak beyond the current moment in time. Scientists should aim to accurately understand the potential societal harms of this rapidly growing technology, and in doing so, help policy-makers, practitioners, and others create a better online world.

References

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.

Abramowitz, A. I., & Webster, S. W. (2018). Negative partisanship: Why Americans dislike parties but behave like rabid partisans. *Political Psychology*, *39*, 119–135.

Acerbi, A., Altay, S., & Mercier, H. (2022). *Research note: Fighting misinformation or fighting for information?*

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *The Journal of Politics*, *76*(3), 607–620.

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2019). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, 0894439319828011.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676.

Altay, S., Hacquin, A.-S., & Mercier, H. (2019). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 1461444820969893.

Amira, K., Wright, J. C., & Goya-Tocchetto, D. (2021). In-group love versus out-group hate: Which is more important to partisans and when? *Political Behavior*, *43*(2), 473–494.

Appiah, O., Knobloch-Westerwick, S., & Alter, S. (2013). Ingroup favoritism and outgroup derogation: Effects of news valence, character race, and recipient race on selective news reading. *Journal of Communication*, *63*(3), 517–534.

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

Arceneaux, K., Gravelle, T. B., Osmundsen, M., Petersen, M. B., Reifler, J., & Scotto, T. J. (2021). Some people just want to watch the world burn: The prevalence, psychology and politics of the 'Need for Chaos.' *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200147.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational Influence Processes*, *58*, 295–303.

Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, *118*(25).

Aslett, K., Godel, W., Sanderson, Z., Persily, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Measuring belief in fake news in real-time. *CEUR Workshop Proceedings*, *2890*.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 1–9.

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

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408–446.

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311.

Barberá, P. (2014). How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. *Job Market Paper, New York University*, *46*.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science*, *14*(2), 292–303.

Bastos, M., Mercea, D., & Baronchelli, A. (2018). The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PloS One*, *13*(11), e0206841.

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2021). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *22*.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497.

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

Bayes, R., Druckman, J. N., Goods, A., & Molden, D. C. (2020a). When and how different motives can drive motivated political reasoning. *Political Psychology*, *41*(5), 1031–1052.

Bayes, R., Druckman, J. N., Goods, A., & Molden, D. C. (2020b). When and how different motives can drive motivated political reasoning. *Political Psychology*, *41*(5), 1031–1052.

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, *22*(7), 891–893.

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

Bhadani, S., Yamaya, S., Flammini, A., Menczer, F., Ciampaglia, G. L., & Nyhan, B. (2022). Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*, 1–11.

Biddlestone, M., Azevedo, F., & van der Linden, S. (2022). Climate of conspiracy: A meta-analysis of the consequences of belief in conspiracy theories about climate change. *Current Opinion in Psychology*, 101390.

Bisgaard, M., & Slothuus, R. (2018a). Partisan elites as culprits? How party cues shape partisan perceptual gaps. *American Journal of Political Science*, *62*(2), 456–469.

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

Bishop, G. F. (2004). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Rowman & Littlefield Publishers.

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin*, *46*(12), 1702–1711.

Boczkowski, P. J., Mitchelstein, E., & Matassi, M. (2018). "News comes across when I'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New Media & Society*, *20*(10), 3523–3539.

Bolland, J. M. (1988). Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, *10*(3), 233–253.

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*(2), 235–262.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298.

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1–18.

Borukhson, D., Lorenz-Spreen, P., & Ragni, M. (2022). When Does an Individual Accept Misinformation? An Extended Investigation Through Cognitive Modeling. *Computational Brain & Behavior*, 1–17.

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.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2019). The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspectives on Psychological Science*, 1745691620917336.

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.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. (2021). *How social learning amplifies moral outrage expression in online social networks*.

Brady, W. J., & Van Bavel, J. J. (2021a). *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis*.

Brady, W. J., & Van Bavel, J. J. (2021b). *Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks*.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*(10), 1802.

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.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2007). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, *20*(2), 172–188.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515.

Brewer, M. B. (2017). *Intergroup discrimination: Ingroup love or outgroup hate?*

Brewer, M. B., Manzi, J. M., & Shaw, J. S. (1993). In-group identification as a function of depersonalization, distinctiveness, and status. *Psychological Science*, *4*(2), 88–92.

Brown, J. R., & Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*, 1–11.

Bullock, J. G. (2020). *Party cues*. Oxford University Press, New York, NY.

Bullock, J. G., & Lenz, G. (2019). Partisan bias in surveys. *Annual Review of Political Science*.

Bunker, C. J., & Varnum, M. E. (2021). How strong is the association between social media use and false consensus? *Computers in Human Behavior*, *125*, 106947.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American Voter New York: Wiley*.

Chu, J., Pink, S. L., & Willer, R. (2021). Religious identity cues increase vaccination intentions and trust in medical experts among American Christians. *Proceedings of the National Academy of Sciences*, *118*(49).

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–313.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9).

Clark, C. J., & Tetlock, P. E. (2021). Adversarial collaboration: The next science reform. *Political Bias in Psychology: Nature, Scope, and Solutions. Springer*.

Clement, J. (2020). Number of social network users worldwide from 2017 to 2025. *Retrieved June*, *4*, 2020.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808.

Costa, L. da F., Oliveira Jr, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., Viana, M. P., & Correa Rocha, L. E. (2011). Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, *60*(3), 329–412.

Costa, M. (2020). Ideology, Not Affect: What Americans Want from Political Representation. *American Journal of Political Science*.

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

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2020). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour*, 1–11.

Druckman, J. N., & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, *83*(1), 114–122.

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119.

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review*, 57–79.

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences*, *114*(36), 9587–9592.

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and twitter data. *Sage Open*, *9*(1), 2158244019832705.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.*

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75–87.

Enders, A., Farhart, C., Miller, J., Uscinski, J., Saunders, K., & Drochon, H. (2022). Are Republicans and Conservatives More Likely to Believe Conspiracy Theories? *Political Behavior*, 1–24.

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). *Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online.*

Evanega, S., Lynas, M., Adams, J., Smolenyak, K., & Insights, C. G. (2020). Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic.' *JMIR Preprints*, *19*(10), 2020.

Fan, R., Xu, K., & Zhao, J. (2020). Weak ties strengthen anger contagion in social media. *ArXiv Preprint ArXiv:2005.01924*.

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).

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Bavel, J. J. V., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Flores, A., Cole, J. C., Dickert, S., Eom, K., Jiga-Boy, G. M., Kogut, T., Loria, R., Mayorga, M., Pedersen, E. J., & Pereira, B. (2022). Politicians polarize and experts depolarize public support for COVID-19 management policies across countries. *Proceedings of the National Academy of Sciences*, *119*(3).

Fowler, A., & Harris, W. G. (2020). Learning from the Opposition. *Working Paper*.

Freeman, D., Loe, B. S., Chadwick, A., Vaccari, C., Waite, F., Rosebrock, L., Jenner, L., Petit, A., Lewandowsky, S., & Vanderslott, S. (2020). COVID-19 vaccine hesitancy in the UK: The Oxford coronavirus explanations, attitudes, and narratives survey (Oceans) II. *Psychological Medicine*, 1–15.

Fridman, A., Gershon, R., & Gneezy, A. (2021). COVID-19 and vaccine hesitancy: A longitudinal study. *PloS One*, *16*(4), e0250123.

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1–12.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Garcia, E. (2022, July 13). *White House is teaming up with GLAAD for a monkeypox briefing*. The Independent. https://www.independent.co.uk/news/world/americas/us-politics/white-house-glaad-lgbtq-monkeypox-b2122397.html

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), eabf1234.

Gawronski, B. (2021). Cognitive Sciences. *Trends in Cognitive Sciences*, *25*(9), 723.

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect. *Statistical Modeling, Causal Inference, and Social Science*.

Gest, J., Reny, T., & Mayer, J. (2018). Roots of the radical right: Nostalgic deprivation in the United States and Britain. *Comparative Political Studies*, *51*(13), 1694–1719.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482.

Gollwitzer, A., Martel, C., Brady, W. J., Pärnamets, P., Freedman, I. G., Knowles, E. D., & Van Bavel, J. J. (2020). Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour*, 1–12.

Green, J., Edgerton, J., Naftel, D., Shoub, K., & Cranmer, S. J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, *6*(28), eabc2717.

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

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science*, *65*(4), 1007–1022.

Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, *118*(14).

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020a). 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.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020b). 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.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586.

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

Hagey, K., & Horwitz, J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *The Wall Street Journal*, *16*.

Haidt, J. (2022). Why the Past 10 Years of American Life Have Been Uniquely Stupid. *The Atlantic*, *11*.

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, *19*(4), 405–411.

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

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling Validated Versus Being Correct:A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

Hetherington, M. J. (2001). Resurgent mass partisanship: The role of elite polarization. *American Political Science Review*, 619–631.

Hodes, L. N., & Thomas, K. G. (2021). Smartphone screen time: Inaccuracy of self-reports and influence of psychological and contextual factors. *Computers in Human Behavior*, *115*, 106616.

Hogg, M. A. (2000). Subjective uncertainty reduction through self-categorization: A motivational theory of social identity processes. *European Review of Social Psychology*, *11*(1), 223–255.

Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imhoff, R., Zimmer, F., Klein, O., António, J. H., Babinska, M., Bangerter, A., Bilewicz, M., Blanuša, N., Bovan, K., & Bužarovska, R. (2022). Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour*, 1–12.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2018). *The Origins and Consequences of Affective Polarization in the United States*.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522.

Jakesch, M., Koren, M., Evtushenko, A., & Naaman, M. (2019). The role of source and expressive responding in political news evaluation. *Computation and Journalism Symposium*.

Jans, L., Postmes, T., & Van der Zee, K. I. (2011). The induction of shared identity: The positive role of individual distinctiveness for groups. *Personality and Social Psychology Bulletin*, *37*(8), 1130–1141.

Jardina, A., & Traugott, M. (2019). The Genesis of the Birther Rumor: Partisanship, racial attitudes, and political knowledge. *Journal of Race, Ethnicity and Politics*, *4*(1), 60–80.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., Zheng, M., Manrique, P., Wuchty, S., & Lupu, Y. (2020a). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230–233.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., Zheng, M., Manrique, P., Wuchty, S., & Lupu, Y. (2020b). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230–233.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, *65*(2), 371–388.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003

Joyce, N., & Harwood, J. (2020). Social identity motivations and intergroup media attractiveness. *Group Processes & Intergroup Relations*, *23*(1), 71–90.

Kahan, D. M. (2012). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, *8*, 407–424.

Kahan, D. M. (2015). 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: An Interdisciplinary, Searchable, and Linkable Resource*, 1–16.

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, *4*(3), 465–505.

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science curiosity and political information processing. *Political Psychology*, *38*, 179–199.

Kahan, D. M., Peters, E., Dawson, E., & Slovic, P. (2013). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *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.

Kalla, J. L., & Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review*, *112*(1), 148–166.

Kanno-Youngs, Z., & Kang, C. (2021, July 16). 'They're Killing People': Biden Denounces Social Media for Virus Disinformation. *The New York Times*. https://www.nytimes.com/2021/07/16/us/politics/biden-facebook-social-media-covid.html

Kim, E., & Ihm, J. (2020). More than virality: Online sharing of controversial news with activated audience. *Journalism & Mass Communication Quarterly*, *97*(1), 118–140.

Klein, E. (2020). *Why We're Polarized*. Avid Reader Press.

Kraft, P. W., Krupnikov, Y., Milita, K., Ryan, J. B., & Soroka, S. (2020a). Social Media and the Changing Information Environment: Sentiment Differences in Read Versus Recirculated News Content. *Public Opinion Quarterly*, *84*(S1), 195–215.

Kraft, P. W., Krupnikov, Y., Milita, K., Ryan, J. B., & Soroka, S. (2020b). Social media and the changing information environment: Sentiment differences in read versus recirculated news content. *Public Opinion Quarterly*, *84*(S1), 195–215.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Kreps, S., Dasgupta, N., Brownstein, J. S., Hswen, Y., & Kriner, D. L. (2021). Public attitudes toward COVID-19 vaccination: The role of vaccine attributes, incentives, and misinformation. *Npj Vaccines*, *6*(1), 1–7.

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, *45*(3), 188–206.

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

Lapowski, I. (2018). Newsguard wants to fight fake news with humans, not algorithms. *Wired, August*, *23*.

Larsen, B., Ryan, T. J., Greene, S., Hetherington, M. J., Maxwell, R., & Tadelis, S. (2022). *Counter-stereotypical Messaging and Partisan Cues: Moving the Needle on Vaccines in a Polarized US*.

Lawler, R. (2022, June 23). *Meta reportedly plans to shut down CrowdTangle, its tool that tracks popular social media posts*. The Verge. https://www.theverge.com/2022/6/23/23180357/meta-crowdtangle-shut-down-facebook-misinformation-viral-news-tracker

Lawson, M. A., & Kakkar, H. (2021). Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General*.

Lazer, D., Green, J., Ognyanova, K., Baum, M., Lin, J., Druckman, J., Perlis, R. H., Santillana, M., & Uslu, A. (2021). *The COVID States Project #57: Social media news consumption and COVID-19 vaccination rates*. OSF Preprints. https://doi.org/10.31219/osf.io/uvqbs

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200143.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, *80*(S1), 392–410.

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. *The Journal of Politics*, *79*(2), 485–501.

Levy, R. (2021a). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–870.

Levy, R. (2021b). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–870.

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.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021a). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021b). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 1–12.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021c). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348.

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*, 1–8. https://doi.org/10.1038/s41562-020-0889-7

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2021). *Digital Media and Democracy: A Systematic Review of Causal and Correlational Evidence Worldwide*.

Lovakov, A., & Agadullina, E. R. (n.d.). *Empirically Derived Guidelines for Effect Size Interpretation in Social Psychology*.

MacDonald, N. E. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, *33*(34), 4161–4164.

Mackie, D. M., Maitner, A. T., & Smith, E. R. (2016). *Intergroup emotions theory.*

Madsen, J. K., Bailey, R. M., & Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports*, *8*(1), 1–8.

Maertens, R., Götz, F., Schneider, C. R., Roozenbeek, J., Kerr, J. R., Stieger, S., McClanahan III, W. P., Drabot, K., & van der Linden, S. (2021). *The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment*.

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*.

Masullo, G. (n.d.). *Bridging Political Divides with Facebook Memes—Center for Media Engagement—Center for Media Engagement*. Retrieved May 9, 2022, from https://mediaengagement.org/research/bridging-political-divides-with-facebook-memes/

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001a). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001b). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., & Hopkins, E. J. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868.

Melnikoff, D. E., & Strohminger, N. (2020). The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour*, *4*(12), 1258–1264.

Melnyk, V., van Herpen, E., & Trijp, H. (2010). The influence of social norms in consumer decision making: A meta-analysis. *ACR North American Advances*.

Mernyk, J., Pink, S., Druckman, J., & Willer, R. (2021). *Correcting Inaccurate Metaperceptions Reduces Americans' Support for Partisan Violence*.

Messing, S., & Weisel, R. (2017). Partisan Conflict and Congressional Outreach. *Pew Research Center Report*.

Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13642–13649.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021a). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*, *118*(7).

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021b). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*, *118*(7), e2022761118. https://doi.org/10.1073/pnas.2022761118

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, *12*(1), 1–10.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos One*, *15*(2), e0228882.

Mosleh, M., & Rand, D. (2021). *Falsehood in, falsehood out: A tool for measuring exposure to elite misinformation on Twitter*.

Mosseri, A. (2018). *News feed FYI: Bringing people closer together. Facebook Newsroom*.

Motta, M., Stecula, D., & Farhart, C. (2020). How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, *53*(2), 335–342.

Mummendey, A., & Otten, S. (1998). Positive–negative asymmetry in social discrimination. *European Review of Social Psychology*, *9*(1), 107–143.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126.

Nicholson, S. P. (2012). Polarizing cues. *American Journal of Political Science*, *56*(1), 52–66.

Orvell, A., Kross, E., & Gelman, S. A. (2020). "You" speaks to me: Effects of generic-you in creating resonance between people and ideas. *Proceedings of the National Academy of Sciences*, *117*(49), 31038–31045.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2020). Partisan polarization is the primary psychological motivation behind "fake news" sharing on Twitter. *American Political Science Review*, 1–17.

Osmundsen, M., Petersen, M. B., Mazepus, H., Toshkov, D., & Dimitrova, A. (2022). *Information battleground: Conflict perceptions motivate the belief in and sharing of fake news about the adversary*.

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614.

Panizza, F., Ronzani, P., Mattavelli, S., Morisseau, T., Martini, C., & Motterlini, M. (2021). *Lateral reading and monetary incentives to sort out scientific (dis) information*.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Parker, M. T., & Janoff-Bulman, R. (2013). Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love." *Social Justice Research*, *26*(1), 81–96.

Pavlović, T., Azevedo, F., De, K., Riaño-Moreno, J. C., Maglić, M., Gkinopoulos, T., Donnelly-Kehoe, P. A., Payán-Gómez, C., Huang, G., & Kantorowicz, J. (2022). Predicting attitudinal and behavioral responses to COVID-19 pandemic using machine learning. *PNAS Nexus*.

Pechar, E., & Kranton, R. (2017). Moderators of intergroup discrimination in the minimal group paradigm: A meta-analysis. *Semantic Scholar*.

Peer, E., Rothschild, D. M., Evernden, Z., Gordon, A., & Damer, E. (2021). MTurk, Prolific or panels? Choosing the right audience for online research. *Choosing the Right Audience for Online Research (January 10, 2021)*.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. (2020). *A practical guide to doing behavioural research on fake news and misinformation*.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018a). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865.

Pennycook, G., Cannon, T., & Rand, D. G. (2018b). *Prior exposure increases perceived accuracy of fake news*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021a). Shifting attention to accuracy can reduce misinformation online. *Nature*, 1–6.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021b). Shifting attention to accuracy can reduce misinformation online. *Nature*, 1–6. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2021). Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning. *Personality and Social Psychology Bulletin*, 01461672211023652.

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. (2021a). *Reducing the spread of fake news by shifting attention to accuracy: Meta-analytic evidence of replicability and generalizability*.

Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*. https://doi.org/10.1016/j.cognition.2018.06.011

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.

Pennycook, G., & Rand, D. G. (2021b). Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review*.

Pennycook, G., & Rand, D. G. (2021c). The psychology of fake news. *Trends in Cognitive Sciences*.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), 1–12.

Pereira, A., Harris, E. A., & Van Bavel, J. J. (2018). *Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news*.

Pereira, A., & Van Bavel, J. (2018). *Identity concerns drive belief in fake news*.

Persily, N., & Tucker, J. A. (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.

Persson, E., Andersson, D., Koppel, L., Västfjäll, D., & Tinghög, G. (2021). A preregistered replication of motivated numeracy. *Cognition*, *214*, 104768.

Petersen, M. B. (2020). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. *Current Opinion in Psychology*, *35*, 71–75.

Petersen, M. B., Osmundsen, M., & Tooby, J. (2020). The Evolutionary Psychology of Conflict and the Functions of Falsehood. *PsyArXiv. August*, *29*.

Peterson, E., & Iyengar, S. (2021). Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading? *American Journal of Political Science*, *65*(1), 133–147.

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*, *12*(1), 1–7.

Pink, S., Chu, J., Druckman, J., Rand, D., & Willer, R. (2021). Elite Party Cues Increase Vaccination Intentions among Republicans. *Proceedings of the National Academy of Sciences*.

Pink, S. L., Chu, J., Druckman, J. N., Rand, D. G., & Willer, R. (2021). Elite party cues increase vaccination intentions among Republicans. *Proceedings of the National Academy of Sciences*, *118*(32), e2106559118.

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2022). *The Misleading count: An identity-based intervention to mitigate the spread of partisan misinformation*.

Pretus, C., Van Bavel, J. J., Brady, W. J., Harris, E. A., Vilarroya, O., & Servin, C. (2021). *The role of political devotion in sharing partisan misinformation*.

Prior, M., Sood, G., & Khanna, K. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science*, *10*(4), 489–518.

Pröllochs, N., & Feuerriegel, S. (2022). Mechanisms of True and False Rumor Sharing in Social Media: Wisdom-of-Crowds or Herd Behavior? *ArXiv Preprint ArXiv:2207.03020*.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106.

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. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Rathje, S. (2022). Letter to the editors of Psychological Science: Meta-analysis reveals that accuracy nudges have little to no effect for US conservatives: Regarding Pennycook et al. (2020). *Psychological Science*.

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

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

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021c). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Relations, U. of O. I. of G., & Sherif, M. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment* (Vol. 10). University Book Exchange Norman, OK.

Ren, Z. B., Dimant, E., & Schweitzer, M. E. (2021). Social Motives for Sharing Conspiracy Theories. *Eugen and Schweitzer, Maurice E., Social Motives for Sharing Conspiracy Theories (September 8, 2021)*.

Resnick, P., Ovadya, A., & Gilchrist, G. (2018). Iffy quotient: A platform health metric for misinformation. *Center for Social Media Responsibility*, *17*.

Robertson, C. E., Pretus, C., Rathje, S., Harris, E., & Van Bavel, J. J. (2022). How Social Identity Shapes Conspiratorial Belief. *Current Opinion in Psychology*, 101423.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Parnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2022). *Negativity drives online news consumption*.

Robertson, C., Pretus, C., Rathje, S., Harris, E. A., & Van Bavel, J. J. (2022). *How Social Identity Shapes Conspiratorial Belief*.

Rogers, N., & Jones, J. J. (2021). Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? *Journal of Social Computing*, *2*(1), 1–13.

Romer, D., & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social Science & Medicine*, *263*, 113356.

Roose, K., Isaac, M., & Frenkel, S. (2020, November 24). Facebook Struggles to Balance Civility and Growth. *The New York Times*. https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

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., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R. H., Sultan, M., & van der Linden, S. (2022a). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547–573.

Roozenbeek, J., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S. (2022b). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3).

Roozenbeek, J., Sander, van der L., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), 201199.

Roozenbeek, J., & van der Linden, S. (2019a). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Roozenbeek, J., & van der Linden, S. (2019b). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Rubin, M., & Hewstone, M. (2016). Social Identity Theory's Self-Esteem Hypothesis: A Review and Some Suggestions for Clarification: *Personality and Social Psychology Review*. https://doi.org/10.1207/s15327957pspr0201_3

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., & Bursalıoğlu, A. (2021). The general fault in our fault lines. *Nature Human Behaviour*, *5*(10), 1369–1380.

Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly*, *82*(1), 135–147.

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affective Science*, *2*(4), 379–390.

Seetharaman, J. H. and D. (2020, May 26). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. *Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255.

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 63–76.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892. https://doi.org/10.1073/pnas.1908369116

Speckmann, F., & Unkelbach, C. (2022). Monetary incentives do not reduce the repetition-induced truth effect. *Psychonomic Bulletin & Review*, *29*(3), 1045–1052.

Stancato, D. M., & Keltner, D. (2021). Awe, ideological conviction, and perceptions of ideological opponents. *Emotion*, *21*(1), 61.

Statista. (2022a). *Daily social media usage worldwide*. Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2022b). Number of social network users worldwide from 2017 to 2025. *Retrieved June*, *4*, 2020.

Stolberg, S. G., & Alba, D. (2021, July 15). Surgeon General Assails Tech Companies Over Misinformation on Covid-19. *The New York Times*. https://www.nytimes.com/2021/07/15/us/politics/surgeon-general-vaccine-misinformation.html

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

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979a). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, 56–65.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979b). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, *56*, 65.

Tappin, B. M., Berinsky, A. J., & Rand, D. G. (2022). *Exposure to Persuasive Messaging Changes Partisan Attitudes Even in the Face of Countervailing Leader Cues*.

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

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020b). 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.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Thompson, D. (2021). Millions are saying no to the vaccines. What are they thinking? *The Atlantic*.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, *26*(1), 1–9.

Traberg, C. S. (2022). Misinformation: Broaden definition to curb its societal influence. *Nature*, *606*(7915), 653–653.

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*, 111269.

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology*, *9*(2), 187–204.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Basil Blackwell.

Valente, T. W. (2012). Network interventions. *Science*, *337*(6090), 49–53.

Van Bavel, J. J., Cichocka, A., Capraro, V., Sjåstad, H., Nezlek, J. B., Pavlović, T., Alfano, M., Gelfand, M. J., Azevedo, F., & Birtel, M. D. (2022). National identity predicts public health support during a global pandemic. *Nature Communications*, *13*(1), 1–14.

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113.

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

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021a). How social media shapes polarization. *Trends in Cognitive Sciences*.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021b). How social media shapes polarization. *Trends in Cognitive Sciences*.

Van Der Linden, S. (2017). The nature of viral altruism and how to make it stick. *Nature Human Behaviour*, *1*(3), 1–4.

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 1–8. https://doi.org/10.1038/s41591-022-01713-6

van der Linden, S., Leiserowitz, A., & Maibach, E. (2018). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, *2*(1), 2–3.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017a). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017b). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2).

van der Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The paranoid style in American politics revisited: An ideological asymmetry in conspiratorial thinking. *Political Psychology*, *42*(1), 23–51.

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society*, *42*(3), 460–470.

van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*, 2928.

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How Can Psychological Science Help Counter the Spread of Fake News? *The Spanish Journal of Psychology*, *24*.

Van Dijck, J. (2013). 'You have one identity': Performing the self on Facebook and LinkedIn. *Media, Culture & Society*, *35*(2), 199–215.

Van Dijcke, D., & Wright, A. L. (2021). Profiling Insurrection: Characterizing Collective Action Using Mobile Device Data. *Available at SSRN 3776854*.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12.

Vegetti, F., & Mancosu, M. (2020). The impact of political sophistication and motivated reasoning on misinformation. *Political Communication*, *37*(5), 678–695.

Verduyn, P., Lee, D. S., Park, J., Shablack, H., Orvell, A., Bayer, J., Ybarra, O., Jonides, J., & Kross, E. (2015). Passive Facebook usage undermines affective well-being: Experimental and longitudinal evidence. *Journal of Experimental Psychology: General*, *144*(2), 480.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wagner, K. (2019, March 8). *Inside Twitter's ambitious plan to clean up its platform*. Vox. https://www.vox.com/2019/3/8/18245536/exclusive-twitter-healthy-conversations-dunking-research-product-incentives

Walker, P., & correspondent, P. W. P. (2020). Boris Johnson says "anti-vaxxers are nuts." *The Guardian*. https://www.theguardian.com/society/2020/jul/24/boris-johnson-says-anti-vaxxers-are-nuts-free-winter-flu-jabs

Waller, I., & Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature*, *600*(7888), 264–268.

Wang, S.-Y. N., & Inbar, Y. (2020). Moral-Language Use by US Political Elites. *Psychological Science*, 0956797620960397.

Wetzels, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2014). Bayesian analysis. *The Encyclopedia of Clinical Psychology*, 1–11.

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2021). *Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter*.

Wojcieszak, M., Winter, S., & Yu, X. (2020). Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. *Mass Communication and Society*, *23*(4), 455–483.

Yaden, D. B., Iwry, J., Slack, K. J., Eichstaedt, J. C., Zhao, Y., Vaillant, G. E., & Newberg, A. B. (2016). The overview effect: Awe and self-transcendent experience in space flight. *Psychology of Consciousness: Theory, Research, and Practice*, *3*(1), 1.

Yu, X., Wojcieszak, M., & Casas, A. (2021). *Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users*.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676.

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PloS One*, *12*(7), e0181821.

8. Supplementary Materials for "Out-group animosity drives engagement on social media"

Table S1. Full Regression Models for Study 1.

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	38.04 *** [36.71, 39.42]	68.68 *** [66.25, 71.20]	41.96 *** [30.85, 57.08]	19.75 *** [17.74, 22.00]
Democrat	1.10 ***	1.29 ***	1.00	1.35 ***
	[1.09, 1.12]	[1.26, 1.31]	[0.99, 1.00]	[1.34, 1.36]
Republican	1.46 ***	1.23 ***	1.57 ***	1.37 ***
	[1.44, 1.48]	[1.20, 1.26]	[1.55, 1.58]	[1.35, 1.38]
NegativeAffect	1.08 ***	1.08 ***	1.05 ***	0.98 ***
	[1.07, 1.09]	[1.06, 1.09]	[1.05, 1.06]	[0.98, 0.99]
PositiveAffect	0.89 ***	0.98 **	0.90 ***	0.94 ***
	[0.88, 0.90]	[0.97, 0.99]	[0.89, 0.90]	[0.93, 0.94]
MoralEmotional	1.17 ***	1.10 ***	1.10 ***	1.17 ***
	[1.16, 1.19]	[1.07, 1.13]	[1.09, 1.11]	[1.16, 1.19]
has_mediaTRUE	1.47 ***	0.54 ***	2.83 ***	5.18 ***
	[1.44, 1.51]	[0.53, 0.55]	[2.08, 3.85]	[4.64, 5.77]
has_URLTRUE	0.73 ***	0.38 ***	1.80 ***	5.04 ***
	[0.70, 0.75]	[0.37, 0.40]	[1.32, 2.44]	[4.53, 5.61]
followers_count	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]		
is_retweetTRUE	1.54 ***	0.79 ***		
	[1.48, 1.60]	[0.76, 0.82]		
`Likes at Posting`			1.00 ***	1.00 ***
			[1.00, 1.00]	[1.00, 1.00]
N	143702	83527	300000	299999
AIC	489760.05	294339.53	1142251.76	1206259.31
BIC	489868.68	294442.19	1142357.88	1206365.43
Pseudo R2	0.19	0.17	0.15	0.10

Table S2. VIFS for Study 1.

	Twitter		Facebook	
	Conservative	Liberal	Conservative	Liberal
	VIF	VIF	VIF	VIF
Democrat	1.02	1.02	1.04	1.02
Republican	1.01	1.01	1.03	1.02
NegativeAffect	1.3	1.24	1.41	1.32
PositiveAffect	1.11	1.07	1.07	1.06
MoralEmotional	1.37	1.29	1.46	1.38
has_media	1.09	1.11	55	370.06
has_URL	1.6	2.74	55.52	370.08
followers_count	1.02	1.01		
is_retweet	1.65	2.72		
Likes at Posting			1.29	1.01

Note. Variance Inflation Factors (VIFS) for study 1. It should be noted that some VIFs are extremely high. This is because, in some cases, these variables are almost perfectly inversely correlated (e.g., if a Facebook post did not have a URL, it had media). These high VIFS were for control variables, not for key variables. We ran the model without control variables and the results were similar (Table S3) and the results were less problematic (Table S5).

Table S3. Study 1 Regression Models Without Control Variables

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	13.50 *** [13.44, 13.55]	8.68 *** [8.64, 8.72]	80.90 *** [80.40, 81.40]	99.70 *** [99.03, 100.37]
Democrat	1.04 ***	3.27 ***	0.97 ***	1.23 ***
	[1.04, 1.05]	[3.23, 3.32]	[0.97, 0.98]	[1.22, 1.24]
Republican	2.41 ***	1.26 ***	1.47 ***	1.27 ***
	[2.38, 2.43]	[1.25, 1.27]	[1.46, 1.48]	[1.26, 1.29]
NegativeAffect	1.36 ***	1.47 ***	1.04 ***	0.98 ***
	[1.35, 1.37]	[1.46, 1.48]	[1.03, 1.04]	[0.98, 0.99]
PositiveAffect	0.91 ***	1.00	0.89 ***	0.92 ***
	[0.91, 0.92]	[1.00, 1.01]	[0.89, 0.90]	[0.92, 0.93]
MoralEmotional	1.11 ***	1.07 ***	1.12 ***	1.18 ***
	[1.10, 1.12]	[1.06, 1.08]	[1.10, 1.13]	[1.17, 1.20]
N	747675	611292	300000	299999
AIC	3069655.63	2530345.65	1179454.34	1232001.33
BIC	3069736.30	2530424.91	1179528.62	1232075.61
Pseudo R2	0.08	0.09	0.03	0.02

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S4. VIFS for Study 1 Regression Without Control Variables

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
	VIF	VIF	VIF	VIF
Democrat	1.01	1.01	1.01	1.02
Republican	1.01	1.01	1.01	1.02
NegativeAffect	1.24	1.29	1.32	1.41
PositiveAffect	1.06	1.08	1.06	1.07
MoralEmotional	1.29	1.37	1.38	1.46

Table S5. Study 1 Robustness Check (Cluster Robust Standard Errors)

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	38.04 ***	68.68 ***	41.96 ***	19.75 ***
	[36.52, 39.62]	[65.73, 71.76]	[30.12, 58.46]	[17.51, 22.29]
Democrat	1.10 ***	1.29 ***	1.00	1.35 ***
	[1.08, 1.12]	[1.26, 1.32]	[0.99, 1.00]	[1.34, 1.36]
Republican	1.46 ***	1.23 ***	1.57 ***	1.37 ***
	[1.43, 1.48]	[1.20, 1.26]	[1.55, 1.58]	[1.35, 1.38]
NegativeAffect	1.08 ***	1.08 ***	1.05 ***	0.98 ***
	[1.07, 1.09]	[1.06, 1.09]	[1.05, 1.06]	[0.98, 0.99]
PositiveAffect	0.89 ***	0.98 **	0.90 ***	0.94 ***
	[0.88, 0.90]	[0.97, 0.99]	[0.89, 0.90]	[0.93, 0.94]
MoralEmotional	1.17 ***	1.10 ***	1.10 ***	1.17 ***
	[1.16, 1.19]	[1.07, 1.13]	[1.09, 1.11]	[1.15, 1.20]
has_mediaTRUE	1.47 ***	0.54 ***	2.83 ***	5.18 ***
	[1.44, 1.51]	[0.53, 0.55]	[2.03, 3.94]	[4.58, 5.85]
has_URLTRUE	0.73 ***	0.38 ***	1.80 ***	5.04 ***
	[0.70, 0.76]	[0.37, 0.40]	[1.29, 2.50]	[4.47, 5.69]
followers_count	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]		
is_retweetTRUE	1.54 ***	0.79 ***		
	[1.48, 1.61]	[0.76, 0.82]		
`Likes at Posting`			1.00 ***	1.00 ***
			[1.00, 1.00]	[1.00, 1.00]
N	143702	83527	300000	299999
AIC	489760.05	294339.53	1142251.76	1206259.31
BIC	489868.68	294442.19	1142357.88	1206365.43
Pseudo R2	0.19	0.17	0.15	0.10

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S5. Study 1 Relative Importance Analysis

	Twitter		Facebook	
	Conservative	Liberal	Conservative	Liberal
Democrat	**0.00734**	0.0005	**0.01082**	0.000076
Republican	0.00439	**0.0095**	0.01014	**0.026482**
NegativeAffect	0.0023	0.0027	0.00032	0.001931
PositiveAffect	0.00042	0.0032	0.00074	0.004462
MoralEmotional	0.00149	0.0026	0.00204	0.001477
has_media	0.02948	0.0073	0.00763	0.010908
has_URL	0.03801	0.0204		
followers_count	0.07837	0.1384		
Likes at Posting			0.06433	0.098183

Note. "lmg" (or estimated R^2) values are shown for each regression mode

Table S6. Liberal Media Reactions Regression Analysis

	Shares	Comments	Likes	Love	Wow	Haha	Sad	Angry	Retweet	Favorites
(Intercept)	41.96 ***	130.60 ***	191.46 ***	12.93 ***	12.59 ***	13.10 ***	12.40 ***	10.44 ***	38.04 ***	80.08 ***
	[30.85, 57.08]	[97.25, 175.37]	[146.85, 249.63]	[9.18, 18.21]	[9.32, 17.00]	[9.60, 17.87]	[8.43, 18.24]	[6.97, 15.64]	[36.71, 39.42]	[77.23, 83.02]
Democrat	1.00	1.57 ***	1.31 ***	1.66 ***	0.91 ***	1.72 ***	0.87 ***	1.15 ***	1.10 ***	1.25 ***
	[0.99, 1.00]	[1.56, 1.58]	[1.30, 1.32]	[1.64, 1.68]	[0.90, 0.92]	[1.71, 1.74]	[0.86, 0.88]	[1.13, 1.16]	[1.09, 1.12]	[1.23, 1.27]
Republican	1.57 ***	2.23 ***	1.45 ***	1.44 ***	1.54 ***	2.92 ***	1.35 ***	3.33 ***	1.46 ***	1.22 ***
	[1.55, 1.58]	[2.21, 2.24]	[1.44, 1.46]	[1.42, 1.45]	[1.52, 1.55]	[2.90, 2.95]	[1.33, 1.36]	[3.30, 3.37]	[1.44, 1.48]	[1.20, 1.24]
NegativeAffect	1.05 ***	1.05 ***	0.99 ***	0.88 ***	1.06 ***	1.01 ***	1.31 ***	1.18 ***	1.08 ***	1.04 ***
	[1.05, 1.06]	[1.05, 1.06]	[0.98, 0.99]	[0.87, 0.88]	[1.05, 1.06]	[1.01, 1.02]	[1.30, 1.32]	[1.17, 1.18]	[1.07, 1.09]	[1.03, 1.05]
PositiveAffect	0.90 ***	0.86 ***	0.98 ***	1.12 ***	0.79 ***	0.92 ***	0.77 ***	0.78 ***	0.89 ***	0.95 ***
	[0.89, 0.90]	[0.86, 0.87]	[0.98, 0.99]	[1.11, 1.13]	[0.79, 0.79]	[0.92, 0.93]	[0.76, 0.77]	[0.77, 0.78]	[0.88, 0.90]	[0.95, 0.96]
MoralEmotional	1.10 ***	1.07 ***	1.05 ***	1.05 ***	1.10 ***	0.91 ***	1.22 ***	1.23 ***	1.17 ***	1.13 ***
	[1.09, 1.11]	[1.06, 1.08]	[1.04, 1.05]	[1.04, 1.06]	[1.09, 1.11]	[0.90, 0.92]	[1.20, 1.23]	[1.21, 1.25]	[1.16, 1.19]	[1.11, 1.15]
has_URLTRUE	1.80 ***	0.78	1.00	0.79	1.09	0.97	1.04	1.14	0.73 ***	0.80 ***
	[1.32, 2.44]	[0.58, 1.05]	[0.76, 1.30]	[0.56, 1.11]	[0.80, 1.47]	[0.71, 1.33]	[0.71, 1.53]	[0.76, 1.71]	[0.70, 0.75]	[0.77, 0.83]
has_mediaTRUE	2.83 ***	1.30	1.68 ***	2.33 ***	1.30	1.27	1.16	1.12	1.47 ***	1.66 ***
	[2.08, 3.85]	[0.97, 1.74]	[1.29, 2.19]	[1.66, 3.29]	[0.96, 1.76]	[0.93, 1.73]	[0.79, 1.70]	[0.75, 1.68]	[1.44, 1.51]	[1.62, 1.70]
`Likes at Posting`	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]		
followers_count									1.00 ***	1.00 ***
									[1.00, 1.00]	[1.00, 1.00]
is_retweetTRUE									1.54 ***	0.01 ***
									[1.48, 1.60]	[0.01, 0.01]
N	300000	300000	300000	300000	300000	300000	300000	300000	143702	143702
AIC	1142251.76	1116558.15	1053241.80	1206829.92	1127482.48	1148172.38	1277786.18	1305894.87	489760.05	494589.33
BIC	1142357.88	1116664.26	1053347.91	1206936.04	1127588.59	1148278.49	1277892.30	1306000.98	489868.68	494697.97

Pseudo R2	0.15	0.26	0.22	0.19	0.17	0.26	0.11	0.17	0.19	0.54

Table S7. Conservative Media Reactions Regression Analysis

	Shares	Comments	Likes	Love	Wow	Haha	Sad	Angry	Retweet	Favorites
(Intercept)	19.75 ***	28.89 ***	71.18 ***	11.15 ***	3.26 ***	3.98 ***	3.15 ***	3.17 ***	38.04 ***	80.08 ***
	[17.74, 22.00]	[26.09, 32.00]	[64.79, 78.19]	[9.94, 12.51]	[2.97, 3.59]	[3.53, 4.49]	[2.82, 3.52]	[2.76, 3.65]	[36.71, 39.42]	[77.23, 83.02]
Democrat	1.35 ***	1.59 ***	1.18 ***	1.08 ***	1.27 ***	2.47 ***	1.16 ***	1.83 ***	1.10 ***	1.25 ***
	[1.34, 1.36]	[1.58, 1.60]	[1.17, 1.19]	[1.06, 1.09]	[1.26, 1.28]	[2.45, 2.50]	[1.15, 1.17]	[1.81, 1.85]	[1.09, 1.12]	[1.23, 1.27]
Republican	1.37 ***	1.81 ***	1.77 ***	2.26 ***	0.99	1.59 ***	0.96 ***	1.47 ***	1.46 ***	1.22 ***
	[1.35, 1.38]	[1.80, 1.83]	[1.76, 1.79]	[2.24, 2.29]	[0.98, 1.00]	[1.57, 1.61]	[0.95, 0.97]	[1.45, 1.49]	[1.44, 1.48]	[1.20, 1.24]
NegativeAffect	0.98 ***	0.95 ***	0.90 ***	0.79 ***	1.01 **	0.93 ***	1.21 ***	1.09 ***	1.08 ***	1.04 ***
	[0.98, 0.99]	[0.94, 0.95]	[0.90, 0.91]	[0.79, 0.80]	[1.00, 1.02]	[0.92, 0.94]	[1.20, 1.22]	[1.08, 1.10]	[1.07, 1.09]	[1.03, 1.05]
PositiveAffect	0.94 ***	0.89 ***	1.05 ***	1.21 ***	0.83 ***	0.92 ***	0.82 ***	0.75 ***	0.89 ***	0.95 ***
	[0.93, 0.94]	[0.88, 0.89]	[1.04, 1.06]	[1.20, 1.21]	[0.82, 0.83]	[0.91, 0.92]	[0.81, 0.82]	[0.74, 0.75]	[0.88, 0.90]	[0.95, 0.96]
MoralEmotional	1.17 ***	1.13 ***	1.12 ***	1.14 ***	1.10 ***	0.91 ***	1.20 ***	1.27 ***	1.17 ***	1.13 ***
	[1.16, 1.19]	[1.12, 1.14]	[1.11, 1.13]	[1.13, 1.15]	[1.09, 1.11]	[0.89, 0.92]	[1.18, 1.21]	[1.25, 1.29]	[1.16, 1.19]	[1.11, 1.15]
has_URLTRUE	5.04 ***	5.59 ***	3.23 ***	0.97	6.02 ***	7.45 ***	3.48 ***	6.98 ***	0.73 ***	0.80 ***
	[4.53, 5.61]	[5.04, 6.19]	[2.94, 3.55]	[0.86, 1.08]	[5.47, 6.62]	[6.60, 8.41]	[3.12, 3.89]	[6.06, 8.03]	[0.70, 0.75]	[0.77, 0.83]
has_mediaTRUE	5.18 ***	5.14 ***	4.04 ***	1.75 ***	3.64 ***	6.75 ***	2.13 ***	4.20 ***	1.47 ***	1.66 ***
	[4.64, 5.77]	[4.63, 5.70]	[3.68, 4.45]	[1.56, 1.97]	[3.30, 4.00]	[5.98, 7.63]	[1.91, 2.38]	[3.64, 4.83]	[1.44, 1.51]	[1.62, 1.70]
`Likes at Posting`	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]		
followers_count									1.00 ***	1.00 ***
									[1.00, 1.00]	[1.00, 1.00]
is_retweetTRUE									1.54 ***	0.01 ***
									[1.48, 1.60]	[0.01, 0.01]
N	299999	299999	299999	299999	299999	299999	299999	299999	143702	143702
AIC	1206259.31	1175256.73	1125262.30	1246255.47	1132021.07	1275974.00	1219208.80	1365334.17	489760.05	494589.33
BIC	1206365.43	1175362.84	1125368.41	1246361.59	1132127.18	1276080.12	1219314.91	1365440.28	489868.68	494697.97
Pseudo R2	0.10	0.18	0.17	0.18	0.14	0.15	0.13	0.08	0.19	0.54

Table S8. Study 1 Descriptive Statistics

	Conservative Twitter		Liberal Twitter		Conservative Facebook		Liberal Facebook	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Democrat	0.18	0.50	0.13	0.45	0.20	0.68	0.31	0.74
2. Republican	0.17	0.44	0.15	0.42	0.27	0.69	0.29	0.65
3. Positive Affect	0.42	0.73	0.55	0.82	0.81	1.14	0.67	0.99
4. Negative Affect	0.41	0.70	0.52	0.79	0.82	1.13	0.76	1.04
5. Moral Emotional	0.18	0.45	0.21	0.49	0.33	0.70	0.31	0.67

Note. Means and standard deviations for each of the language categories in each dataset.

Table S9. Liberal Media Facebook Reactions

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Shares	491.17	3984.80							
2. Likes	864.71	4124.87	.63** [.62, .63]						
3. Comments	410.85	1009.18	.43** [.43, .44]	.49** [.48, .49]					
4. Love	169.58	1290.41	.32** [.32, .32]	.79** [.78, .79]	.32** [.31, .32]				
5. Wow	82.22	686.84	.51** [.50, .51]	.45** [.44, .45]	.33** [.32, .33]	.17** [.16, .17]			
6. Haha	106.47	571.40	.16** [.16, .17]	.22** [.21, .22]	.43** [.43, .44]	.14** [.13, .14]	.12** [.12, .13]		
7. Sad	181.40	1343.43	.42** [.42, .42]	.28** [.28, .29]	.26** [.25, .26]	.10** [.09, .10]	.23** [.22, .23]	.02** [.01, .02]	
8. Angry	231.56	1123.14	.22** [.22, .23]	.07** [.07, .08]	.46** [.46, .47]	.00** [.00, .01]	.18** [.17, .18]	.14** [.14, .15]	.13** [.13, .14]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

Table S10. Conservative Media Correlations and Descriptive Statistics

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Shares	755.51	4580.27							
2. Likes	1491.57	7194.08	.55** [.55, .55]						
3. Comments	866.29	3159.10	.34** [.34, .34]	.38** [.37, .38]					
4. Love	264.84	2143.31	.49** [.49, .49]	.84** [.84, .84]	.31** [.31, .31]				
5. Wow	89.36	435.82	.48** [.48, .48]	.33** [.32, .33]	.29** [.28, .29]	.27** [.27, .28]			
6. Haha	367.68	2074.96	.26** [.26, .27]	.17** [.17, .17]	.55** [.55, .55]	.10** [.10, .11]	.19** [.19, .19]		
7. Sad	138.69	1447.45	.26** [.26, .27]	.09** [.09, .10]	.13** [.13, .13]	.06** [.06, .07]	.18** [.18, .18]	.01** [.01, .02]	
8. Angry	414.01	2054.33	.21** [.21, .22]	.05** [.05, .06]	.60** [.60, .61]	.01** [.01, .01]	.29** [.29, .30]	.29** [.29, .30]	.14** [.14, .15]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$

Table S11. Study 2 Regression Models

	Facebook		Twitter	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	8.79 ***	7.83 ***	9.70 ***	6.25 ***
	[8.65, 8.94]	[7.71, 7.95]	[9.63, 9.78]	[6.20, 6.30]
Democrat	1.02 ***	1.65 ***	0.75 ***	2.80 ***
	[1.01, 1.03]	[1.64, 1.67]	[0.75, 0.75]	[2.77, 2.84]
Republican	1.58 ***	1.20 ***	2.13 ***	0.85 ***
	[1.57, 1.59]	[1.19, 1.20]	[2.11, 2.15]	[0.84, 0.85]
NegativeAffect	1.14 ***	1.12 ***	1.33 ***	1.45 ***
	[1.13, 1.14]	[1.11, 1.12]	[1.33, 1.34]	[1.44, 1.45]
PositiveAffect	0.96 ***	0.98 ***	0.95 ***	1.04 ***
	[0.95, 0.96]	[0.97, 0.98]	[0.95, 0.95]	[1.04, 1.05]
MoralEmotional	1.06 ***	1.05 ***	1.10 ***	1.06 ***
	[1.05, 1.06]	[1.04, 1.05]	[1.10, 1.11]	[1.05, 1.07]
has_URLTRUE	1.24 ***	1.00	0.95 ***	0.83 ***
	[1.21, 1.26]	[0.98, 1.01]	[0.94, 0.96]	[0.82, 0.83]
has_mediaTRUE	1.09 ***	0.92 ***	1.08 ***	1.22 ***
	[1.08, 1.11]	[0.91, 0.94]	[1.07, 1.09]	[1.20, 1.23]
`Likes at Posting`	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]		
followers_count			1.00 ***	1.00 ***
			[1.00, 1.00]	[1.00, 1.00]
is_retweetTRUE			5.10 ***	5.36 ***
			[5.04, 5.16]	[5.30, 5.42]
N	354814	410313	747675	611292
AIC	1253213.12	1474731.77	2861220.18	2343833.57
BIC	1253320.92	1474841.02	2861346.95	2343958.13
Pseudo R2	0.33	0.13	0.30	0.33

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S12. VIFS to Study 2

	Twitter		Facebook	
	Conservative	Liberal	Conservative	Liberal
	VIF	VIF	VIF	VIF
Democrat	1.05	1.13	1.09	1.02
Republican	1.14	1.03	1.08	1.1
NegativeAffect	1.23	1.32	1.63	1.52
PositiveAffect	1.24	1.22	1.41	1.36
MoralEmotional	1.4	1.47	1.95	1.8
has_media	1.33	1.33	3.27	3.67
has_URL	1.38	1.37	3.32	3.71
followers_count	1.01	1.01		
is_retweet	1.34	1.42		
Likes at Posting			1	1.01

Note. Variance Inflation Factors (VIFS) for study 2.

Table S13. Study 2 Models Without Control Variables

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	31.36 *** [31.13, 31.60]	23.97 *** [23.72, 24.22]	80.90 *** [80.40, 81.40]	99.70 *** [99.03, 100.37]
Democrat	1.03 **	1.29 ***	0.97 ***	1.23 ***
	[1.01, 1.05]	[1.26, 1.32]	[0.97, 0.98]	[1.22, 1.24]
Republican	1.35 ***	1.26 ***	1.47 ***	1.27 ***
	[1.32, 1.37]	[1.23, 1.29]	[1.46, 1.48]	[1.26, 1.29]
NegativeAffect	1.08 ***	1.10 ***	1.04 ***	0.98 ***
	[1.06, 1.09]	[1.08, 1.12]	[1.03, 1.04]	[0.98, 0.99]
PositiveAffect	0.89 ***	1.04 ***	0.89 ***	0.92 ***
	[0.89, 0.90]	[1.02, 1.05]	[0.89, 0.90]	[0.92, 0.93]
MoralEmotional	1.17 ***	1.11 ***	1.12 ***	1.18 ***
	[1.15, 1.19]	[1.08, 1.14]	[1.10, 1.13]	[1.17, 1.20]
N	143702	83527	300000	299999
AIC	517325.84	307824.60	1179454.34	1232001.33
BIC	517394.97	307889.93	1179528.62	1232075.61
Pseudo R2	0.02	0.02	0.03	0.02

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S14. VIFS for Study 2 Models Without Control Variables

	Twitter		Facebook	
	Liberal	Conservative	Liberal	Conservative
	VIF	VIF	VIF	VIF
Democrat	1.01	1.04	1.02	1.09
Republican	1.02	1.03	1.08	1.06
NegativeAffect	1.31	1.22	1.5	1.6
PositiveAffect	1.2	1.2	1.36	1.41
MoralEmotional	1.47	1.4	1.81	1.95

Table S15. Study 2 With Cluster Robust Standard Errors

	Facebook		Twitter	
	Liberal	Conservative	Liberal	Conservative
(Intercept)	8.79 ***	7.83 ***	9.70 ***	6.25 ***
	[8.65, 8.94]	[7.70, 7.96]	[9.63, 9.78]	[6.20, 6.30]
Democrat	1.02 ***	1.65 ***	0.75 ***	2.80 ***
	[1.01, 1.03]	[1.62, 1.68]	[0.74, 0.76]	[2.75, 2.86]
Republican	1.58 ***	1.20 ***	2.13 ***	0.85 ***
	[1.56, 1.59]	[1.18, 1.21]	[2.10, 2.15]	[0.84, 0.86]
NegativeAffect	1.14 ***	1.12 ***	1.33 ***	1.45 ***
	[1.13, 1.14]	[1.11, 1.12]	[1.33, 1.34]	[1.43, 1.46]
PositiveAffect	0.96 ***	0.98 ***	0.95 ***	1.04 ***
	[0.95, 0.96]	[0.97, 0.98]	[0.95, 0.95]	[1.04, 1.05]
MoralEmotional	1.06 ***	1.05 ***	1.10 ***	1.06 ***
	[1.05, 1.06]	[1.04, 1.05]	[1.10, 1.11]	[1.05, 1.07]
has_mediaTRUE	1.09 ***	0.92 ***	1.08 ***	1.22 ***
	[1.07, 1.11]	[0.91, 0.94]	[1.07, 1.09]	[1.21, 1.23]
has_URLTRUE	1.24 ***	1.00	0.95 ***	0.83 ***
	[1.21, 1.26]	[0.98, 1.02]	[0.94, 0.96]	[0.82, 0.83]
followers_count			1.00 ***	1.00 ***
			[1.00, 1.00]	[1.00, 1.00]
is_retweetTRUE			5.10 ***	5.36 ***
			[5.03, 5.18]	[5.28, 5.44]
`Likes at Posting`	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]		
N	354814	410313	747675	611292
AIC	1253213.12	1474731.77	2861220.18	2343833.57
BIC	1253320.92	1474841.02	2861346.95	2343958.13
Pseudo R2	0.33	0.13	0.30	0.33

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table S16. Study 2 Relative Importance Analysis

	Twitter		Facebook	
	Conservative	Liberal	Conservative	Liberal
	lmg	lmg	lmg	lmg
Democrat	**0.046435128**	0.003990291	**0.037326404**	0.000695989
Republican	0.004320336	**0.035396725**	0.010754976	**0.045910789**
NegativeAffect	0.020092764	0.021551454	0.018937441	0.022214026
PositiveAffect	0.000413394	0.002722643	0.001024292	0.002541912
MoralEmotional	0.003008898	0.004656238	0.005519233	0.006538567
has_media	0.002206309	0.00426283	0.000992487	0.003629467
has_URL	0.01654347	0.006841593	0.002193025	0.002415596
followers_count	0.131899268	0.136219852		
is_retweet	0.10321281	0.08377927		
Likes at Posting			0.048458365	0.244020805

Table S17. Study 2 Conservative Congress Facebook Reactions

	Shares	Comments	Likes	Loves	Wow	Haha	Sad	Angry	Retweet	Favorite
(Intercept)	7.83 ***	31.95 ***	64.91 ***	4.56 ***	1.56 ***	2.36 ***	2.03 ***	3.08 ***	6.25 ***	12.08 ***
	[7.71, 7.95]	[31.38, 32.52]	[63.96, 65.87]	[4.50, 4.63]	[1.55, 1.57]	[2.33, 2.38]	[2.01, 2.04]	[3.04, 3.13]	[6.20, 6.30]	[11.99, 12.18]
Democrat	1.65 ***	1.58 ***	1.32 ***	1.13 ***	1.31 ***	1.43 ***	1.26 ***	1.68 ***	2.80 ***	1.93 ***
	[1.64, 1.67]	[1.56, 1.60]	[1.31, 1.33]	[1.12, 1.14]	[1.30, 1.31]	[1.42, 1.43]	[1.25, 1.27]	[1.67, 1.69]	[2.77, 2.84]	[1.91, 1.95]
Republican	1.20 ***	1.41 ***	1.26 ***	1.32 ***	1.05 ***	1.28 ***	1.04 ***	1.25 ***	0.85 ***	1.18 ***
	[1.19, 1.20]	[1.40, 1.42]	[1.25, 1.27]	[1.31, 1.33]	[1.04, 1.05]	[1.27, 1.29]	[1.04, 1.04]	[1.24, 1.26]	[0.84, 0.85]	[1.17, 1.19]
NegativeAffect	1.12 ***	1.10 ***	1.05 ***	1.00	1.04 ***	1.00	1.10 ***	1.06 ***	1.45 ***	1.37 ***
	[1.11, 1.12]	[1.09, 1.10]	[1.05, 1.05]	[0.99, 1.00]	[1.03, 1.04]	[1.00, 1.00]	[1.10, 1.10]	[1.06, 1.06]	[1.44, 1.45]	[1.36, 1.38]
PositiveAffect	0.98 ***	0.99 ***	1.01 ***	1.02 ***	0.98 ***	0.98 ***	0.97 ***	0.96 ***	1.04 ***	1.16 ***
	[0.97, 0.98]	[0.99, 0.99]	[1.01, 1.02]	[1.02, 1.02]	[0.98, 0.98]	[0.98, 0.98]	[0.97, 0.97]	[0.96, 0.96]	[1.04, 1.05]	[1.16, 1.16]
MoralEmotional	1.05 ***	1.03 ***	1.02 ***	1.05 ***	1.00	1.00	1.03 ***	1.02 ***	1.06 ***	1.03 ***
	[1.04, 1.05]	[1.03, 1.04]	[1.02, 1.03]	[1.04, 1.05]	[1.00, 1.00]	[0.99, 1.00]	[1.03, 1.04]	[1.02, 1.03]	[1.05, 1.07]	[1.03, 1.04]
has_URLTRUE	1.00	0.64 ***	0.69 ***	0.69 ***	0.99 **	0.82 ***	0.77 ***	0.79 ***	0.83 ***	0.81 ***
	[0.98, 1.01]	[0.63, 0.65]	[0.68, 0.70]	[0.68, 0.71]	[0.98, 1.00]	[0.81, 0.83]	[0.77, 0.78]	[0.78, 0.81]	[0.82, 0.83]	[0.80, 0.82]
has_mediaTRUE	0.92 ***	0.65 ***	0.95 ***	0.96 ***	0.92 ***	0.84 ***	0.73 ***	0.74 ***	1.22 ***	1.52 ***
	[0.91, 0.94]	[0.64, 0.67]	[0.94, 0.97]	[0.95, 0.98]	[0.91, 0.92]	[0.83, 0.84]	[0.73, 0.74]	[0.73, 0.75]	[1.20, 1.23]	[1.50, 1.53]
`Likes at Posting`	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]		
is_retweetTRUE									5.36 ***	0.07 ***
									[5.30, 5.42]	[0.07, 0.07]
followers_count									1.00 ***	1.00 ***
									[1.00, 1.00]	[1.00, 1.00]
N	410313	410313	410313	410313	410313	410313	410313	410313	611292	611292
AIC	1474731.77	1608010.64	1457229.58	1422293.10	900551.43	1183686.06	1069002.13	1375557.79	2343833.57	2265816.97
BIC	1474841.02	1608119.89	1457338.82	1422402.35	900660.68	1183795.31	1069111.38	1375667.04	2343958.13	2265941.53
Pseudo R2	0.13	0.11	0.11	0.09	0.13	0.13	0.12	0.13	0.33	0.43

Table S18. Study 2 Liberal Congress Facebook Reactions

	Shares	Comments	Likes	Loves	Wow	Haha	Sad	Angry	Retweet	Favorite
(Intercept)	8.79 ***	26.67 ***	81.16 ***	7.74 ***	1.73 ***	2.76 ***	2.73 ***	2.87 ***	9.70 ***	25.87 ***
	[8.65, 8.94]	[26.14, 27.21]	[79.87, 82.47]	[7.61, 7.88]	[1.71, 1.74]	[2.73, 2.80]	[2.69, 2.77]	[2.82, 2.91]	[9.63, 9.78]	[25.69, 26.05]
Democrat	1.02 ***	1.16 ***	1.12 ***	1.23 ***	0.97 ***	1.15 ***	0.89 ***	0.93 ***	0.75 ***	1.04 ***
	[1.01, 1.03]	[1.15, 1.17]	[1.11, 1.13]	[1.22, 1.24]	[0.97, 0.98]	[1.14, 1.16]	[0.89, 0.90]	[0.92, 0.93]	[0.75, 0.75]	[1.04, 1.05]
Republican	1.58 ***	1.75 ***	1.28 ***	1.01	1.32 ***	1.45 ***	1.42 ***	2.24 ***	2.13 ***	1.69 ***
	[1.57, 1.59]	[1.74, 1.76]	[1.27, 1.29]	[1.00, 1.01]	[1.32, 1.33]	[1.44, 1.45]	[1.41, 1.42]	[2.22, 2.25]	[2.11, 2.15]	[1.68, 1.70]
NegativeAffect	1.14 ***	1.17 ***	1.08 ***	1.01 ***	1.07 ***	1.04 ***	1.26 ***	1.17 ***	1.33 ***	1.30 ***
	[1.13, 1.14]	[1.17, 1.18]	[1.07, 1.08]	[1.01, 1.01]	[1.07, 1.07]	[1.04, 1.04]	[1.25, 1.26]	[1.17, 1.17]	[1.33, 1.34]	[1.29, 1.30]
PositiveAffect	0.96 ***	0.94 ***	1.02 ***	1.03 ***	0.96 ***	0.95 ***	0.92 ***	0.90 ***	0.95 ***	1.05 ***
	[0.95, 0.96]	[0.94, 0.95]	[1.02, 1.02]	[1.03, 1.03]	[0.96, 0.97]	[0.95, 0.95]	[0.92, 0.92]	[0.90, 0.90]	[0.95, 0.95]	[1.05, 1.05]
MoralEmotional	1.06 ***	1.09 ***	1.04 ***	1.05 ***	1.00	1.02 ***	1.05 ***	1.06 ***	1.10 ***	1.05 ***
	[1.05, 1.06]	[1.08, 1.10]	[1.03, 1.04]	[1.04, 1.05]	[1.00, 1.00]	[1.02, 1.03]	[1.04, 1.05]	[1.06, 1.07]	[1.10, 1.11]	[1.05, 1.06]
has_URLTRUE	1.24 ***	0.70 ***	0.66 ***	0.61 ***	1.24 ***	0.81 ***	1.04 ***	1.28 ***	0.95 ***	0.87 ***
	[1.21, 1.26]	[0.68, 0.71]	[0.65, 0.67]	[0.60, 0.62]	[1.23, 1.26]	[0.80, 0.82]	[1.03, 1.06]	[1.26, 1.30]	[0.94, 0.96]	[0.86, 0.87]
has_mediaTRUE	1.09 ***	0.66 ***	0.85 ***	0.94 ***	0.99	0.82 ***	0.74 ***	0.87 ***	1.08 ***	1.15 ***
	[1.08, 1.11]	[0.65, 0.68]	[0.84, 0.87]	[0.92, 0.96]	[0.98, 1.00]	[0.81, 0.84]	[0.73, 0.75]	[0.85, 0.88]	[1.07, 1.09]	[1.14, 1.16]
`Likes at Posting`	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***	1.00 ***		
	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]		
is_retweetTRUE									5.10 ***	0.04 ***
									[5.04, 5.16]	[0.04, 0.04]
followers_count									1.00 ***	1.00 ***
									[1.00, 1.00]	[1.00, 1.00]
N	354814	354814	354814	354814	354814	354814	354814	354814	747675	747675
AIC	1253213.12	1391799.69	1228236.29	1266431.32	878864.12	1059362.91	1147845.85	1222456.91	2861220.18	2760433.97
BIC	1253320.92	1391907.49	1228344.08	1266539.11	878971.92	1059470.70	1147953.64	1222564.70	2861346.95	2760560.74
Pseudo R2	0.33	0.22	0.25	0.17	0.42	0.23	0.36	0.39	0.30	0.51

Table S19. Liberal Congress Facebook Reactions

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Shares	134.14	1590.63							
2. Likes	333.68	1722.16	.60** [.59, .60]						
3. Comments	124.52	506.20	.45** [.44, .45]	.57** [.57, .58]					
4. Love	41.35	325.31	.43** [.43, .43]	.80** [.80, .81]	.48** [.48, .49]				
5. Wow	7.59	80.29	.52** [.52, .52]	.41** [.41, .42]	.38** [.38, .38]	.15** [.15, .16]			
6. Haha	8.46	88.31	.24** [.24, .24]	.30** [.30, .31]	.40** [.40, .40]	.21** [.21, .21]	.24** [.24, .25]		
7. Sad	28.75	272.60	.38** [.38, .38]	.35** [.35, .36]	.34** [.34, .34]	.12** [.12, .13]	.40** [.40, .40]	.09** [.09, .10]	
8. Angry	67.91	672.91	.41** [.41, .42]	.31** [.30, .31]	.43** [.43, .43]	.04** [.04, .05]	.63** [.63, .63]	.20** [.20, .20]	.45** [.45, .45]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

Table S20. Conservative Congress Facebook Reactions

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Shares	67.82	5292.50							
2. Likes	241.78	1659.19	.67** [.66, .67]						
3. Comments	97.68	512.71	.41** [.41, .42]	.54** [.54, .54]					
4. Love	17.66	153.78	.12** [.11, .12]	.44** [.44, .45]	.35** [.35, .35]				
5. Wow	2.09	35.43	.73** [.73, .73]	.55** [.55, .55]	.46** [.46, .46]	.17** [.17, .17]			
6. Haha	4.69	45.48	.06** [.06, .06]	.18** [.18, .18]	.34** [.34, .34]	.23** [.23, .23]	.19** [.19, .19]		
7. Sad	5.07	109.59	.07** [.06, .07]	.12** [.12, .12]	.21** [.21, .21]	.09** [.09, .10]	.29** [.29, .29]	.04** [.04, .04]	
8. Angry	15.60	212.02	.47** [.47, .47]	.38** [.38, .39]	.64** [.64, .64]	.13** [.13, .13]	.67** [.67, .67]	.22** [.22, .23]	.22** [.21, .22]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

Table S21. Descriptive Statistics – Congress

	Conservative Twitter		Liberal Twitter		Conservative Congress		Liberal Congress	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Democrat	0.1	0.38	0.21	0.66	0.13	0.54	0.21	0.58
2. Republican	0.23	0.63	0.15	0.44	0.24	0.67	0.27	0.72
3. PositiveAffect	1.16	1.28	1.19	1.28	2.48	2.78	2.31	2.46
4. NegativeAffect	0.35	0.72	0.54	0.88	0.79	1.66	0.98	1.61
5. MoralEmotional	0.29	0.6	0.42	0.72	0.65	1.22	0.78	1.25

Table S22. Meta-Analyzed Effect Sizes (Facebook Reactions)

name	estimate	conf.low	conf.high	model
Shares	1.53	1.41	1.66	Outgroup
Comments	1.77	1.51	2.08	Outgroup
Likes	1.30	1.20	1.42	Outgroup
Love	1.15	0.99	1.33	Outgroup
Haha	1.96	1.40	2.75	Outgroup
Wow	1.36	1.28	1.44	Outgroup
Sad	1.29	1.19	1.40	Outgroup
Angry	2.19	1.68	2.84	Outgroup
Retweets	1.83	1.35	2.49	Outgroup
Favorites	1.49	1.22	1.81	Outgroup
Shares	1.14	0.99	1.30	Ingroup
Comments	1.47	1.24	1.74	Ingroup
Likes	1.34	1.13	1.60	Ingroup
Love	1.57	1.24	2.00	Ingroup
Haha	1.42	1.20	1.68	Ingroup
Wow	0.98	0.92	1.04	Ingroup
Sad	0.94	0.86	1.03	Ingroup
Angry	1.18	0.99	1.42	Ingroup
Retweets	0.96	0.81	1.14	Ingroup
Favorites	1.20	1.09	1.32	Ingroup

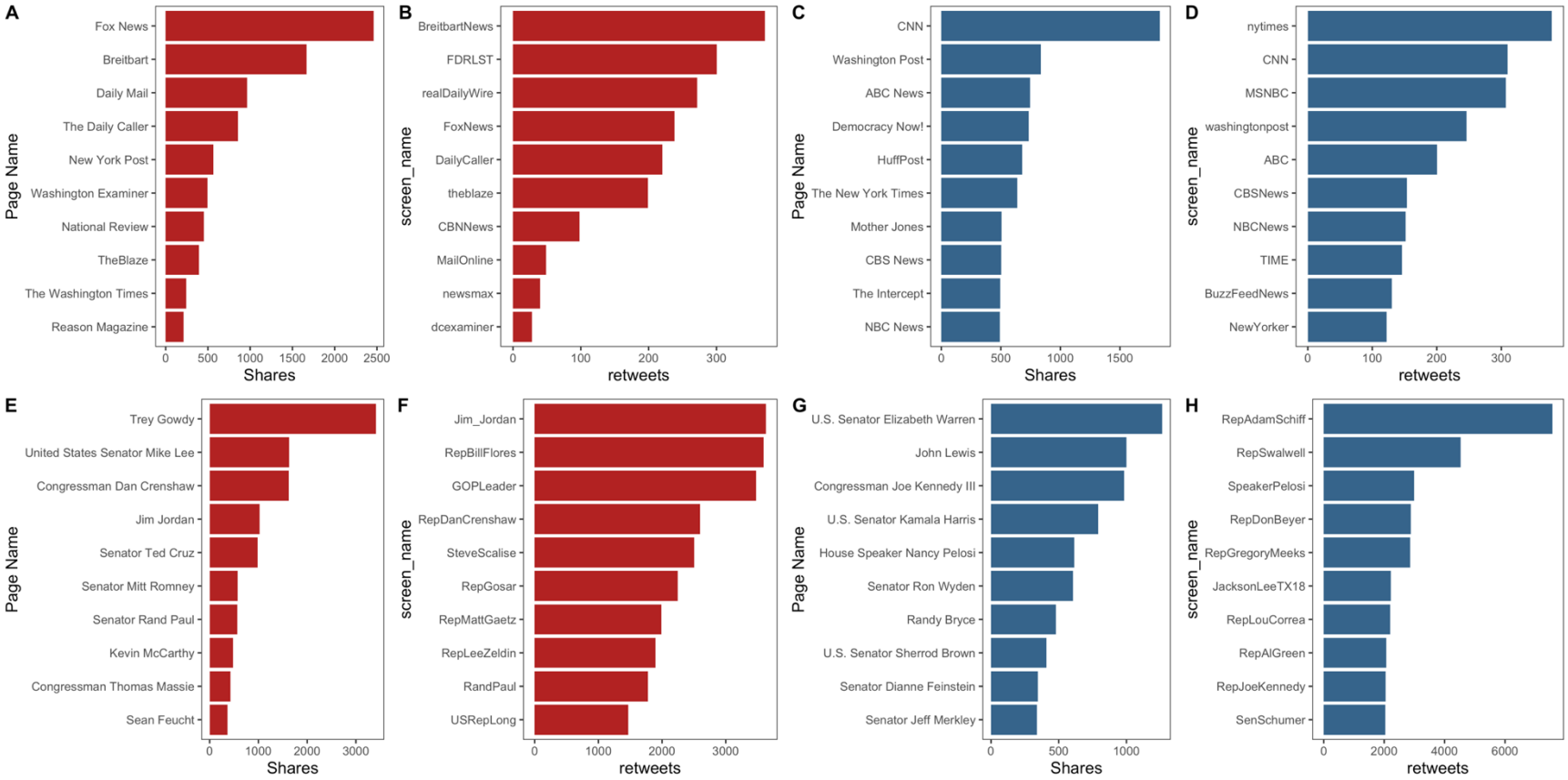


Figure S1. Pages Associated with the Most Engagement on Facebook and Twitter. Panels represent (A) conservative media Facebook, (B) conservative media Twitter, (C) liberal media Facebook, (D), liberal media Twitter, (E) conservative congress Facebook, (F) conservative congress Twitter, (G) liberal congress Facebook, and (H) liberal congress Twitter.

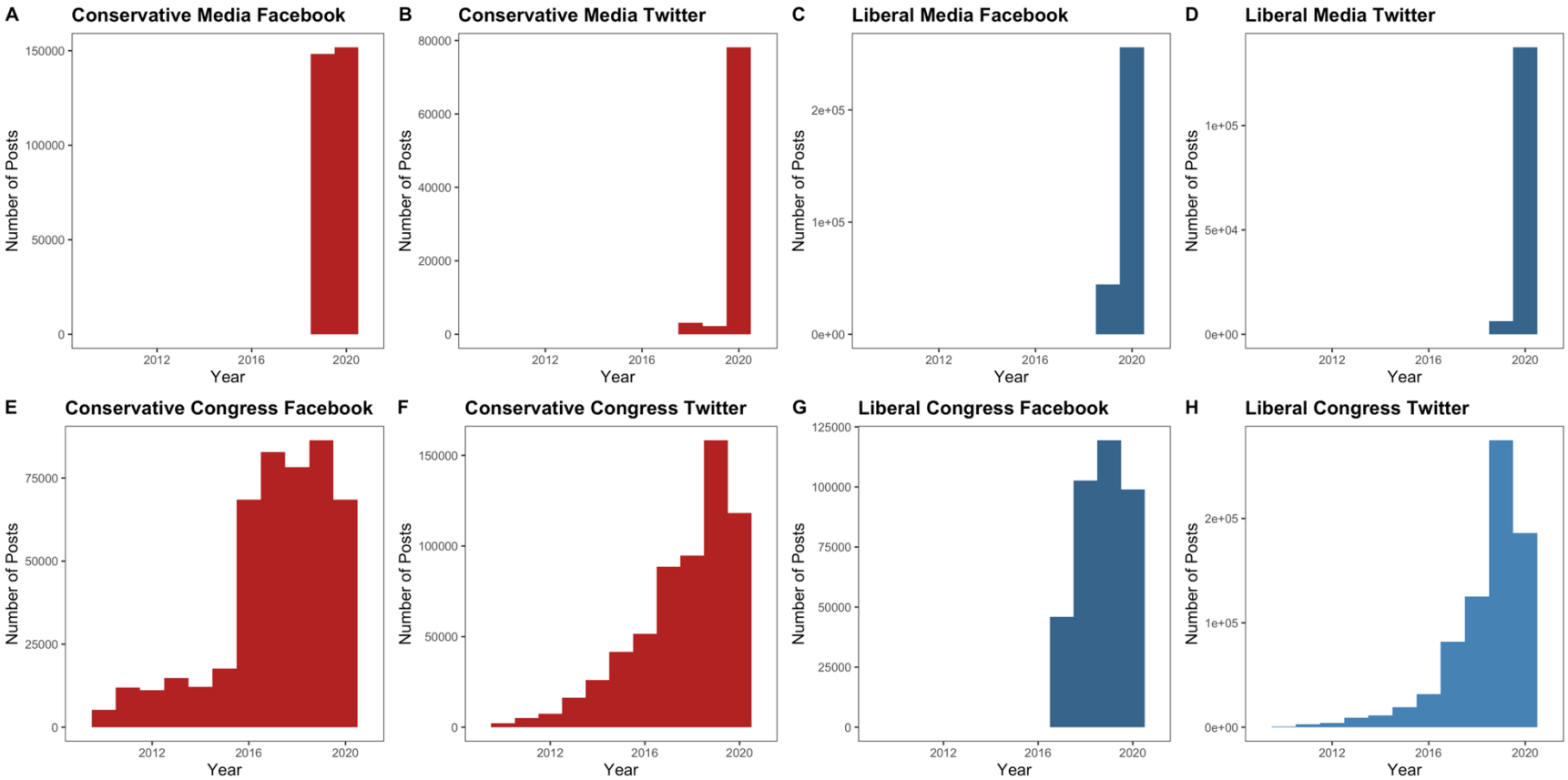


Figure S2. Histograms of the time the tweets and Facebook posts were created. The media tweets were retrieved on May 4 and July 11, 2020; the congress tweets were retrieved on July 2, 2020; the media Facebook posts were retrieved on August 14, 2020, and the congress media posts were retrieved on August 18, 2020.

Figure S3: AllSides Media Bias Chart. The above 2019 AllSides Media Bias Chart (retrieved from AllSides.com) was used retrieve Twitter handles and Facebook accounts. The left and right media accounts (but not the centrist ones) were used.

9. Supplementary Materials for "Accuracy and Social Motivations Shape Judgements of (Mis)Information"

S1: Extended Results

Study 1

Analysis of Type of Headlines Impacted. To explore what kind of headlines the incentives impacted specifically, we conducted a 2 (incentives vs. control condition) X 2 (true headlines vs. false headlines) X 2 (politically-congruent versus politically-incongruent) mixed-design ANOVA with the percentage of articles rated as accurate as the dependent variable. There was a main effect of condition, $F(1, 460) = 12.71$, $p < 0.001$, $\eta^2_G = 0.01$, political congruence, $F(1, 460) = 263.50$, $p < 0.001$, $\eta^2_G = 0.11$, and veracity of the headlines, $F(1, 460) = 5.58$, $p < 0.001$, $\eta^2_G = 0.27$. There was also an interaction effect between the incentives and political congruence of the headlines, $F(1, 460) = 8.00$, $p = 0.01$, $\eta^2_G = 0.004$, and between the incentives and the veracity of the headlines $F(1, 460) = 7.77$, $p = 0.003$, $\eta^2_G = 0.004$.

Following up on these interaction effects with Tukey HSD post-hoc tests, we found that the incentives primarily increased belief in politically-incongruent true news ($M = 51.53$, 95% CI = [47.36, 55.70]) when compared to the control condition ($M = 38.25$, 95% CI = [34.41, 42.08]), $p < 0.001$, $d = 0.43$. When controlling for multiple comparisons with Tukey HSD post-hoc tests, incentives had no effect on politically-incongruent false news ($p = 0.444$), politically-congruent false news ($p = 0.999$), or politically-congruent true news ($p = 0.472$). In other words, the effect of the incentives was driven by an increase in belief in news from the opposing party.

Study 2

Analysis of Type of Headlines Impacted. To test what types of headlines were affected by the incentives, we ran a 2 (accuracy incentive vs. no incentive) X 2 (social incentive vs. no incentive) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA with the percent of articles rated as accurate as the dependent variable. There was a significant main effect of the accuracy incentives, $F(1, 994) = 23.44$, $p < 0.001$, $\eta^2_G = 0.01$, veracity, $F(1, 994) = 550.43$, $p < 0.001$, $\eta^2_G = 0.20$, and political congruence, $F(1, 994) = 8.99$, $p = 0.003$, $\eta^2_G = 0.002$. Furthermore, there was a significant interaction between accuracy incentives and political congruence, $F(1,994) = 8.99$, $p = 0.003$, $\eta^2_G = 0.00$,

between accuracy incentives and veracity, $F(1, 994) = 29.06$, $p < 0.001$, $\eta^2_G = 0.01$, and between social incentives and veracity, $F(1, 994) = 7.613$, $p = 0.006$, $\eta^2_G = 0.00$. All other *ps* > 0.085.

Tukey HSD post-hoc tests found that there was a significant difference in the amount of *incongruent true* articles rated as accurate between the accuracy incentives condition (*M* = 55.61%, 95% CI = [51.68, 59.54]) and the control condition (M = 37.65%, 95% CI = [33.83, 41.46]), $p < 0.001$, $d = 0.58$. However, the mixed incentives condition (M = 46.07%, 95% CI = [42.04, 51.10]) did not differ from the control condition (M =), $p = 0.092$, once again supporting the idea that social incentives distract from accuracy incentives. The incentives once again did not impact congruent true news, incongruent false news, or congruent false news (ps > 0.148).

Analysis of Sharing Behavior. To test how incentives influenced sharing intentions, we ran another 2X2X2X2 mixed ANOVA on sharing intentions. Here, there was no effect of accuracy incentives, $F(1, 994) = 0.05$, $p = 0.830$, $\eta^2 = 0.00$, but there was a significant main effect of social incentives, $F(1, 994) = 10.07$, $p = 0.002$, $\eta^2 = 0.00$, political congruence, $F(1, 994) = 368.96$, $p < 0.001$, $\eta^2 = 0.05$, and veracity, $F(1, 994) = 173.71$, $p < 0.001$, $\eta^2 = 0.01$. Furthermore, there was a significant interaction between social incentives and political congruence, $F(1, 994) = 20.41$, $p < 0.001$, $\eta^2 = 0.00$.

Following up on the interaction between the social incentives and political congruence, we found that those in the social incentives condition shared more politically congruent news (either true or false) (*M* = 1.98, 95% CI = [1.90, 2.05]) as compared to the control condition (*M* = 1.80, 95% CI = [1.74, 1.87]), $p = 0.015$, $d = 0.21$. Additionally, those in the mixed condition (*M* = 2.02, 95% CI = [1.94, 2.10]) shared more politically congruent news (true or false) as compared to the control condition, $p < 0.001$, $d = 0.26$. Thus, thinking about whether an article will be liked by one's party, whether or not one is incentivized to be accurate, appears to indiscriminately increase sharing of both true and false news that appeal to one's political party.

Study 3

Analysis of Type of Headlines Impacted. We ran a 2 (accuracy incentive vs. no incentive) X 2 (social incentive vs. no incentive) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA. Here, we saw a significant main effect of political congruence on perceived accuracy, $F(1, 917) = 457.79$, $p < 0.001$, $\eta^2 = 0.09$. There was also a significant main effect of veracity on perceived accuracy, $F(1,$

917) = 945.35, $p < 0.01$, $\eta^2 = 0.20$. As in Study 2, there was a significant interaction between the accuracy incentives condition and political congruence, $F(1, 917) = 18.22$, $p < 0.001$, $\eta^2 = 0.00$, a significant interaction between veracity and political congruence, $F(1, 917) = 5.65$, $p = 0.018$, $\eta^2 = 0.00$, a significant interaction between the accuracy incentives and source cues, $F(1, 917) = 4.71$, $p = 0.030$, $\eta^2 = 0.00$, and a significant interaction between the accuracy incentives and the veracity of the headline, $F(1, 917) = 4.71$, $p = 0.036$, $\eta^2 = 0.00$.

We then followed up on these interactions with Tukey post-hoc tests. Replicating the results of studies 1 and 2, there was a large difference in the percentage of *incongruent true* headlines rated as accurate in the accuracy incentive (with sources) condition ($M = 51.20$, 95% CI = [47.28, 55.12]) versus the control (with sources) condition ($M = 39.47$, 95% CI = [35.69, 43.34]), $p < 0.001$, $d = 0.39$. However, without sources present beside the headlines, there was no difference in the percentage of incongruent true headlines rated as accurate when comparing the accuracy incentive and control condition ($p = 0.605$). No other post-hoc tests were significant ($ps > 0.864$). These results replicate the finding that the effects are driven by an increase in belief in politically-incongruent headlines, but also show these effects depend upon source cues being present beside the headlines.

Like in Experiment 2, there was once again no significant impact of accuracy incentives on sharing discernment ($p = 0.906$). However, there was an effect of the source cues on sharing discernment such that source cues improved sharing discernment, $F(1, 917) = 4.92$, $p = 0.027$, $\eta^2 = 0.01$. There was also no interaction between accuracy incentives and source cues on sharing discernment ($p = 0.124$).

Integrative Data Analysis

Reaction Time Data. To further examine whether the accuracy incentives increased the amount of effort people put into discerning the accuracy of headlines, we also examined reaction time data (in seconds) in the pooled dataset. We ran a 2X2X2 ANOVA (Condition X Political Congruence X Veracity) and found a main effect of condition such that people spent more time on each item in the accuracy ($M = 18.27$, 95% CI = [17.58, 18.95]) as opposed to control condition ($M = 15.88$, 95% CI = [15.45, 16.30]), $F(1, 1458) = 20.50$, $p < 0.001$, $\eta^2 = 0.01$. There was also a main effect of veracity such that people spent more time on true news ($M = 18.77$,

95% CI = [18.15, 19.40]) than false news (M = 16.56, 95% CI = [15.82, 17.30]), $F(1, 1458)$ = 17.28, p = 0.002, $\eta^2 = 0.00$.

Partisan Differences. Conservatives also showed more partisan bias than liberals: partisan bias was 2.55 points for unincentivized conservatives and 1.16 points for unincentivized liberals – a 1.40 point difference, 95% CI = [1.15, 1.64], $t(906.25) = 11.23$, $p < .001$, $d = 0.70$. Yet, this difference became 0.65 points when conservatives were incentivized to be accurate, 95% CI = [0.40, 0.90], $t(834.30) = 5.02$, $p = 0.001$, $d = 0.32$. In other words, while conservatives initially expressed more partisan bias, incentives for accuracy closed this gap in partisan bias by 53.57%.

Finally, liberals believed 1.78 (out of 4) true news headlines from the opposing party, whereas conservatives believed 1.14 (out of 4) true news headlines from the opposing party – a 0.65-point difference, 95% CI [0.50, 0.79], t(1026.59) = 9.00, p < .001, $d = 0.55$. But, when conservatives were incentivized to be accurate, they correctly identified 1.83 (out of 4) true news headlines from the opposing party, eliminating this gap, difference = 0.05, 95% CI [-0.10, 0.20], t(918.14) = 0.63, $p = 0.528$, $d = 0.04$

Addressing Multiple Interpretations. One alternate interpretation of our results is that participants were simply guessing what fact-checkers would say is true in the accuracy incentives condition rather than expressing their genuine beliefs about accuracy. However, in studies 2 and 3, we asked participants in the accuracy incentive condition whether they answered in a way that did not reflect their true beliefs just to receive the payment, and told participants that their answer to this question would not affect their final payment (See *S7* for question wording). Only 3% of participants said "yes" to this question, indicating that people reported responding in a way that reflected their true beliefs.

Another interpretation is that accuracy incentives inhibit motivated responding (also known as "expressive responding" or "partisan cheerleading")(Peterson & Iyengar, 2021; Schaffner & Luks, 2018), or partisans' tendency to purposely give incorrect answers just to express support for their own party. To address this interpretation, in study 2, we asked participants who were not in the accuracy incentives condition whether they ever said an article was true (or false) not because they actually believed it was true (or false), but because they liked (or disliked it). Only 5% of participants admitted to engaging in this kind of motivated

responding, indicating that most participants reported answering in line with their genuine beliefs.

While our designs cannot fully tease apart these explanations, the self-report questions (if we assume people are being truthful in their answers) indicate these alternate interpretations are unlikely. Instead, it appears that the accuracy incentives motivate people to put more effort into providing correct responses rather than giving a quick response based on the partisan-lean of the source and content. Further supporting this interpretation, supplementary analysis indicated that participants responded more slowly in the accuracy incentives condition, presumably because they were putting more effort into their responses.

We also directly asked participants whether they believed their responses were influenced by the treatment conditions at the end of the experiment. In total, 20% of participants said they believed their judgements were influenced by the accuracy incentives, and 24% of participants said they believed their judgements were influenced by the task of identifying politically-congruent articles. Thus, while some participants were aware that the experimental conditions impacted how they responded, the majority were not aware of this. In most cases, it seems as if participants were responding in a way that they thought reflected their true beliefs and were blind to the impact of incentives. Furthermore, only 7% of participants said they would knowingly share fake news on social media.

S2: Item-by-Item Analysis

Type	Headline	Source	Mean (Accuracy)	Mean (Control)	Mean Difference	CI Low	CI High	t	df	p	Cohen's D
Democrat True	Trump allies are handing out cash to black voters	Politico	0.40	0.30	-0.10	0.05	0.15	3.97	1412.44	0.000	0.21
Democrat True	Trump targets Reagan foundation after it asks campain, RNC to stop using former president's likeness	CBS News	0.69	0.55	-0.14	0.09	0.19	5.48	1424.16	0.000	0.29
Democrat True	Facebook removes Trump ads with symbols once used by Nazis	AP News	0.67	0.58	-0.09	0.04	0.14	3.51	1425.70	0.000	0.19
Democrat True	Melania Trump was praised for acknowledging racism. But she has also spread false 'birther' claims about Trump.	Washington Post	0.70	0.60	-0.10	0.05	0.15	3.87	1424.70	0.000	0.21
Democrat False	White House Chef Quits because Trump Has Only Eaten Fast Food For 6 Months	HalfWay Post	0.21	0.18	-0.03	-0.01	0.07	1.25	1415.68	0.212	0.07
Democrat False	Trump's Top Scientist Pick: "Scientists Are Just Dumb Regular People That Think Dinosaurs Existed and the Earth is Getting Warmer"	USPoln	0.27	0.23	-0.04	-0.01	0.08	1.65	1416.00	0.099	0.09
Democrat False	Hispanic Women Claims, "Donald Trump Paid Me For Sex in Cancun, This Is Our Love Child"	Now8News	0.18	0.11	-0.06	0.03	0.10	3.36	1363.43	0.001	0.18
Democrat False	Donald Trump Signs Executive Order Allowing the Hunting of Bald Eagles		0.18	0.18	0.01	-0.05	0.03	-0.33	1425.48	0.745	-0.02
Republican True	Plant a Million Trees: Republicans Offer Fossil-Friendly Climate Fix	Reuters	0.56	0.45	-0.11	0.05	0.16	3.99	1424.56	0.000	0.21
Republican True	UPSP Flashback: Obama administration removed thousands of mailboxes	Fox News	0.49	0.34	-0.15	0.09	0.20	5.64	1415.93	0.000	0.30
Republican True	Trump gets support of NYC police union, warns 'no one will be safe in Biden's America'	NBC News	0.84	0.78	-0.06	0.02	0.10	2.73	1416.80	0.006	0.15
Republican True	Chinese dissedent brought to US by Obama administration praises Trump at RNC	CNN	0.52	0.44	-0.07	0.02	0.13	2.79	1423.73	0.005	0.15
Republican False	UPDATE: Malia Obama Among 10 Arrested In Racist Antifa Attach -- US NEWS	PoliceUS.Info	0.09	0.10	0.01	-0.04	0.02	-0.35	1425.91	0.730	-0.02
Republican False	Hillary Clinton Accepted $30,000 Donation From NXIVM Child Sex Cult	YourNewsWire	0.30	0.27	-0.03	-0.02	0.08	1.22	1420.17	0.222	0.06

Republican	The 'Obama Foundation' Just Broke Its First							-			
False	Federal Law	WeaponStricks	0.28	0.29	0.01	-0.05	0.04	0.30	1425.01	0.767	-0.02
Republican	Donald Trump Sent His Own Plane To Transport							-			
False	200 Stranded Marines	Uconservative	0.42	0.44	0.02	-0.07	0.03	0.78	1424.86	0.437	-0.04

Note: Data is from the integrative data analysis.

S3: Full Regression Models for Integrative Data Analysis

	Truth Discernment	Partisan Bias	Incongruent True News
(Intercept)	2.298 ***	1.615 ***	1.663 ***
	[2.092, 2.503]	[1.437, 1.793]	[1.553, 1.773]
conditionRecode	0.700 ***	-0.538 ***	0.581 ***
	[0.468, 0.931]	[-0.738, -0.337]	[0.457, 0.705]
PoliticalOrientation	-0.633 ***	0.631 ***	-0.230 ***
	[-0.753, -0.513]	[0.527, 0.736]	[-0.294, -0.166]
CRSum	0.371 ***	-0.157 **	0.116 ***
	[0.249, 0.493]	[-0.263, -0.052]	[0.051, 0.181]
PKSum	0.444 ***	-0.034	0.074 *
	[0.319, 0.568]	[-0.142, 0.074]	[0.007, 0.141]
outgrouphate	0.159 *	0.521 ***	-0.119 ***
	[0.035, 0.282]	[0.414, 0.628]	[-0.185, -0.053]
Education	-0.025	-0.087	0.081 *
	[-0.148, 0.099]	[-0.194, 0.021]	[0.015, 0.147]
Age	-0.161 **	0.202 ***	-0.106 **
	[-0.280, -0.041]	[0.099, 0.306]	[-0.170, -0.043]
Income	0.045	-0.016	0.028
	[-0.078, 0.168]	[-0.123, 0.091]	[-0.038, 0.094]
GenderRecode	-0.170	0.324 **	-0.249 ***
	[-0.409, 0.068]	[0.118, 0.531]	[-0.376, -0.122]
N	1385	1385	1385
AIC	6117.177	5722.786	4380.750
BIC	6174.745	5780.354	4438.318
Pseudo R2	0.192	0.187	0.145

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

S4: Full Relative Importance Analysis for Integrative Data Analysis

Truth Discernment

term	lmg	conf.low	conf.high
conditionRecode	0.020532317	0.008935026	0.036625856
PoliticalOrientation	0.071745706	0.049429337	0.100798511
CRSum	0.035210002	0.02058485	0.054147144
PKSum	0.040337337	0.024215731	0.061294932
outgrouphate	0.013212485	0.005012211	0.025773805
Education	0.000615348	0.000255952	0.004317704
Age	0.005573851	0.001001938	0.014396754
Income	0.000445728	0.000133143	0.004277437
GenderRecode	0.002093208	0.000285042	0.008885336

Partisan Bias

term	lmg	conf.low	conf.high
conditionRecode	0.015910588	0.00633209	0.030680641
PoliticalOrientation	0.079438354	0.053838595	0.106793351
CRSum	0.007990016	0.00192017	0.018036401
PKSum	0.000988131	0.000618156	0.005037605
outgrouphate	0.048763912	0.031689179	0.071379885
Education	0.004500754	0.000611908	0.014214294
Age	0.0174529	0.007333396	0.032545215
Income	0.000862162	0.000238516	0.005147937
GenderRecode	0.008326444	0.002133619	0.019153783

Incongruent True News

term	lmg	conf.low	conf.high
conditionRecode	0.053477661	0.033399849	0.078825372
PoliticalOrientation	0.031479192	0.015709584	0.051636357
CRSum	0.012995727	0.004533847	0.02632355
PKSum	0.003228492	0.000391516	0.010660455

outgrouphate	0.006170344	0.000933994	0.015800282
Education	0.007324363	0.001636777	0.018196119
Age	0.010170418	0.002677802	0.021930849
Income	0.002166289	0.000335553	0.008761043
GenderRecode	0.012443196	0.004029015	0.026493235

S5: Example Stimuli

Example Republican-Leaning Real News:

Example Democrat-Leaning Real News:

Example Democrat-Leaning Fake News

Example Republican-Leaning Fake News:

Example headline without source:

The full set of stimuli are available on our OSF: https://osf.io/75sqf.

S6: Manipulation Text

Accuracy Incentives Manipulation Text

You will be presented with a series of real and fake news headlines. There are 16 headlines in total.

We are interested in your opinion about the following:

1) How accurate is the headline?
2) How likely would you be to share the headline on social media?

You will be given 60 seconds to answer these two questions about each headline.

Note: You will receive a **BONUS PAYMENT** of up to **$1.00** (£0.75) based on how many **CORRECT** answers you provide regarding the accuracy of the articles. Correct answers are based on the expert evaluations of non-partisan fact-checkers.

More specifically, if you answer 15 or more of the out 16 questions correctly, you will receive the full bonus payment of $1.00. If you answer 13 or more of the 16 questions correctly, you will receive a partial bonus payment of $0.50. Your bonus payment will be delivered to your Prolific ID. It may take a few weeks to calculate your scores and for you to receive your bonus payment.

We ask you about accuracy on a 6-point scale ranging from "extremely inaccurate" to "extremely accurate." For the purpose of this study, if the headline describes a true event, either "slightly accurate," "moderately accurate," or "extremely accurate" constitute correct responses. Similarly, if the headline describes a false event, either "extremely inaccurate," "moderately inaccurate," or "slightly inaccurate" constitute "correct" responses.

Your answers to all other questions will not contribute to your bonus payment.

After seeing each headline, questions in this condition appeared as follows:

Note: If you answer the question below about accuracy correctly, you have a higher chance of receiving a bonus payment.

To the best of your knowledge, is the claim in the above headline accurate?

Not at all accurate	Moderately innaccurate	Slightly inaccurate	Slightly accurate	Moderately accurate	Extremely accurate

If you were to see the above article on social media, how likely would you be to share it?

Extremely unlikely	Moderately unlikely	Slightly unlikely	Slightly likely	Moderately likely	Extremely likely

Control Text

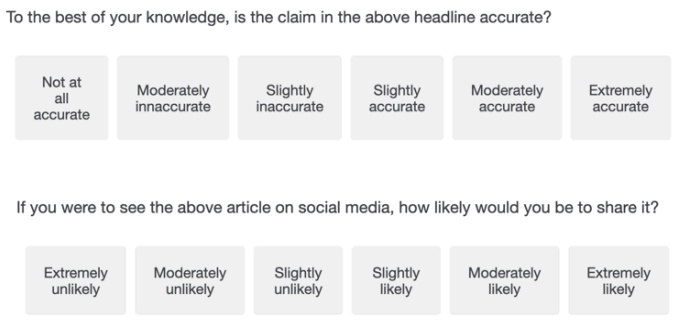
You will be presented with a series of real and fake news headlines. There are 16 headlines in total.

We are interested in your opinion about the following:

1) How accurate is the headline?
2) How likely would you be to share the headline on social media?

You will be given 60 seconds to answer these two questions about each headline.

After seeing each headline, questions in this condition appeared as follows:



Social Incentive Manipulation Text

You will be presented with a series of real and fake news headlines. There are 16 headlines in total.

We are primarily interested in your opinion about the following:

1) How likely is this article to appeal to [Democrats/Republicans]?

We want to see how well you can identify articles that appeal to [Democrats/Republicans]? You will receive a **BONUS PAYMENT** of up to $1.00 (£0.75) based on **how well you identify articles that are likely to appeal to [Democrats/Republicans]?**

More specifically, we have pre-tested these articles to see how much they are liked by [Democrats/Republicans]. We want to see how close your answers are to the answers we identified in the pre-test. If you correctly identify 15 or more out of 16 articles that are liked by [Democrats/Republicans], you will receive the full bonus payment of $1.00. If you correctly identify 13 or more of the 16 articles that are liked by [Democrats/Republicans], you will receive a partial bonus payment of $0.50.

We will also ask you:

2) How accurate is the headline?
3) How likely would you be to share the headline on social media?

But, we will not give you a bonus payment based on your response to these questions. Your answers to all other questions in the survey will not contribute to your bonus payment.

The images may take a second to load. Please wait for the images to load before answering the questions.

After seeing each headline, questions in this condition appeared as follows:

Note: If you correctly predict whether this headline will be liked by Democrats in the below question, you have a higher chance of receiving a bonus payment.

If you shared this article on social media, how likely is it that it would receive a positive reaction from Democrats (e.g., likes, shares, and positive comments)?

Very unlikely	Moderately unlikely	Slightly unlikely	Slightly likely	Moderately likely	Very likely

To the best of your knowledge, is the claim in the above headline accurate?

Not at all accurate	Moderately innaccurate	Slightly inaccurate	Slightly accurate	Moderately accurate	Extremely accurate

If you were to see the above article on social media, how likely would you be to share it?

Extremely unlikely	Moderately unlikely	Slightly unlikely	Slightly likely	Moderately likely	Extremely likely

Mixed Incentives Manipulation Text

You will be presented with a series of real and fake news headlines. There are 16 headlines in total.

We are primarily interested in your opinion about the following:

1) How likely is this article to appeal to [Democrats/Republicans]?
2) How accurate is the headline?

We want to see how well you can identify articles that appeal to [Democrats/Republicans]? You will receive a **BONUS PAYMENT** of up to **$1.00 (£0.75)** based on **how well you identify articles that are likely to appeal to** [Democrats/Republicans]?

You will receive an **ADDITIONAL BONUS PAYMENT** of up to **$1.00 (£0.75)** based on how many **CORRECT** answers you provide regarding the accuracy of the articles. Correct answers are based on the expert evaluations of non-partisan fact-checkers.

More specifically, we have pre-tested these articles to see how much they are liked by [Democrats/Republicans]. We want to see how close your answers are to the answers we identified in the pre-test. If you correctly identify 15 or more out of 16 articles that are liked by [Democrats/Republicans], you will receive the full bonus payment of $1.00. If you correctly identify 13 or more of the 16 articles that are liked by [Democrats/Republicans], you will receive a partial bonus payment of $0.50.

Additionally, if you answer 15 or more out of the 16 questions about accuracy correctly, you will receive the full bonus payment of $1.00. If you answer 13 or more of the 16 questions correctly, you will receive a partial bonus payment of $0.50. Your bonus payment will be delivered to your Prolific ID. It may take a few weeks to calculate your scores and for you to receive your bonus payment.

We ask you about accuracy on a 6-point scale ranging from "extremely inaccurate" to "extremely accurate." For the purpose of this study, if the headline describes a true event, either "slightly

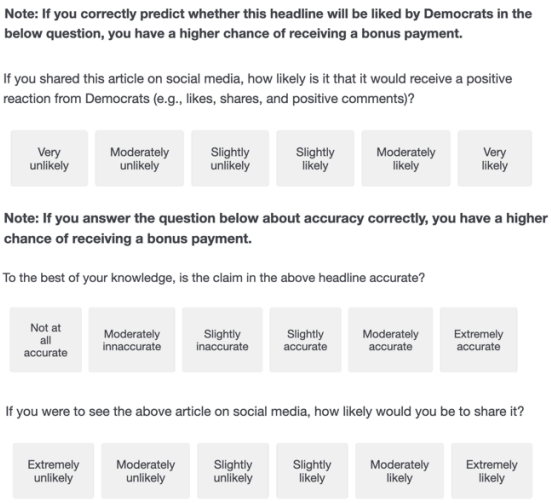
accurate," "moderately accurate," or "extremely accurate" constitute correct responses. Similarly, if the headline describes a false event, either "extremely inaccurate," "moderately inaccurate," or "slightly inaccurate" constitute "correct" responses.

We will also ask you:

3) How likely would you be to share the headline on social media?

But, we will not give you a bonus payment based on your response to this question. Your answers to all other questions in the survey will not contribute to your bonus payment.

After seeing each headline, questions in this condition appeared as follows:



S7: Question Wording

Cognitive Reflection Test.

The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?

If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take 50 printers to print out 50 pages of paper?

On a loaf of bread, there is a patch of mold. Every day, the patch doubles in size. If it takes 40 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?

Affective Polarization.

How favorable do you feel towards Democrats?

How favorable do you feel towards Republicans?

Note: Affective polarization was measured as positive feelings toward the in-party minus negative feelings toward the out-party.

Political Knowledge.

Whose responsibility is it to determine if a law is constitutional or not – is it the President, the Congress, or the Supreme Court?

President

Congress

Supreme Court

How much of a majority is required for the U.S. Senate and House to override a presidential veto?

1/2 majority

1/3 majority

3/4 majority

What party currently has the most members in the House of Representatives in Washington?

Democrats

Republicans

Neither

Would you say that one of the major parties is more conservative than the other at the national level? If so, which party is more conservative?

Democrats

Republicans

Neither

How many justices are on the U.S. Supreme Court?

9

12

18

Political Conservatism.

Which of the following best describes your political preference?

Strongly Democratic

Democratic

Lean Democratic

Lean Republican

Republican

Strongly Republican

Education.

What is the highest level of school you have completed or the highest degree you have received?

Less than high school degree

High school graduate (high school diploma or equivalent including GED)

Some college but no degree

Associate degree in college (2-year)

Bachelor's degree in college (4-year)

Master's degree

Doctoral degree

Professional degree (JD, MD)

Income.

Information about income is very important to understand. Would you please give your best guess?

Please indicate the answer that includes your entire household income in (previous year) before taxes.

Less than $10,000

$10,000 to $19,999

$20,000 to $29,999

$30,000 to $39,999

$40,000 to $49,999

$50,000 to $59,999

$60,000 to $69,999

$70,000 to $79,999

$80,000 to $89,999

$90,000 to $99,999

$100,000 to $149,999

$150,000 or more

Age.

What is your age?

Gender (Recoded for Regression Analysis as Female/Not Female)

What is your gender?

Male

Female

Transgender Female

Transgender Male

Trans/Non-Binary

Not Listed:

Additional Questions

Did you respond randomly at any point during the study?

Note: Please be honest! You will get your payment regardless of your response.

Yes

No

…

Please answer honestly:

Would you ever share an article on social media that you know is false?

Yes

No

Note: These questions were only shown to those in the accuracy incentives condition:

Please answer honestly: this will not affect your payment, and it is important to the researchers to understand how you responded.

Did you ever say an article was accurate simply because you thought it would get you a higher payment and not because you genuinely believed it was accurate?

Or did all answers about accuracy reflect your true beliefs?

Yes

No

…

Please answer honestly: do you think your answers were influenced by the extra financial incentive to be accurate?

Yes

No

Note: These questions were only shown to those in the social incentives condition:

Please answer honestly: this will not affect your payment, and it is important to understand how you responded.

Did you ever say an article was accurate simply because you liked it (or say an article was inaccurate because you disliked it)?

Or did all answers about accuracy reflect your true beliefs?

Yes

No

…

Please answer honestly: do you think your answers to other questions were influenced by your task of identifying articles that would appeal to your political party?

Yes

No

S8: Results for Continuously Coded Outcome Variables

Variable	Mean (Accuracy)	Mean (Control)	Difference	CI_low	CI_high	t	df	p	d	d_CI_low	d_CI_high
Truth Discernment	0.78	0.68	-0.11	0.03	0.18	2.77	1385.76	0.006	0.15	0.04	0.25
Partisan Bias	4.95	6.37	1.42	-2.36	-0.49	-3.00	1401.89	0.003	-0.16	0.26	0.06
Incongruent True News	3.99	3.79	-0.20	0.09	0.31	3.55	1414.72	0.000	0.19	0.08	0.29

Note: Above are the results for truth discernment, partisan bias, and incongruent true news when coded on a continuous, rather than dichotomous scale. Results shown above are from the integrative data analysis. Results do not change the conclusions, but the effect sizes are smaller. This smaller effect can likely be attributed to the fact that incentives rewarded responses as accurate regardless of whether people answered "slightly accurate," "moderately accurate," or "extremely accurate." In other words, it appeared as though people were ignoring magnitude.

S9: Study 3 Results Including Additional News Items

Below, we show the results for Study 3 when including the 8 additional fake news stimuli. These additional analyses do not change our conclusions.

Truth Discernment. A 2X2 (Accuracy X Source) ANOVA found that there was a significant effect of the accuracy incentives on truth discernment such that the accuracy incentives improved discernment, $F(1, 917) = 4.08$, $p = 0.04$, $\eta^2_G = 0.004$. Additionally, there was a significant impact of source cues on truth discernment such that source cues improved accuracy, $F(1, 917) = 8.13$, $p = 0.004$, $\eta^2_G = 0.009$. However, there was no significant interaction between the accuracy incentive condition and source cues, $p = 0.649$.

Partisan Bias. A 2X2 (Accuracy X Source) ANOVA found that there was a significant effect of the accuracy incentives on partisan bias such that the incentives reduced bias, $F(1, 917) = 16.79$, $p < 0.001$, $\eta^2_G = 0.02$. The source cues did not impact partisan bias ($p = 0.603$), and there was no interaction between source cues and partisan bias ($p = 0.438$).

Sharing Discernment. A 2X2 (Accuracy X Source) ANOVA found that there was no significant effect of accuracy ($p = 0.733$), but there was a significant effect of the source cues, $F(1, 917) = 5.50$, $p = 0.019$, $\eta^2_G = 0.01$, and no interaction between accuracy incentives and source cues ($p = 0.533$).

Effects on Sharing Intentions Broken Down by Headline Type. We then ran a 2 (accuracy incentive vs. no incentive) X 2 (social incentive vs. no incentive) X 2 (politically congruent vs. politically incongruent) X 2 (true headlines vs. false headlines) mixed-design ANOVA. We found no main effect of source cues ($p = 0.524$), but did find a main effect of the incentive, $F(1,917) = 0.396$, $p = 0.047$, $\eta^2_G = 0.00$, political congruence, $F(1, 917) = 5.79$, $p = 0.016$, $\eta^2_G = 0.00$, and veracity, $F(1, 917) = 1330.58$, $p < 0.001$, $\eta^2_G = 0.24$. There was also a significant interaction between the incentives and political congruence, $F(1, 917) = 4.12$, $p = 0.043$, $\eta^2_G = 0.00$, the source cues and veracity, $F(1, 917) = 8.18$, $p = 0.004$, $\eta^2_G = 0.00$, incentives and veracity, $F(1, 917) = 4.06$, $p = 0.044$, $\eta^2_G = 0.00$, and political congruence and veracity, $F(1, 917) = 361.44$, $p < 0.001$, $\eta^2_G = 0.09$. There were also significant three-way interactions between source cues, incentives, and political congruence, $F(1, 917) = 5.00$, $p = 0.026$, $\eta^2_G = 0.00$, and incentives, political congruence, and veracity, $F(1, 917) = 16.74$, $p = 0.00$, $\eta^2_G = 0.01$.

Importantly, post-hoc Tukey HSD tests revealed that there was a difference between belief in *incongruent true* headlines rated as accurate in the accuracy incentive (with sources) condition (M = 56.33, 95% CI = [52.92, 59.75]) versus the control (with sources) condition (M = 46.24, 95% CI = [42.79, 49.70]), $p < 0.001$, $d = 0.39$. However, there was no difference between the accuracy incentives condition (without sources cues) as compared to the control condition (without source cues), p = 0.311.

10. Supplementary Materials for "Partisan Differences in the Effectiveness of Priming Accuracy"

Supplementary Section S2. Linear regression analyses at the headline rating level

We primarily report participant-level results in the form of t-tests and moderation analyses in the main body. As a robustness check, we also replicate Pennycook et al.'s rating-level linear regression analyses (clustered on participants and headlines), both to check for differences between Republicans and Democrats (and other measures of conservatism) and for interaction effects between sharing discernment, condition and various measures of conservatism, as well as for performance on the cognitive reflection test (see *Tables S6-S13*; we refer to our OSF page for the full STATA output and analysis script). These analyses broadly show the same results as those reported in the main body: the accuracy prime is effective for Democrats/non-Trump voters, and mostly ineffective for Republicans/Trump voters (see *Tables S7* and *S8*). Further, when pooling the data from all 5 studies together, there is a significant three-way interaction between sharing discernment, condition and all 5 measures of conservatism (Democrat/Republican; Trump voters vs non-Trump voters; social conservatism; economic conservatism; and the average of social and economic conservatism; see *Tables S8-S12)*. These interactions are significant both when participants who indicate not sharing political news on social media are included and excluded. Finally, we find no significant interactions between discernment, condition and CRT performance (see *Table S12)*.

Supplementary Section S2: Asymmetries between Democrats and Republicans

In the replication study by Roozenbeek, Freeman and van der Linden, we find that conservatism does not correlate with the amount of attention checks passed, r = -0.01, 95% CI [-0.06, 0.04], $t(1581) = -0.32$, $p = 0.747$. Additionally, in the Epstein et al. preprint the authors asked us to re-analyze, we find that conservatism very slightly correlates with the number of attention checks passed, $r = 0.07$, 95% CI [0.02, 0.11], $t(1768) = 2.78$, $p = 0.001$. Thus, our current data does not support the idea that differences in (in)attention explain different reactions to the accuracy nudge among liberals and conservatives.

Table S1. Re-Analysis Including Data from Epstein et al., (2021).

	Democrats				Republicans			
Study	Cohen's *d*	95% CI	*p*	*n*	Cohen's *d*	95% CI	*p*	*n*
Study 3 (Pennycook et al., 2021)	0.38	[0.16; 0.60]	0.001	318	0.13	[-0.22; 0.48]	0.483	133
Study 4 (Pennycook et al., 2021)	0.42	[0.21; 0.63]	0.000	364	0.23	[-0.10; 0.55]	0.175	153
Study 5 (Pennycook et al., 2021)	0.23	[-0.02; 0.48]	0.077	277	0.07	[-0.21; 0.35]	0.624	211
Psych Science Paper (Pennycook, McPhetres et al., 2020)	0.32	[0.10; 0.54]	0.005	326	0.17	[-0.07; 0.41]	0.157	271
Psych Science Replication Study (Roozenbeek et al., under review)	0.27	[0.11; 0.42]	0.001	636	0.06	[-0.12; 0.24]	0.522	486
Accuracy Nudge Toolkit Study (Epstein et al., Preprint)	0.19	[0.05; 0.34]	0.021	741	0.08	[-0.12, 0.24]	0.310	599
Mean Effect Size (Pennycook et al. Studies 3-5, 2021)	**0.35**	**[0.22, 0.49]**	**0.001**	**671**	**0.13**	**[-0.05; 0.32]**	**0.150**	**497**
Mean Effect Size (All Studies)	**0.28**	**[0.20; 0.36]**	**0.001**	**2662**	**0.09**	**[0.01; 0.19]**	**0.028**	**1853**

Note: At the request of the authors, we added an additional dataset to the analysis. Specifically, we added waves 2 and 3 from Epstein et al., (2021) including only the accuracy nudge and control conditions in those waves. When adding this additional data, the results for the pooled dataset become significant for Republicans at the $p = 0.05$ level, though the effect size still remains negligible (average $d = 0.09$). While Epstein et al. report moderation analyses by political party in their paper, they pooled all treatment conditions for the moderation analyses (including an accuracy prime condition, but also a "social norms" condition, a "partisan norms" condition, and a "tips" condition). These are very different interventions that are conceivably moderated differently by political partisanship and other covariates. Thus, their moderation analyses are ultimately inconclusive about how partisanship moderates the effect of accuracy primes.

Table S2. Full moderation models for different operationalizations of partisanship (five datasets)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	-0.02	-0.03	-0.03	-0.03	0.01	-0.39
	[-0.23, 0.18]	[-0.26, 0.21]	[-0.21, 0.16]	[-0.23, 0.17]	[-0.23, 0.24]	[-0.84, 0.05]
Condition	0.42 ***	0.38 ***	0.42 ***	0.39 ***	0.25 **	0.61 ***
	[0.29, 0.55]	[0.23, 0.53]	[0.30, 0.53]	[0.26, 0.51]	[0.10, 0.40]	[0.32, 0.89]
Conservatism	0.00					
	[-0.06, 0.07]					
Condition*Conservatism	-0.08 ***					
	[-0.12, -0.04]					
Republican (Dichotomous)		0.00				
		[-0.17, 0.18]				
Condition*Republication		-0.15 **				
		[-0.26, -0.04]				
Social Conservatism			0.00			
			[-0.06, 0.06]			
Condition*Social Conservatism			-0.08 ***			
			[-0.12, -0.04]			
Economic Conservatism				0.00		
				[-0.06, 0.07]		
Condition*Economic Conservatism				-0.06 **		
				[-0.10, -0.02]		
Democrat-Republican (Continuous)					-0.02	
					[-0.20, 0.16]	
Condition*Democrat-Republican					-0.05	
					[-0.16, 0.07]	
Clinton-Trump (Binary)						0.25
						[-0.06, 0.57]
Condition*ClintonTrump						-0.23 *
						[-0.43, -0.03]
	4487	4497	4493	4495	4495	1452
R2	0.04	0.03	0.05	0.03	0.02	0.03

*** p < 0.001; ** p < 0.01; * p < 0.05.

Note: As shown above, there is a significant interaction between condition and partisan affiliation across various measures of partisanship, including: 1) a continuous measure of conservatism (average of social and economic conservatism), 2) a dichotomous measure of whether or not the participant is a Republican, 3) social conservatism (continuous), 4) economic conservatism (continuous), and 5) a dichotomous measure of preference for Clinton or Trump. The final measure was not included in all datasets. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Note: for Study 5, the "accuracy prime" and "accuracy importance" treatment were pooled, following the authors' pre-registered analysis, but the significant interaction effects did not change when excluding the "accuracy importance" treatment.

Table S3. Full moderation models for different operationalizations of partisanship (with the five pooled datasets and the Epstein et al. data)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	-0.18	-0.23	-0.22 *	-0.17	-0.20	-0.89 ***
	[-0.38, 0.03]	[-0.46, 0.01]	[-0.41, -0.04]	[-0.38, 0.03]	[-0.44, 0.03]	[-1.37, -0.41]
Condition	0.41 ***	0.41 ***	0.42 ***	0.36 ***	0.44 ***	0.67 ***
	[0.28, 0.54]	[0.26, 0.56]	[0.31, 0.54]	[0.23, 0.49]	[0.29, 0.59]	[0.37, 0.97]
Conservatism	-0.04					
	[-0.11, 0.03]					
Condition*Conservatism	-0.07 ***					
	[-0.11, -0.03]					
Republican (Dichotomous)		-0.06				
		[-0.23, 0.11]				
Condition*Republication		-0.16 **				
		[-0.27, -0.05]				
Social Conservatism			-0.03			
			[-0.09, 0.04]			
Condition*Social Conservatism			-0.08 ***			
			[-0.12, -0.04]			
Economic Conservatism				-0.04		
				[-0.10, 0.02]		
Condition*Economic Conservatism				-0.05 **		
				[-0.09, -0.01]		
Democrat-Republican (Continuous)					-0.07	
					[-0.22, 0.09]	
Condition*Democrat-Republican					-0.17 ***	
					[-0.27, -0.07]	
Clinton-Trump (Binary)						0.29
						[-0.05, 0.63]
Condition*ClintonTrump						-0.25 *
						[-0.47, -0.04]
N	6257	6267	6263	6265	6252	1452
R2	0.04	0.03	0.05	0.03	0.04	0.03

Table S4. Analysis at each level of political conservatism.

Conservatism Score	*d*	CI low	CI high	*p*	*n*
5	-0.06	-0.23	0.11	0.475	580
4.5	-0.14	-0.41	0.13	0.312	221
4	-0.12	-0.26	0.03	0.108	773
3.5	-0.27	-0.49	-0.06	0.013	353
3	-0.14	-0.23	-0.04	0.004	1917
2.5	-0.29	-0.48	-0.10	0.002	459
2	-0.31	-0.44	-0.18	< 0.001	917
1.5	-0.31	-0.52	-0.11	0.003	381
1	-0.28	-0.42	-0.13	< 0.001	769

Note: The above table shows the effect of the accuracy nudge for each level of political conservatism (total $N = 6{,}472$, all six studies combined, including the additional Epstein et al. data). As shown above, the accuracy nudge is not significant at the highest levels of conservatism, but has small effects for people who score somewhat above the midpoint on the scale. Thus, if the accuracy nudge has small effects for Republicans, this is likely driven by more moderate Republicans.

Table S5. Rating-level linear regressions, clustered on participants and headlines, for Democrats and Republicans, separated by study.

		Democrats			Republicans		
Study	**Variables**	*B*	SE	*p*	*B*	SE	*p*
Study 3 (Pennycook et al., 2021)	real	0.238	0.152	0.119	0.329	0.178	0.0638
	treatment	-0.248	0.09	0.00611	-0.317	0.126	0.0122
	realxtreatment	**0.324**	0.06	6.71E-08	0.148	0.086	0.0847
	Constant	2.372	0.113	0	2.527	0.141	0
	Observations	11,647			5,746		
	R-squared	0.02			0.021		
Study 4 (Pennycook et al., 2021)	real	0.239	0.215	0.265	0.217	0.22	0.325
	treatment	-0.336	0.103	0.00108	-0.172	0.128	0.179
	realxtreatment	**0.368**	0.096	0.00012	**0.204**	0.099	0.0396
	Constant	2.498	0.166	0	2.72	0.159	0
	Observations	12,997			5,656		
	R-squared	0.02			0.009		
Study 5 (Pennycook et al., 2021)	real	0.168	0.175	0.335	0.109	0.207	0.597
	treatment	-0.275	0.177	0.122	-0.423	0.173	0.0148
	realxtreatment	**0.245**	0.103	0.0174	0.183	0.098	0.0611
	Constant	3.168	0.126	0	3.514	0.162	0
	Observations	7,333			6,007		
	R-squared	0.005			0.008		
Psych Science paper (Pennycook, McPhetres et al., 2020)	real	0.184	0.073	0.011	-0.0043	0.1	0.965
	treatment	-0.021	0.139	0.879	0.0596	0.147	0.685
	realxtreatment	**0.227**	0.065	0.00053	0.106	**0.039**	0.00687
	Constant	3.369	0.113	0	3.301	0.118	0
	Observations	13,665			11,812		
	R-squared	0.008			0.001		
Psych Science replication study (Roozenbeek et al., 2021)	real	0.31	0.142	0.029	0.011	0.109	0.92
	treatment	-0.225	0.09	0.0123	-0.057	0.104	0.584
	realxtreatment	**0.132**	0.043	0.00237	0.0261	0.034	0.442
	Constant	3.024	0.108	0	3.059	0.098	0
	Observations	27,120			20,370		
	R-squared	0.013			0		

All 5 studies combined	real	0.245	0.092	0.00786	0.0838	0.081	0.298
	treatment	-0.247	0.057	1.43E-05	-0.138	0.069	0.0461
	realxtreatment	**0.242**	0.043	1.73E-08	0.0828	0.046	0.0732
	Constant	2.922	0.071	0	3.091	0.068	0
	Observations	72,762			49,591		
	R-squared	0.012			0.002		

Note: Significant discernment predictors marked in bold. "Real" refers to the sharing discernment dummy. As shown above, the accuracy prime treatment is broadly effective for Democrats, but not for Republicans, in support of the results reported in the main body.

Table S6. Rating-level linear regressions, clustered on participants and headlines, for Trump voters and non-Trump voters, separated by study.

		Non-Trump voters			Trump voters		
Study	**Variables**	*B*	SE	*p*	*B*	SE	*p*
Study 3 (Pennycook et al., 2021)	real	0.258	0.13	0.0476	0.307	0.188	0.103
	treatment	-0.255	0.083	0.00212	-0.287	0.165	0.0819
	realxtreatment	**0.323**	0.06	5.99E-08	0.0077	0.055	0.888
	Constant	2.377	0.098	0	2.581	0.161	0
	Observations	13,659			3,734		
	R-squared	0.021			0.016		
Study 4 (Pennycook et al., 2021)	real	0.202	0.204	0.322	0.309	0.287	0.28
	treatment	-0.301	0.1	0.0025	-0.2	0.156	0.201
	realxtreatment	**0.335**	0.093	0.00031	**0.308**	0.098	0.00157
	Constant	2.513	0.159	0	2.752	0.205	0
	Observations	14,640			3,989		
	R-squared	0.016			0.018		
Study 5 (Pennycook et al., 2021)	real	0.074	0.155	0.634	0.265	0.213	0.212
	treatment	-0.335	0.157	0.0328	-0.334	0.197	0.0894
	realxtreatment	**0.233**	0.09	0.00918	0.156	0.1	0.12
	Constant	3.24	0.105	0	3.459	0.17	0
	Observations	8,423			4,917		
	R-squared	0.004			0.01		
Psych Science paper (Pennycook, McPhetres et al., 2020)	real	0.13	0.078	0.0963	0.0473	0.098	0.628
	treatment	0.039	0.125	0.753	-0.016	0.171	0.928
	realxtreatment	**0.203**	0.054	0.00019	0.0804	0.041	0.0507
	Constant	3.277	0.106	0	3.444	0.127	0
	Observations	16,395			9,202		
	R-squared	0.006			0.001		
Psych Science replication study (Roozenbeek et al., under review)	real	0.251	0.132	0.0573	0.034	0.116	0.77
	treatment	-0.22	0.082	0.00718	-0.02	0.124	0.871
	realxtreatment	**0.116**	0.039	0.00302	0.0188	0.042	0.653
	Constant	3.008	0.099	0	3.11	0.114	0
	Observations	32,490			15,000		
	R-squared	0.01			0		
All 5 studies combined	real	0.198	0.085	0.02	0.137	0.086	0.11

treatment	-0.228	0.054	2.68E-05	-0.126	0.079	0.112
realxtreatment	**0.226**	0.044	3.37E-07	0.0495	0.044	0.255
Constant	2.915	0.066	0	3.164	0.075	0
Observations	85,607			36,842		
R-squared	0.009			0.003		

Note: Significant discernment predictors marked in bold. "Real" refers to the sharing discernment dummy. As shown above, the accuracy prime treatment is broadly effective for non-Trump voters, but not for Trump voters, in support of the results reported in the main body.

Table S7. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) for Democrats vs Republicans (z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.18	0.161
	SE	-0.0723	-0.0806
	p	0.013	0.0454
treatment	*B*	-0.203	-0.154
	SE	-0.0459	-0.0389
	p	9.93E-06	7.33E-05
zdemrep	*B*	0.0832	0.0679
	SE	-0.0419	-0.0379
	p	0.047	0.0731
realxtreatment	*B*	0.177	0.191
	SE	-0.0367	-0.0311
	p	1.36E-06	9.46E-10
realXzdemrep	*B*	-0.0789	-0.0524
	SE	-0.0495	-0.0457
	p	0.111	0.252
treatmentXzdemrep	*B*	0.0532	0.0763
	SE	-0.042	-0.0362
	p	0.205	0.0349
realXtreatmentxzdemrep	*B*	**-0.078**	**-0.0819**
	SE	-0.0247	-0.0223

	p	0.0016	0.000246
Constant		2.99	2.765
		-0.0557	-0.0588
		0	0
Observations		122,353	156,275
R2		0.009	0.008

Note: 'Real' refers to the sharing discernment dummy. Relevant significant predictors (three-way interaction between discernment, treatment & being Republican) are marked in bold.

Table S8. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) for social conservatism (z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.177	0.16
	SE	-0.0714	-0.08
	p	0.013	0.0458
treatment	*B*	-0.206	-0.152
	SE	-0.0449	-0.0378
	p	4.38E-06	5.67E-05
zsoccon	*B*	0.117	0.104
	SE	-0.0419	-0.0403
	p	0.00513	0.0096
realxtreatment	*B*	0.18	0.189
	SE	-0.0354	-0.0297
	p	3.56E-07	2.14E-10
realXzsoccon	*B*	-0.101	-0.0779
	SE	-0.0496	-0.0485
	p	0.0424	0.108
treatmentXzsoccon	*B*	0.126	0.153
	SE	-0.0418	-0.0388
	p	0.00267	7.73E-05
realXtreatmentxzsoccon	*B*	**-0.0982**	**-0.104**
	SE	-0.0248	-0.0269

	p	7.66E-05	0.000113
Constant		2.993	2.766
		-0.0547	-0.058
		0	0
Observations		122,298	156,084
R2		0.012	0.012

Note: 'Real' refers to the sharing discernment dummy. Relevant significant predictors (three-way interaction between discernment, treatment & social conservatism) are marked in bold.

Table S9. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) for economic conservatism (z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.179	0.16
	SE	-0.0732	-0.0819
	p	0.0147	0.0506
treatment	*B*	-0.205	-0.158
	SE	-0.0456	-0.0387
	p	6.99E-06	4.59E-05
zeconcon	*B*	-0.0278	-0.0302
	SE	-0.0395	-0.0375
	p	0.481	0.421
realxtreatment	*B*	0.179	0.19
	SE	-0.0358	-0.0304
	p	5.68E-07	3.63E-10
realXzeconcon	*B*	-0.0739	-0.0588
	SE	-0.0459	-0.0447
	p	0.107	0.188
treatmentXzeconcon	*B*	0.15	0.162
	SE	-0.0417	-0.0382
	p	0.00031	2.32E-05
realXtreatmentxzeconcon	*B*	**-0.0851**	**-0.0836**
	SE	-0.0244	-0.0245

	p	0.000482	0.00064
Constant		2.992	2.767
		-0.0564	-0.0597
		0	0
Observations		122,395	156,257
R2		0.009	0.008

Note: 'Real' refers to the sharing discernment dummy. Relevant significant predictors (three-way interaction between discernment, treatment & economic conservatism) are marked in bold.

Table S10. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) for conservatism (combined measure of social and economic conservatism, z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.178	0.161
	SE	-0.0727	-0.0813
	p	0.0143	0.0477
treatment	*B*	-0.207	-0.156
	SE	-0.0452	-0.0381
	p	4.88E-06	4.30E-05
zcons	*B*	0.0475	0.0399
	SE	-0.042	-0.0406
	p	0.258	0.325
realxtreatment	*B*	0.18	0.189
	SE	-0.0355	-0.0298
	p	4.16E-07	2.45E-10
realXzcons	*B*	-0.0921	-0.073
	SE	-0.0501	-0.0494
	p	0.0661	0.14
treatmentXzcons	*B*	0.143	0.165
	SE	-0.0418	-0.0388
	p	0.000617	2.08E-05
realXtreatmentxzcons	*B*	**-0.0947**	**-0.0982**

	SE	-0.0247	-0.0261
	p	0.000131	0.00017
Constant		2.993	2.766
		-0.0558	-0.0591
		0	0
Observations		122,527	156,569
R2		0.01	0.009

Note: 'Real' refers to the sharing discernment dummy. Relevant significant predictors (three-way interaction between discernment, treatment & conservatism) are marked in bold.

Table S11. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) and for voting for Trump (z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.18	0.161
	SE	-0.0722	-0.08
	p	0.0126	0.0442
treatment	*B*	-0.198	-0.149
	SE	-0.0458	-0.0389
	p	1.54E-05	0.000122
ztrumpvote	*B*	0.113	0.117
	SE	-0.0397	-0.0366
	p	0.00427	0.00141
realxtreatment	*B*	0.175	0.187
	SE	-0.0369	-0.0314
	p	2.08E-06	2.30E-09
realXztrumpvote	*B*	-0.0275	-0.0114
	SE	-0.0454	-0.0428
	p	0.544	0.789
treatmentXztrumpvote	*B*	0.0462	0.0521
	SE	-0.0424	-0.0373
	p	0.276	0.163
realXtreatmentxztrumpvote	*B*	**-0.0803**	**-0.0775**

	SE	-0.0242	-0.0217
	p	0.000906	0.000345
Constant		2.987	2.764
		-0.0556	-0.0583
		0	0
Observations		122,449	156,423
R2		0.01	0.01

Note: 'Real' refers to the sharing discernment dummy. Relevant significant predictors (three-way interaction between discernment, treatment & voting for Trump) are marked in bold.

Table S12. Interaction analyses at the rating level (linear regression, clustered on participants and headlines) and CRT score (cognitive reflection test performance, z-scored).

		All 5 studies	All 5 studies (never-sharers included)
Variables	**Statistic**		
real	*B*	0.18	0.158
	SE	-0.0648	-0.0747
	p	0.00561	0.0347
treatment	*B*	-0.21	-0.145
	SE	-0.0417	-0.0363
	p	4.97E-07	6.52E-05
zcrt	*B*	-0.411	-0.379
	SE	-0.0289	-0.0248
	p	0	0
realxtreatment	*B*	0.178	0.19
	SE	-0.0323	-0.0301
	p	3.66E-08	2.38E-10
realXzcrt	*B*	0.105	0.0831
	SE	-0.0233	-0.0241
	p	6.36E-06	0.000583
treatmentXzcrt	*B*	-0.057	-0.0631
	SE	-0.0382	-0.033
	p	0.135	0.0562
realXtreatmentxzcrt	*B*	0.00416	0.0255

	SE	-0.0245	-0.025
	p	0.866	0.307
Constant		2.99	2.762
		-0.0494	-0.0534
		0	0
Observations		122,551	158,280
R2		0.05	0.046

Note: 'Real' refers to the sharing discernment dummy.

Supplemental References

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*.

Pennycook, G., & Rand, D. (2021). Reducing the spread of fake news by shifting attention to accuracy: Meta-analytic evidence of replicability and generalizability. *Working Paper*.

11. Supplementary Materials for "Social Media Behavior is Correlated with Vaccine Hesitancy"

Section S1: Question Wording

Below we list the exact question wording for the survey questions participants were asked that are relevant to this study for Sample 1 and Sample 2.

Sample 1:

COVID-19 Vaccination Intentions:

Have you or do you intend to receive the COVID-19 vaccine when you are eligible to do so?

- Yes, I have already received at least one dose of the COVID-19 vaccine
- Yes, I haven't received my first dose, but I intend to do so
- No, I haven't received my first dose and I am uncertain about whether I will get one
- No, I haven't received my first dose and I do not intend to get one
- I cannot get vaccinated against COVID-19 due to medical reasons

COVID-19 Vaccine Safety & Efficacy:

The currently available COVID-19 vaccines are…:

- Safe [1 "strongly disagree", 7 "strongly agree]
- Effective in preventing the disease [1 "strongly disagree", 7 "strongly agree]

Political Ideology

[United States]

Which of the following best describes your political preference?

- Extremely liberal
- Liberal
- Slightly liberal
- Moderate
- Slightly conservative
- Conservative
- Extremely conservative

[United Kingdom]

Which of the following best describes your political preference?

- Extremely left-wing/liberal
- Left-wing/liberal
- Slightly left-wing/liberal
- Middle of the road
- Slightly right-wing/conservative
- Right-wing/conservative
- Extremely right-wing/conservative

Age:

What is your year of birth?

Gender:

What is your gender?

- Male
- Female
- Transgender Female
- Transgender Male
- Trans/Non-Binary
- Not Listed
- Prefer not to Say

Education:

[United States]

What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some college but no degree
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Doctoral degree
- Professional degree (JD, MD)

[United Kingdom]

What is the highest level of education that you completed?

- No formal education above age 16
- Professional or technical qualifications above age 16
- School education up to age 18
- Degree (Bachelor's) or equivalent
- Degree (Master's) or other postgraduate qualification
- Doctorate

US or UK:

Are you a resident of the United Kingdom or the United States?

Country:

In which country do you currently reside?

[drop-down list of countries]

Sample 2:

COVID-19 Vaccine Confidence:

How likely are you to get vaccinated for COVID-19 when it becomes available? (0 = very unlikely and 100 = very likely)? If you have already received the vaccine, you may select 100.

Political Orientation:

What is your political orientation?

- Extremely liberal
- Liberal
- Slightly liberal
- Moderate
- Slightly conservative
- Conservative
- Extremely conservative

Gender:

What is your gender?

- Male
- Female

- Transgender Male
- Transgender Female
- Non-Binary/Other

Age:

How old are you?

Education:

What is the highest level of education you've completed?

- High School or Less
- Some College
- Bachelor's Degree
- Higher Degree

Please choose whichever ethnicity that you identify with (you may choose more than one option):

- White/Caucasian
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Other

Table S1. Demographics of Each Sample

Study 1 (Full Sample) (N=1246)		Study 1 (Twitter Handles Only) (N=464)		Study 2 (App Dataset) (N = 1600)	
Vaccine Confidence		**Vaccine Confidence**		**Likelihood of Getting Vaccine**	
Mean (SD)	5.34 (3.29)	Mean (SD)	5.32 (1.49)	Mean (SD)	93.3 (21.6)
Median [Min, Max]	5.50 [1.00, 100]	Median [Min, Max]	6.00 [1.00, 7.00]	Median [Min, Max]	100 [0, 100]
Missing	34 (2.7%)	Missing	1 (0.2%)	**Age**	
Will Get Vaccine		**Will Get Vaccine**		Mean (SD)	38.4 (12.6)
Will not get vaccine	349 (27.3%)	Will not get vaccine	118 (25.4%)	Median [Min, Max]	36.0 [-53.0, 77.0]
Will get vaccine	896 (70.0%)	Will get vaccine	345 (74.4%)	Missing	503 (31.4%)
Missing	35 (2.7%)	Missing	1 (0.2%)	**Political Conservatism**	
Country		**Country**		Mean (SD)	2.64 (1.50)
UK	548 (42.8%)	UK	225 (48.5%)	Median [Min, Max]	2.00 [1.00, 7.00]
US	454 (35.5%)	US	158 (34.1%)	Missing	206 (12.9%)
Other/Missing	278 (21.7%)	Other/Missing	81 (17.6%)	**Gender**	
UK or US		**UK or US**		Male	749 (46.8%)
United Kingdom	281 (22.0%)	UK	120 (25.9%)	Female	619 (38.7%)
United States	372 (29.1%)	US	123 (26.5%)	Transgender Female	2 (0.1%)
Neither	15 (1.2%)	Other	6 (1.3%)	Transgender Male	1 (0.1%)
Missing	612 (47.8%)	Missing	215 (46.3%)	Non-Binary/Other	30 (1.9%)
Age		**Age**		Missing	199 (12.4%)
Mean (SD)	34.9 (12.1)	Mean (SD)	37.7 (12.5)	**Education**	

Median [Min, Max]	32.0 [18.0, 73.0]	Median [Min, Max]	36.0 [18.0, 73.0]	High School or Less	69 (4.3%)
Missing	251 (19.6%)	Missing	74 (16.1%)	Some College	218 (13.6%)
Gender		**Gender**		Bachelor's Degree	421 (26.3%)
Male	465 (36.3%)	Male	175 (38.0%)	Higher Degree	686 (42.9%)
Female	556 (43.4%)	Female	210 (45.3%)	Missing	206 (12.9%)
Transgender Female	3 (0.2%)	Transgender Female	0 (0.0%)	**Ethnicity**	
Transgender Male	9 (0.7%)	Transgender Male	5 (1.1%)	White/Caucasian	1148 (71.8%)
Non-Binary/Other	2 (0.2%)	Non-Binary/Other	1 (0.2%)	Black or African American	33 (2.1%)
Prefer Not To Answer	1 (0.1%)	Not Listed	0 (0%)	American Indian or Alaska Native	3 (0.2%)
Missing	244 (19.1%)	Missing	73 (15.8%)	Asian	71 (4.4%)
Political Conservatism		**Political Conservatism**		Native Hawaiian or Pacific Islander	3 (0.2%)
Mean (SD)	3.97 (1.95)	Mean (SD)	4.10 (1.93)	Other	67 (4.2%)
Median [Min, Max]	5.00 [1.00, 7.00]	Median [Min, Max]	5.00 [1.00, 7.00]	Multiple Options Selected	66 (4.1%)
Missing	612 (47.8%)	Missing	214 (46.4%)	Missing	209 (13.1%)
Bachelors		**Bachelors**		**Followers**	
No Bachelor's Degree	256 (20.0%)	No Bachelor's Degree	99 (21.3%)	Mean (SD)	1370 (3810)
Bachelor's Degree	412 (32.2%)	Bachelor's Degree	150 (32.3%)	Median [Min, Max]	330 [0, 45900]
Missing	612 (47.8%)	Missing	215 (46.3%)	**# of Accounts Followed**	
		# of Accounts Followed		Mean (SD)	1040 (1700)
		Mean (SD)	572 (1010)	Median [Min, Max]	544 [0, 34300]
		Median [Min, Max]	189 [1.00, 5000]	**Number of Tweets**	
				Mean (SD)	2030 (1290)
				Median [Min, Max]	2840 [1.00, 3200]
				Number of Hyperpartisan Sites Shared	

Mean (SD)	0.767 (5.34)
Median [Min, Max]	0 [0, 151]

Note. Descriptive Statistics for all samples. Some descriptive statistics from Study 1 are missing because 1) these questions were added to the survey partway through data collection, or 2) participants chose not to answer certain questions. Because we had different questions assessing education in the US and the UK, we re-coded this particular variable to measure whether a participant had a bachelor's degree for the purpose of presenting descriptives for the overall sample. Additionally, since we had two variables measuring country (a drop-down list versus a UK/US/other question), when subsetting the data for network analysis, we included participants who said US on either of the questions in the US dataset for network analysis, and included participants who said UK on either of the questions in the UK dataset for network analysis.

Table S2. Study 1 Regression Models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
(Intercept)	0.00 [-0.09, 0.09]	0.00 [-0.09, 0.09]	0.00 [-0.09, 0.09]	0.00 [-0.09, 0.09]	-0.15 [-0.38, 0.08]	-0.14 [-0.38, 0.09]	-0.16 [-0.39, 0.07]	-0.15 [-0.38, 0.08]
Republicans	-0.12 * [-0.21, -0.03]				-0.18 ** [-0.30, -0.05]			
Tory		0.06 [-0.04, 0.15]				0.03 [-0.10, 0.16]		
HyperPartisan			-0.15 ** [-0.24, -0.06]				-0.20 ** [-0.32, -0.08]	
BadHandles				-0.19 *** [-0.28, -0.10]				-0.19 ** [-0.32, -0.06]
Politics					-0.30 *** [-0.43, -0.17]	-0.32 *** [-0.45, -0.19]	-0.29 *** [-0.42, -0.17]	-0.29 *** [-0.42, -0.16]
Age					0.11 [-0.02, 0.24]	0.09 [-0.04, 0.22]	0.11 [-0.02, 0.24]	0.13 [-0.01, 0.26]
GenderRecode					-0.02 [-0.27, 0.23]	0.01 [-0.25, 0.27]	0.00 [-0.25, 0.25]	0.00 [-0.25, 0.26]
Bachelors					0.26 * [0.00, 0.51]	0.23 [-0.03, 0.49]	0.25 * [0.00, 0.51]	0.25 [-0.01, 0.50]
followers					0.06 [-0.10, 0.21]	0.06 [-0.10, 0.22]	0.07 [-0.09, 0.22]	0.06 [-0.09, 0.22]
friends					-0.07	-0.08	-0.08	-0.07

					[-0.23, 0.09]	[-0.24, 0.08]	[-0.24, 0.08]	[-0.23, 0.09]
N	460	460	460	460	231	231	231	231
R2	0.01	0.00	0.02	0.03	0.15	0.12	0.15	0.15

Note: above are regression models examining how the number of Republican Accounts, UK Conservative Accounts, Hyperpartisan Accounts, and NewsGuard "Low Quality" accounts one follows predict vaccine confidence without (Models 1-4) and with (Models 5-8) control variables.

Table S3. Study 1 Supplementary Regression Models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
(Intercept)	0.00	0.00	0.00	0.00	0.01	-0.00	0.00	0.00
	[-0.09, 0.09]	[-0.09, 0.09]	[-0.09, 0.09]	[-0.09, 0.09]	[-0.16, 0.18]	[-0.17, 0.17]	[-0.17, 0.18]	[-0.17, 0.17]
Republicans	-0.12 *				-0.14 *			
	[-0.21, -0.03]				[-0.26, -0.02]			
Tory		0.06				0.05		
		[-0.04, 0.15]				[-0.07, 0.17]		
HyperPartisan			-0.15 **				-0.14 *	
			[-0.24, -0.06]				[-0.26, -0.02]	
BadHandles				-0.19 ***				-0.15 *
				[-0.28, -0.10]				[-0.27, -0.02]
Politics					-0.33 ***	-0.35 ***	-0.33 ***	-0.32 ***
					[-0.46, -0.21]	[-0.48, -0.22]	[-0.45, -0.20]	[-0.45, -0.20]
GenderRecode					-0.02	0.00	-0.01	-0.00
					[-0.26, 0.22]	[-0.25, 0.25]	[-0.25, 0.23]	[-0.25, 0.24]
Age					0.09	0.07	0.09	0.10
					[-0.03, 0.22]	[-0.05, 0.19]	[-0.03, 0.21]	[-0.02, 0.22]
N	463	463	463	463	245	245	245	245
R2	0.01	0.00	0.02	0.03	0.13	0.11	0.13	0.13

Note: above are regression models examining how the number of Republican Accounts, UK Conservative Accounts, Hyperpartisan Accounts, and NewsGuard "Low Quality" accounts one follows predict vaccine confidence without (Models 1-4) and with (Models 5-8) a different set of control variables used in a prior draft of this manuscript.

Table S4. Study 1 Additional Supplementary Regression Models

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.00	0.00	0.00	0.00
	[-0.09, 0.09]	[-0.09, 0.09]	[-0.09, 0.09]	[-0.09, 0.09]
UK Conservatives (Tories) Followed - UK Liberals (Labor Party) Followed	0.02			
	[-0.08, 0.11]			
US Conservatives (Republicans) Followed - US Liberals (Democrats) Followed		-0.10 *		
		[-0.19, -0.01]		
# of US Democrats Followed			0.05	
			[-0.04, 0.14]	
# of UK Labor Party Members Followed				0.06
				[-0.03, 0.16]
N	460	460	460	460
R2	0.00	0.01	0.00	0.00

Supplementary analysis showing how 1) the number of UK conservatives minus the number of UK liberals one follows predicts vaccine confidence, 2) how the number US conservatives minus the number of US liberals one follows predicts vaccine confidence, 3) how the number of Democrats one follows predicts vaccine confidence, and 4) how the number of labor party members one follows predicts vaccine confidence.

Table S4. Network Statistics

	US Twitter Network		UK Twitter Network	
Network Level Statistics				
Average Path length	3.398		3.25	
Number of influencers followed by at least 3 participants	614		2409	
Modularity between political attitude communities	0.228		0.214	
Modularity between vaccine attitude communities	0.194		0.186	
Assortativity based on political conservatism	0.726		0.566	
Assortativity based on vaccine confidence	0.549		0.391	
Community Level Statistics				
Structural Communities	Community A	Community B	Community A	Community B
Number of nodes (participants and "influencers")	583	140	1728	799
Number of participants	95	14	107	11
Nominal assortativity between communities	0.79		0.77	
Average political conservatism among participants	3.43	6.07	4.41	5.09
	[3.00, 3.86]	[5.80, 6.35]	[4.10, 4.72]	[4.33, 5.85]

Community difference in political conservatism	-2.64***		-0.68 NS	
	[-3.14, -2.14]		[-1.49, 0.13]	
Average vaccine confidence	5.98	4.46	6.06	6.50
	[5.72, 6.24]	[3.45, 5.47]	[5.87, 6.25]	[6.24, 6.76]
Community difference in vaccine confidence	1.51**		-0.44**	
	[0.48, 2.55]		[-0.75, -0.13]	
Node Level Statistics				
Correlation between centrality within structural community and political conservatism	-0.23*	0.16 NS	-0.02 NS	0.18 NS
	[-0.40, -0.04]	[-0.03, 0.33]	[-0.20, 0.16]	[-0.001, 0.35]
Correlation between centrality within structural community and vaccine confidence	0.08 NS	-0.22*	0.07 NS	0.15 NS
	[-0.11, 0.27]	[-0.39, -0.03]	[-0.12, 0.24]	[-0.04, 0.32]
Correlation between pairwise structural distance and pairwise difference in political conservatism	0.08***		0.02 NS	
	[0.06, 0.11]		[-0.002, 0.05]	
Correlation between pairwise structural distance and pairwise difference	0.05***		0.04 **	
	[0.02, 0.07]		[0.01, 0.06]	

in vaccine confidence	

Note: Network descriptive and inferential statistics in the US and the UK at the network level, community level, and pair level, are shown above. "Network level" refers to overall descriptive statistics about the US and the UK networks as a whole, community level refers to the statistics within each community in the US and the UK (Community A and Community B), and pair-level refers to the statistics about each pair of nodes in the US and UK networks. Assortativity based on an attitude measures the likelihood that connected nodes have a similar attitude; nominal assortativity measures the likelihood that connected nodes belong to the same cluster. 95% confidence intervals are shown in square brackets where applicable. NS denotes not significant, * denotes $p < 0.05$, ** denotes $p < 0.01$, *** denotes $p < 0.001$.

Section S2. Supplementary Network Analysis.

Description of Network Statistics. The average pathlength statistic is calculated by averaging the lengths of the shortest path between every pair of nodes in a network. A lower average pathlength reflects greater connectivity within a network. Here, a "path" refers to the number of edges connecting one node to another and does not refer to the absolute distance between two nodes on a graph visualization. Modularity is a measure for graph partitioning quality – in other words, the strength of division of a network into different sections, or modules. We assigned community memberships to nodes based on their self-reported political or vaccine attitudes (or for influencers, the average political or vaccine attitudes of the participants following them), and used the modularity coefficient to measure the extent to which political or vaccine opinion differences structurally partition the networks. The assortativity coefficient measures the likelihood that connected nodes have a similar attribute (i.e., political conservatism and vaccine confidence). We were not able to perform statistical tests for the differences between these descriptive statistics between the US and the UK. We also calculated the nominal assortativity between the communities in each country, or the likelihood that connected nodes belong to the same structural community.

Different Network Sizes. A visual inspection of the network visualizations in our main text suggests that the UK network is much denser than the US network. This is because the UK has many more influencers than in the US. With our exclusion criteria of eliminating influencers with less than 3 participants, the US has 614 influencers, while the UK has 2,409 influencers. We further found that while 90% of the "influencers" in the UK only have one follower among our participants, 95% of the US "influencers" only have one among our participants. This means that even by only filtering out the influencers followed by less than 2 participants, we lose more influencers in the US than in the UK. This difference in the effect of filtering, in addition to the US raw dataset already containing fewer influencers (despite more participants; see **Table S5** for details) than the UK, explains the apparent difference in the size of the US and the UK networks.

Table S6. Network size statistics under different influencer inclusion criteria

	US			UK		
Number of	Influencer	Participant	Edge	Influencer	Participant	Edge
N ≥ 5	150	99	1067	709	112	5417
N ≥ 4	279	105	1583	1247	116	7569
N ≥ 3	614	109	2588	2409	118	11055
N ≥ 2	2414	117	6188	6465	120	19167
N ≥ 1	46502	124	50276	64458	123	77160

Node: "N ≥ *x*" denotes influencer inclusion criteria where influencers followed by less than x participants are excluded. "N ≥ 1" is the same as not filtering; "N ≥ 3" is the inclusion criteria chosen in the main analysis.

Robustness Check of Influencer Inclusion Criteria in Network Analysis. In the main analysis, we set an inclusion criterion that the nodes in the network graphs only include 1) influencers that are followed by at least 3 participants, and 2) participants following at least 1 influencer. To ensure that the inclusion criteria did not significantly alter the network structures, we conducted the following additional robustness check on how different inclusion criteria for influencers may impact the network topology.

First, we must acknowledge that, in the strictest definition, network topology will be changed when removing any node or edge from the network (Easley & Kleinberg, 2010). Details of the network topology that are irrelevant to analysis, however, can be safely reduced without impacting the network topologies of interest, in our case, polarized communities. In fact, many graph partitioning algorithms that identify topological structures (or structural communities) in networks, such as the Girvan-Newman method (Girvan & Newman, 2002), function by selectively removing nodes and edges to reduce topological details of the network and highlight

structural communities. Our inclusion criteria for influencers aim to reduce structural details irrelevant to community detection while retaining as much relevant information as possible.

We can compare assortativity coefficients as a descriptive measure of structural separation of opinions (i.e., polarization) at different conditions, as shown in **Table S6.**

Table S7. Assortativity coefficients under different influencer inclusion criteria

	Assortativity coefficient for political conservatism		**Assortativity coefficient for vaccine confidence**	
Country	US	UK	US	UK
Including N ≥ 5	0.704	0.448	0.456	0.254
Including N ≥ 4	0.716	0.506	0.495	0.315
Including N ≥ 3	0.726	0.566	0.549	0.391
Including N ≥ 2	0.773	0.670	0.690	0.531
Including N ≥ 1	0.978	0.945	0.967	0.916

Note: "Including N ≥ x" denotes inclusion criteria for influencers, where we filter out influencers that are followed by less than x participants. "Including N ≥ 1" is the same as not filtering, as all influencers are followed by at least one participant. "Including N ≥ 3" is our chosen inclusion criteria in the main analysis.

As shown in ***Table S6***, as inclusion criteria decrease, assortativity scores increase for both political conservatism and vaccine confidence in both countries. This is due to a lower

inclusion criterion including more influencers that we assign attributes to based on their follower's attributes, which inflates the likelihood that connected nodes will have similar attributes (the definition of assortativity). When not filtering, the assortativity coefficient approaches 1, because the majority of edges in the graph are participants following a lone "influencer" that is not followed by anyone else, meaning that most of the connected nodes will have similar attributes. Because we do not know the influencers' attributes, we cannot perform network homophily or polarization analysis on a dataset where more than 90% of the attributes are inferred. However, it is worth noting that even though the UK has far more edges than the US under all inclusion criteria (***Table S5***), making it more likely to have connected nodes that have similar attributes, it appears to have lower assortativity than the US in all conditions. This shows that while changing inclusion criteria does change certain measures of network topology such as assortativity, the comparison between network topologies under the same criteria still reveals the same patterns of greater structural separation in the US.

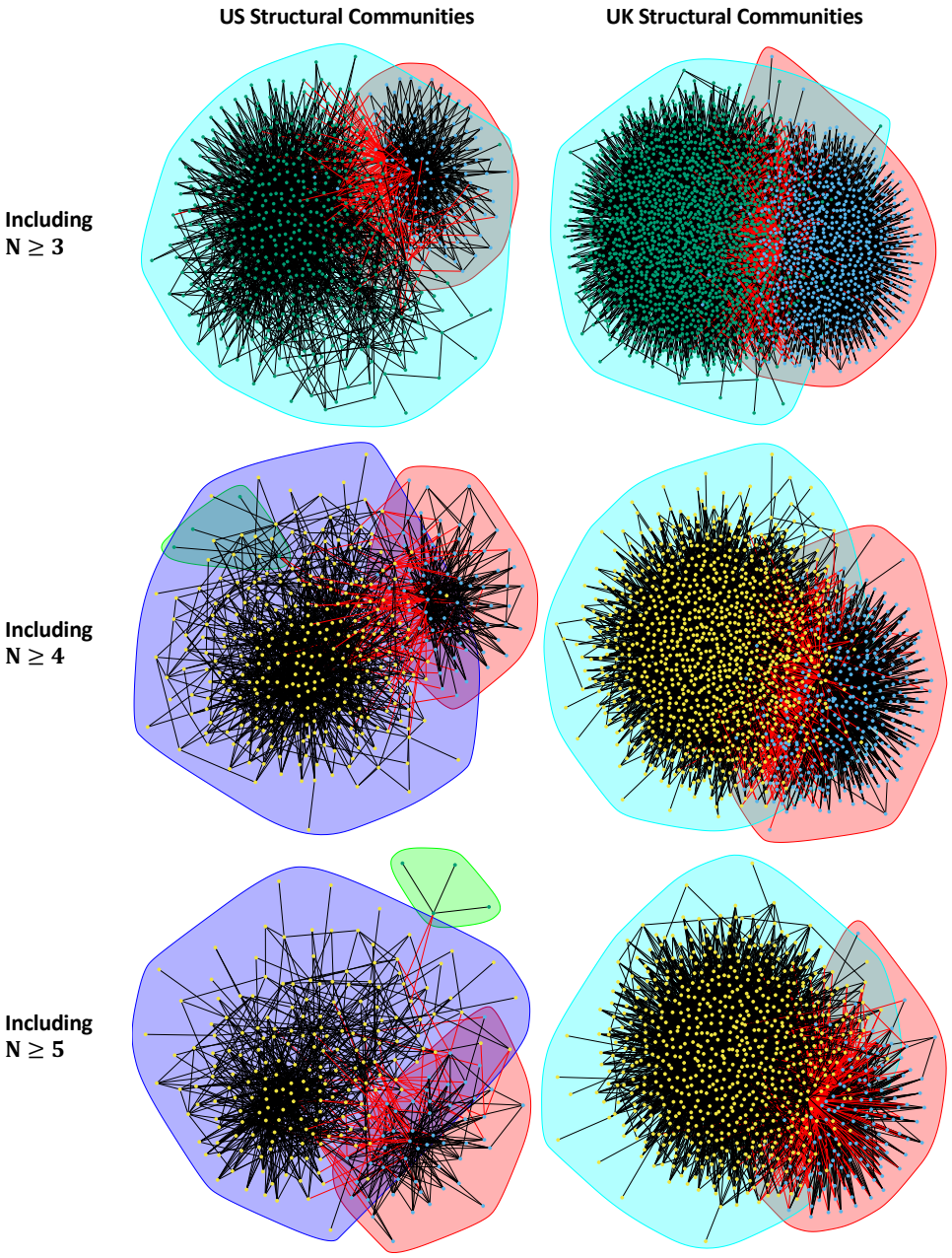
Table S8. Network size statistics under different influencer inclusion criteria

	US			**UK**		
Number of	Influencer	Participant	Edge	Influencer	Participant	Edge
N >= 5	150	99	1067	709	112	5417
N >= 4	279	105	1583	1247	116	7569
N >= 3	614	109	2588	2409	118	11055
N >= 2	2414	117	6188	6465	120	19167
N >= 1	46502	124	50276	64458	123	77160

Node: "N >= x" denotes influencer inclusion criteria where influencers followed by less than x participants are excluded. "N >= 1" is the same as not filtering; "N >= 3" is the inclusion criteria chosen in the main analysis.

While certain topological details are lost during filtering of influencers, the structural communities remain consistent across different inclusion criteria. **Figure S1** demonstrates how different inclusion criteria impact the outcomes of community detection analysis in the networks, where stricter inclusion criteria reduce the size of networks visibly for both the US and the UK. Stricter criteria also disintegrate the larger cluster in the US ("Cluster A" in the main text) but not in the UK. However, we can see that the overall structural communities remain consistent under the same algorithmic graph partitioning, where only a few nodes in the US are partitioned into their own structural community. Therefore, structural communities remain relatively consistent under different inclusion criteria.

Figure S1. Changes of network visualizations under different influencer inclusion criteria



Note. "Including N ≥ x" denotes the influencer inclusion criteria, where influencers followed by less than x participants are filtered out. "Including N ≥ 3" is the inclusion criteria used in the main analysis. Consistent with the main analysis, network graphs in this figure are created using the large-graph-layout algorithm in the igraph package in R; different graphs have slightly different layout orientations because they are created from different datasets under different

inclusion criteria. Consistent with the main analysis, structural communities are identified using the same label-propagation graph partitioning algorithm. Lower inclusion criteria of $x = 1$ or 2 are not visualized due to limited computational power.

Table S9. Influencers associated with low and high vaccine confidence in the US and the UK (using a threshold of influencers followed by 5+ participants).

Twitter Handle	Vaccine Confidence	% Getting Vaccinated	Twitter Handle	Vaccine Confidence	% Getting Vaccinated
Influencers Associated with Low Vaccine Confidence					
PrisonPlanet	1.83 (1.13)	0.00 (0.00)	britishlibrary	6.83 (0.26)	100.00 (0.00)
Timcast	2.43 (1.99)	28.57 (48.80)	Metro_Ents	6.80 (0.27)	100.00 (0.00)
KatTimpf	2.50 (1.90)	16.67 (40.82)	ThePoke	6.80 (0.27)	100.00 (0.00)
laurenboebert	2.58 (1.77)	16.67 (40.82)	mackies_crisps	6.80 (0.27)	100.00 (0.00)
RudyGiuliani	2.70 (1.04)	0.00 (0.00)	fact_covid	6.70 (0.45)	100.00 (0.00)
EricTrump	2.78 (1.64)	11.11 (33.33)	OrendaBooks	6.70 (0.27)	100.00 (0.00)
parscale	2.80 (1.89)	20.00 (44.72)	educationgovuk	6.70 (0.67)	100.00 (0.00)
cvpayne	2.83 (1.86)	16.67 (40.82)	Tryanuary	6.70 (0.27)	100.00 (0.00)
RepMattGaetz	2.86 (1.57)	14.29 (37.80)	FitbitUK	6.67 (0.41)	100.00 (0.00)
RubinReport	2.90 (2.07)	20.00 (44.72)	JimHarris	6.60 (0.55)	100.00 (0.00)
ChanelRion	2.90 (2.07)	20.00 (44.72)	SouthernRailUK	6.60 (0.65)	100.00 (0.00)
TulsiGabbard	2.90 (2.46)	20.00 (44.72)	HeathrowAirport	6.60 (0.42)	100.00 (0.00)
MariaBartiromo	2.92 (2.04)	16.67 (40.82)	MensHealthUK	6.60 (0.42)	100.00 (0.00)
catturd2	2.93 (1.72)	14.29 (37.80)	toyotires_uk	6.60 (0.22)	100.00 (0.00)
kimguilfoyle	3.10 (1.64)	20.00 (44.72)	DesignMuseum	6.60 (0.89)	100.00 (0.00)
Influencers Associated with High Vaccine Confidence					
CoryBooker	6.90 (0.22)	100.00 (0.00)	TomCruise	3.40 (1.34)	60.00 (54.77)
DrBiden	6.79 (0.39)	100.00 (0.00)	CapitalOfficial	3.50 (1.26)	16.67 (40.82)
ezraklein	6.75 (0.42)	100.00 (0.00)	JLSOfficial	3.80 (1.15)	60.00 (54.77)
AnnaKendrick47	6.71 (0.39)	100.00 (0.00)	leonalewis	3.80 (1.79)	40.00 (54.77)
NotAltWorld	6.70 (0.45)	100.00 (0.00)	nicolerichie	3.80 (1.79)	60.00 (54.77)
Nate_Cohn	6.70 (0.45)	100.00 (0.00)	BrunoMars	3.83 (1.44)	55.56 (52.70)
ChrisEvans	6.70 (0.45)	100.00 (0.00)	PerezHilton	3.83 (1.60)	33.33 (51.64)

neilhimself	6.67 (0.41)	83.33 (40.82)	MrsSOsbourne	3.83 (1.60)	50.00 (54.77)
politico	6.67 (0.41)	100.00 (0.00)	BiffyClyro	3.90 (1.14)	60.00 (54.77)
VP	6.62 (0.48)	100.00 (0.00)	JamesGShore	3.90 (1.29)	40.00 (54.77)
NPR	6.60 (0.42)	100.00 (0.00)	HollyGShore	3.90 (1.29)	40.00 (54.77)
TheEconomist	6.60 (0.55)	100.00 (0.00)	Usher	3.90 (1.43)	40.00 (54.77)
LilNasX	6.60 (0.55)	80.00 (44.72)	mishacollins	3.90 (1.85)	60.00 (54.77)
Lin_Manuel	6.60 (0.42)	100.00 (0.00)	ddlovato	3.94 (1.12)	37.50 (51.75)
FLOTUS44	6.58 (0.49)	100.00 (0.00)	freemoneylotto	4.00 (1.17)	0.00 (0.00)

Note. In the above table, we replicated the analysis presented in Table 1 of the paper, showing the average vaccine confidence of influencers associated with low and high vaccine hesitancy, but this time we show influencers followed by at least 5 people.

Table S10. Influencers associated with low and high vaccine confidence in the US and the UK (using a threshold of influencers followed by 25+ participants).

United States			United Kingdom		
Twitter Handle	**Vaccine Confidence**	**% Getting Vaccinated**	**Twitter Handle**	**Vaccine Confidence**	**% Getting Vaccinated**
Influencers Associated with Low Vaccine Confidence					
elonmusk	4.41 (2.07)	55.17 (50.61)	RealDMitchell	6.04 (0.89)	92.86 (26.23)
BarackObama	5.96 (1.37)	85.71 (35.63)	daraobriain	6.02 (0.88)	100.00 (0.00)
			SoVeryBritish	5.91 (0.93)	89.29 (31.50)
			stephenfry	5.90 (1.11)	95.83 (20.19)
			BBCNews	5.82 (1.05)	85.71 (35.63)
			JoeBiden	5.75 (1.30)	88.46 (32.58)
			jackwhitehall	5.73 (0.98)	91.89 (27.67)
			MartinSLewis	5.72 (1.08)	87.50 (33.42)
			GaryLineker	5.71 (0.92)	79.41 (41.04)
			BBCBreaking	5.69 (1.15)	81.48 (39.21)
			AldiUK	5.67 (1.11)	80.77 (40.19)
			jeremycorbyn	5.67 (1.21)	83.33 (37.90)
			rickygervais	5.64 (1.18)	84.62 (36.55)
			piersmorgan	5.59 (1.23)	85.71 (35.63)
			Twitter	5.58 (1.39)	90.00 (30.51)
Influencers Associated with High Vaccine Confidence					
BarackObama	5.96 (1.37)	85.71 (35.63)	hollywills	5.22 (1.51)	80.00 (40.51)
elonmusk	4.41 (2.07)	55.17 (50.61)	davidwalliams	5.30 (1.25)	74.07 (44.66)
			Fearnecotton	5.33 (1.33)	88.46 (32.58)
			elonmusk	5.40 (1.48)	85.29 (35.95)
			antanddec	5.41 (1.26)	78.95 (41.32)
			AmazonUK	5.48 (1.36)	75.00 (44.10)
			jk_rowling	5.48 (1.62)	82.14 (39.00)
			JKCorden	5.48 (1.09)	84.85 (36.41)
			Lord_Sugar	5.50 (1.33)	85.71 (35.63)
			jimmycarr	5.50 (1.24)	92.31 (27.17)
			JeremyClarkson	5.50 (0.93)	90.32 (30.05)
			BorisJohnson	5.52 (1.37)	85.71 (35.42)
			BarackObama	5.53 (1.36)	82.61 (38.32)
			VancityReynolds	5.53 (1.49)	76.67 (43.02)
			SkyNews	5.53 (1.31)	83.33 (37.90)

Note: In the above table, we replicated the analysis presented in Table 1 of the paper, showing the average vaccine confidence of influencers associated with low and high vaccine hesitancy,

but this time we show influencers followed by at least 25 people. Note that in the US, only two influencers (Barack Obama and Elon Musk) are followed by more than 25 people.

Table S11. Study 2 Regression Models

	Hyperpartisan	Hyperpartisan 2	NewsGuard Shares	NewsGuard Shares 2	NewsGuard Favorites	NewsGuard Favorites 2
(Intercept)	0.00 [-0.05, 0.05]	0.09 [-0.11, 0.28]	0.00 [-0.06, 0.06]	-0.43 ** [-0.69, -0.16]	0.00 [-0.06, 0.06]	-0.33 ** [-0.58, -0.08]
vaccineLikely	-0.07 ** [-0.12, -0.02]	-0.10 * [-0.18, -0.02]	0.23 *** [0.17, 0.29]	0.19 *** [0.09, 0.29]	0.23 *** [0.17, 0.28]	0.11 * [0.02, 0.21]
Liberalism		0.03 [-0.06, 0.12]		0.08 [-0.03, 0.18]		0.24 *** [0.13, 0.35]
Polarization		0.08 [-0.00, 0.16]		-0.03 [-0.13, 0.07]		-0.05 [-0.15, 0.05]
Conspiracy		0.01 [-0.07, 0.09]		-0.05 [-0.14, 0.04]		-0.12 ** [-0.21, -0.03]
mentalHealth		-0.07 [-0.17, 0.03]		0.00 [-0.12, 0.12]		-0.00 [-0.12, 0.12]
lifeSatisfaction		-0.02 [-0.11, 0.08]		0.04 [-0.08, 0.16]		-0.07 [-0.18, 0.05]
Male		-0.03 [-0.18, 0.12]		0.03 [-0.15, 0.21]		0.06 [-0.12, 0.24]
Age		0.21 *** [0.14, 0.29]		0.04 [-0.05, 0.13]		0.08 [-0.02, 0.17]
Bachelors		-0.09 [-0.28, 0.10]		0.48 *** [0.22, 0.74]		0.35 ** [0.11, 0.59]
followers_count		-0.02 [-0.10, 0.06]		0.06 [-0.05, 0.18]		0.03 [-0.08, 0.14]
friends_count		0.11 ** [0.03, 0.19]		0.05 [-0.06, 0.17]		-0.05 [-0.16, 0.06]
N	1600	734	1036	475	1064	480
R2	0.00	0.08	0.05	0.13	0.05	0.16

Note: Study 2 regression models, showing (from left-right) how one's likelihood of vaccine predicts the sharing of Hyperpartisan news (with and without controls), how one's likelihood of getting the vaccine predicts the quality of news URLS shared as rated by NewsGuard (with and without controls), and how one's likelihood of getting the vaccine predicts the quality of news URLs favorited as rated by NewsGuard (with and without controls).

Table S12. Study 2 Regression Models With Different Control Variables

	Hyperpartisan	Hyperpartisan 2	NewsGuard Shares	NewsGuard Shares 2	NewsGuard Favorites	NewsGuard Favorites 2
(Intercept)	0.00	0.01	0.00	-0.05	0.00	-0.04
	[-0.05, 0.05]	[-0.07, 0.10]	[-0.06, 0.06]	[-0.16, 0.06]	[-0.06, 0.06]	[-0.15, 0.07]
vaccine Likely	-0.07 **	-0.14 ***	0.23 ***	0.18 ***	0.23 ***	0.17 ***
	[-0.12, -0.02]	[-0.20, -0.07]	[0.17, 0.29]	[0.10, 0.26]	[0.17, 0.28]	[0.09, 0.24]
Liberalism		0.04		0.16 ***		0.20 ***
		[-0.02, 0.10]		[0.08, 0.24]		[0.12, 0.27]
Male		-0.02		0.09		0.06
		[-0.14, 0.09]		[-0.06, 0.24]		[-0.08, 0.21]
Age		0.20 ***		0.07 *		0.07
		[0.14, 0.26]		[0.00, 0.15]		[-0.00, 0.14]
N	1600	1093	1036	690	1064	705
R2	0.00	0.05	0.05	0.08	0.05	0.10

Note. Study 2 regression models, showing (from left-right) how one's likelihood of vaccine predicts the sharing of hyperpartisan news (with and without controls), how one's likelihood of getting the vaccine predicts the quality of news URLS shared as rated by NewsGuard (with and without controls), and how one's likelihood of getting the vaccine predicts the quality of news URLs favorited as rated by NewsGuard (with and without controls).

Table S13. Specific URLs shared or favorited associated with low self-reported likelihood of receiving the vaccine (using threshold of 25+ shares or favorites)

News Website Shared	Likelihood of Getting Vaccine	News Website Favorited	Likelihood of Getting Vaccine
espn.com	88.00 (30.37)	foxnews.com	88.21 (30.62)
dailymail.co.uk	88.84 (27.76)	metro.co.uk	88.94 (27.43)
metro.co.uk	89.80 (25.55)	dailymail.co.uk	89.81 (27.00)
mashable.com	90.82 (23.72)	nypost.com	90.09 (25.55)
boingboing.net	90.87 (25.05)	variety.com	90.91 (25.74)
msn.com	92.20 (27.01)	chicago.suntimes.com	91.85 (19.92)
nypost.com	92.35 (23.98)	standard.co.uk	91.93 (23.63)
on.wsj.com	92.47 (22.64)	a.msn.com	92.06 (24.02)
huffingtonpost.co.uk	92.69 (20.90)	chicagotribune.com	92.14 (23.61)
on.ft.com	92.72 (22.45)	truthout.org	92.19 (27.15)
salon.com	93.27 (20.72)	thesun.co.uk	92.31 (21.87)
foxnews.com	93.41 (21.90)	deadline.com	92.42 (24.17)
nationalreview.com	93.57 (23.76)	mol.im	92.50 (23.84)
ajc.com	93.60 (22.89)	yahoo.com	92.67 (23.34)
dailykos.com	93.62 (22.43)	people.com	92.67 (23.80)

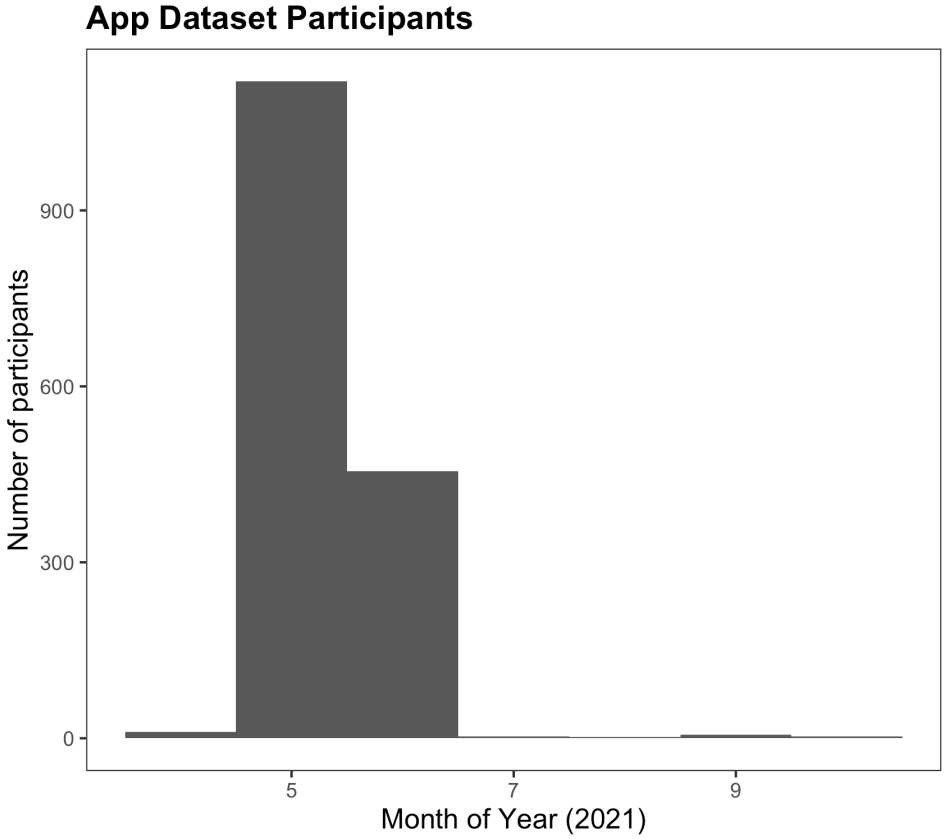
Note. In the above table, we replicate the analysis shown in Table 2, but show URLs that are shared or favorited by at least 25 participants.

Table S14. Specific URLs shared or favorited associated with low self-reported likelihood of receiving the vaccine (using threshold of 5+ shares or favorites)

News Website Shared	Likelihood of Getting Vaccine	News Website Favorited	Likelihood of Getting Vaccine
redstate.com	58.00 (53.10)	thegatewaypundit.com	45.44 (45.42)
6abc.com	61.40 (52.91)	chroniclelive.co.uk	54.60 (42.61)
breitbart.com	65.78 (43.88)	cancer.gov	60.00 (54.77)
thepostmillennial.com	66.00 (34.89)	newsmax.com	66.71 (46.03)
billboard.com	69.08 (42.21)	billboard.com	70.07 (41.76)
zerohedge.com	72.62 (42.52)	deviantart.com	72.30 (41.00)
dailycaller.com	72.86 (46.45)	westernjournal.com	72.40 (43.69)
dailywire.com	72.86 (43.48)	zerohedge.com	73.94 (39.67)
thewrap.com	73.00 (43.53)	thepostmillennial.com	75.23 (38.98)
fox32chicago.com	73.40 (39.88)	khou.com	77.00 (40.67)
foxbusiness.com	73.50 (41.05)	realclearpolitics.com	77.00 (40.67)
crooksandliars.com	74.40 (35.17)	thewrap.com	77.73 (38.93)
washingtonexaminer.com	75.92 (35.16)	breitbart.com	78.07 (37.21)
theepochtimes.com	76.67 (40.82)	tvline.com	78.33 (43.01)
treehugger.com	76.67 (31.80)	fox26houston.com	79.20 (44.31)

Note. In the above table, we replicate the analysis shown in Table 2, but show URLs that are shared or favorited by at least 5 participants.

Figure S2. Recruitment of App Participants.



Note. Histogram of when app participants were recruited. Most participants were recruited when the app was first sent out on Twitter in May, 2021. However, a second wave of participants also used the app when Study 1 participants were recommended to use it in June, 2021.

12. Supplementary Materials for "How Social Media (Unfollowing) Behavior Influences Affective Polarization and Well-Being: Results from a Social Media Field Experiment"

S1: Accounts Associated with Low and High Favorability Toward Democrats and Republicans

Accounts Associated With The Lowest Favorability Toward Democrats					Accounts Associated With The Lowest Favorability Toward Republicans				
Twitter Accounts	Number of Followers in Sample	Favorability Toward Republicans	Favorability Toward Democrats	Followers	Twitter Accounts	Number of Followers in Sample	Favorability Toward Republicans	Favorability Toward Democrats	Followers
ThomasSowell	27	42.56	21.33	1030313	kashanacauley	26	4.42	81.31	123656
TheBabylonBee	35	42.43	35.06	1556775	PhilosophyTube	29	4.48	56.59	314382
MrAndyNgo	31	36.90	35.81	1028024	caslernoel	38	4.92	82.55	391237
DonaldJTrumpJr	33	43.79	37.91	8488730	biologistimo	27	5.11	77.19	71731
DanCrenshawTX	27	43.63	39.41	1249378	NULL	27	5.41	59.78	0
RealCandaceO	28	44.71	39.89	3154969	KatiePhang	31	5.48	84.06	226606
jordanbpeterson	55	36.91	41.98	2879088	corvidresearch	36	5.50	72.97	71942
benshapiro	53	37.98	42.38	4535667	LincolnsBible	30	5.57	78.10	200848
TuckerCarlson	33	39.24	44.45	5257576	BrandyLJensen	31	5.58	67.10	152154
TitaniaMcGrath	38	32.18	45.82	722084	SenJeffMerkley	38	5.63	84.26	512709
megynkelly	28	33.14	46.29	2612372	DrGJackBrown	26	5.73	85.50	166996
RandPaul	35	38.43	46.29	4024448	andylassner	26	5.73	85.77	476988
tedcruz	33	36.12	46.58	5279294	SpiroAgnewGhost	32	5.75	82.63	203257
naval	39	27.85	46.64	1881997	joshscampbell	39	5.77	85.05	242562
NRO	28	27.82	46.68	338973	pareene	27	5.78	74.33	122061
joerogan	56	28.36	47.04	9346143	jpbrammer	26	5.81	77.96	173728
DouglasKMurray	27	32.59	47.04	503537	AmyMcGrathKY	68	5.88	85.74	555953
BretWeinstein	26	27.88	47.38	734395	CecileRichards	31	5.90	81.97	200467
seanhannity	28	40.75	47.57	5913330	claudiamconwayy	52	5.90	78.27	530498
TulsiGabbard	28	29.86	48.18	1592666	ProBirdRights	39	5.92	71.56	405273
EricRWeinstein	28	34.93	48.68	691094	TomJChicago	29	5.93	83.69	178210
lhfang	26	22.62	49.42	184190	mcmansionhell	34	6.06	60.59	92396
normmacdonald	30	21.63	49.73	1149518	gregolear	35	6.11	80.91	197890
IvankaTrump	30	33.03	49.93	10271520	eliehonig	34	6.24	82.47	300711
KellyannePolls	29	37.83	51.10	3458283	JordanUhl	26	6.27	61.00	276992
dailystoic	26	27.62	51.38	366328	NARAL	42	6.29	79.93	227246
marcorubio	29	39.03	53.45	4417266	grantstern	27	6.30	88.67	187261
sentientist	31	21.52	53.77	58860	Hegemommy	31	6.32	74.71	77026
kanyewest	65	26.12	54.00	30838775	richardmarx	37	6.38	80.16	348825
BenSasse	28	19.43	55.04	289742	AkilahObviously	28	6.39	75.96	240101
balajis	34	24.68	55.29	679492	brikeilarcnn	31	6.45	82.84	351598
GravelInstitute	37	6.59	55.46	387804	TheDweck	37	6.46	78.03	245414
reason	26	26.19	55.46	277097	DanaSchwartzzz	26	6.50	76.65	202721
Quillette	33	22.12	55.48	216848	goldengateblond	71	6.51	80.61	366503
EconTalker	26	19.23	55.73	73903	briantylercohen	36	6.53	82.33	570605
Ayaan	32	24.63	55.84	451312	girlsreallyrule	51	6.55	81.96	349967
clairlemon	27	24.67	55.96	237163	RespectableLaw	36	6.58	69.78	174934
GlennLoury	33	22.70	56.21	169429	GravelInstitute	37	6.59	55.46	387804
mtaibbi	46	17.33	56.22	664047	NULL	38	6.63	80.00	0
nfergus	26	26.85	56.38	265975	JBeachFerrara	41	6.71	82.85	151008
ggreenwald	69	21.01	56.45	1864288	JamaalBowmanNY	52	6.71	69.85	357776
thomaschattwill	31	22.90	56.45	126611	KillerMartinis	30	6.73	72.23	90493
PhilosophyTube	29	4.48	56.59	314389	SecDebHaaland	27	6.81	83.26	157906
DemSocialists	39	6.82	56.62	380282	DemSocialists	39	6.82	56.62	380281
slatestarcodex	30	20.97	56.73	118291	electrolemon	30	6.83	68.03	232570
sullydish	40	22.10	56.88	276926	PhilippeReines	30	6.83	83.57	115337
JeremyClarkson	60	27.63	57.18	7864664	Afro_Herper	33	6.85	77.52	53351
BorisJohnson	50	31.34	57.20	4626071	MaxKennerly	27	6.85	82.37	85860
GOP	28	30.14	57.21	3002715	sfpelosi	30	6.87	85.60	191179
PsychRabble	30	28.30	57.23	21176	TomPerez	30	6.87	77.83	227725
MrBeast	40	20.45	57.35	15125211	TrueFactsStated	26	6.88	82.19	228845
MrJamesMay	41	20.78	57.56	3223698	EricIdle	39	6.90	77.79	506979
DudespostingWs	33	26.09	57.70	1816098	_SJPeace_	26	6.92	76.96	311778
billburr	36	21.47	57.72	2015092	Booker4KY	45	6.98	69.56	445845
animalcrossing	28	11.04	57.75	1488895	itsJeffTiedrich	77	7.00	82.17	956962
colesprouse	26	24.62	57.92	9530632	TopherSpiro	33	7.03	80.76	104195
SkySportsNews	29	23.41	58.14	11341498	RacismDog	71	7.04	75.04	638049
PlayStation	33	20.94	58.24	26605943	JortsTheCat	39	7.10	75.56	215788
GordonRamsay	41	24.15	58.59	7679123	maziehirono	35	7.11	83.97	520983
pulte	53	25.13	58.74	3284180	joelockhart	32	7.13	85.34	194906
TigerWoods	30	23.27	58.83	6673646	beatonna	28	7.18	73.64	159395
IGN	29	23.48	58.86	9196435	Eleven_Films	39	7.21	81.54	174816
wikileaks	82	20.07	58.95	5615102	ChrisWarcraft	29	7.21	74.17	153195
CatoInstitute	28	24.25	58.96	365740	grimalkina	28	7.21	76.64	15816
TheFIREorg	26	16.00	59.04	70991	Sarahchadwickk	28	7.21	79.61	263198
piersmorgan	44	36.73	59.07	7943177	michpoligal	27	7.22	83.19	135023
romanyam	28	20.36	59.21	113108	BrunoAmato_1	36	7.22	87.58	238232
ESPNStatsInfo	35	19.26	59.43	1774753	Rschooley	27	7.22	80.52	115093
FoxNews	47	30.89	59.45	22269966	feministabulous	26	7.23	80.42	172586
jacobin	38	11.29	59.45	374179	pixelatedboat	57	7.25	64.42	330224

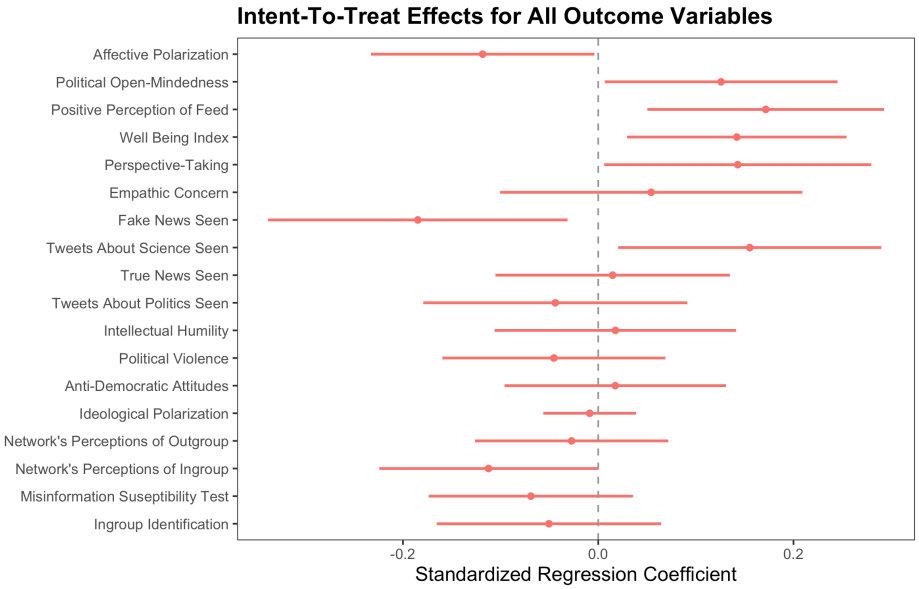
MartinSLewis	36	29.03	59.56	1683204	NaomiBiden	32	7.25	85.97	322044
Mike_Pence	30	27.57	59.63	5863931	galendruke	27	7.26	78.00	52195
NULL	27	5.41	59.78	0	SICKOFWOLVES	45	7.29	70.73	170712
ContraPoints	59	7.71	59.81	581560	B52Malmet	27	7.30	78.63	300837
WilliamShatner	42	28.31	59.86	2570498	JuliaDavisNews	37	7.30	81.51	306664
bariweiss	48	22.27	59.90	476785	AoDespair	74	7.30	78.23	336200
Xbox	42	24.64	59.93	19324988	duty2warn	37	7.30	81.05	316383
LindseyGrahamSC	29	23.00	60.17	2134619	YoYo_Ma	26	7.31	76.23	228733
RapSheet	36	21.47	60.25	3543826	seditiontrack	29	7.31	82.66	71036
PressSec45	56	25.88	60.30	5842091	NickKnudsenUS	29	7.31	89.38	235377
kendricklamar	27	17.67	60.48	12139746	TheRaDR	46	7.33	78.15	162308
mcmansionhell	34	6.06	60.59	92395	AmyEGardner	32	7.34	82.22	125056
danieltosh	44	22.09	60.61	24164355	EdMarkey	40	7.35	70.45	240095
TheJusticeDept	36	17.14	60.81	2064675	giselefetterman	26	7.38	84.31	144130
KDTrey5	35	23.29	60.83	20378821	davejorgenson	35	7.40	80.60	105777
tferriss	34	22.88	60.91	1845338	GeoffRBennett	34	7.41	83.09	187699
Reductress	32	10.84	60.94	301809	SaraGideon	63	7.44	83.92	310449
JordanUhl	26	6.27	61.00	276995	JustinMcElroy	28	7.46	76.29	356829
JonahDispatch	41	20.44	61.02	356219	PossumEveryHour	28	7.50	66.96	576761
PhDForum	31	15.77	61.42	116720	DrDenaGrayson	46	7.50	76.52	317109
DegenRolf	33	18.64	61.64	53445	Hannahgadsby	26	7.50	77.77	266070
RichardHammond	33	29.09	61.64	3080778	_FloridaMan	57	7.51	71.56	358167
notthefakeSVP	29	21.76	61.66	2077602	MaryRobinette	26	7.54	79.88	73992
VP45	65	27.63	61.78	9908131	SparkNotes	32	7.56	74.44	378361
nntaleb	67	23.66	61.84	885895	capitalweather	28	7.57	76.50	1113623
MLB	48	19.60	61.88	10074476	mikamckinnon	29	7.59	71.00	69588
BrandsOwned	26	13.54	61.92	290177	anne_theriault	35	7.60	80.74	91815
SarahTheHaider	26	20.12	61.92	115824	socialistdogmom	35	7.60	73.03	132875
ATabarrok	32	19.22	61.94	66215	PPact	49	7.61	76.92	505551
YouTube	59	21.71	61.95	76077719	NASAClimate	34	7.62	74.38	355563

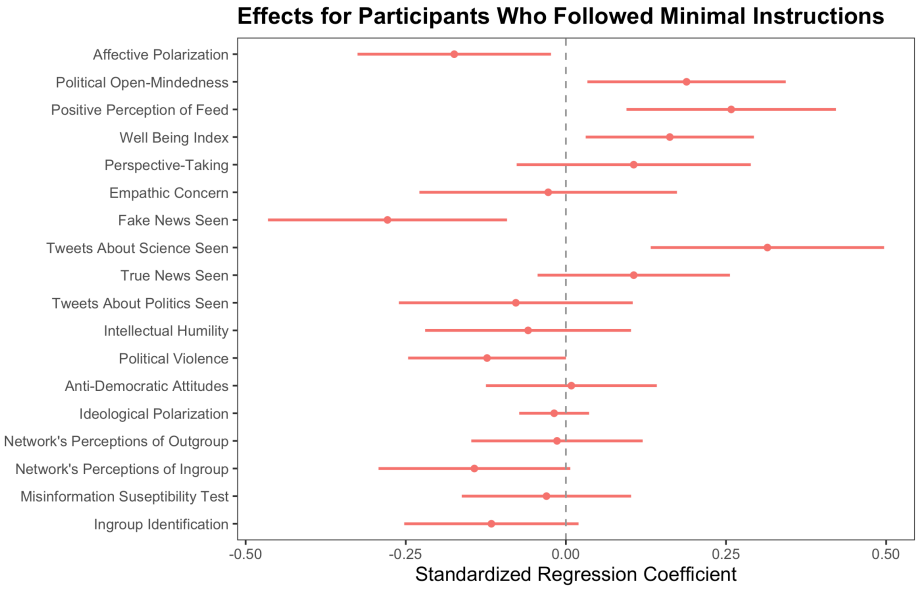
Note. Shown above are a subset of Twitter "influencers" that had at least 25 followers in our sample, ordered by lowest to highest favorability toward Democrats (on a scale of 1-100), and lowest to highest favorability toward Republicans, ordered by lowest to highest favorability toward Republicans (on a scale of 1-100). Lists such as these were used to help select the 60 accounts that we asked participants to unfollow.

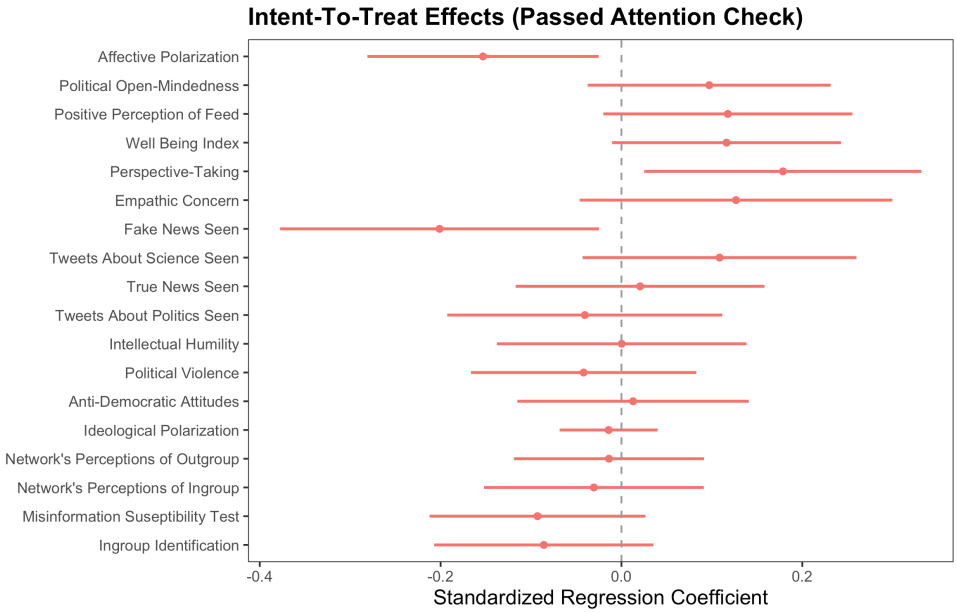
S2: Multiple Regression Table – Predictors of Twitter Toxicity

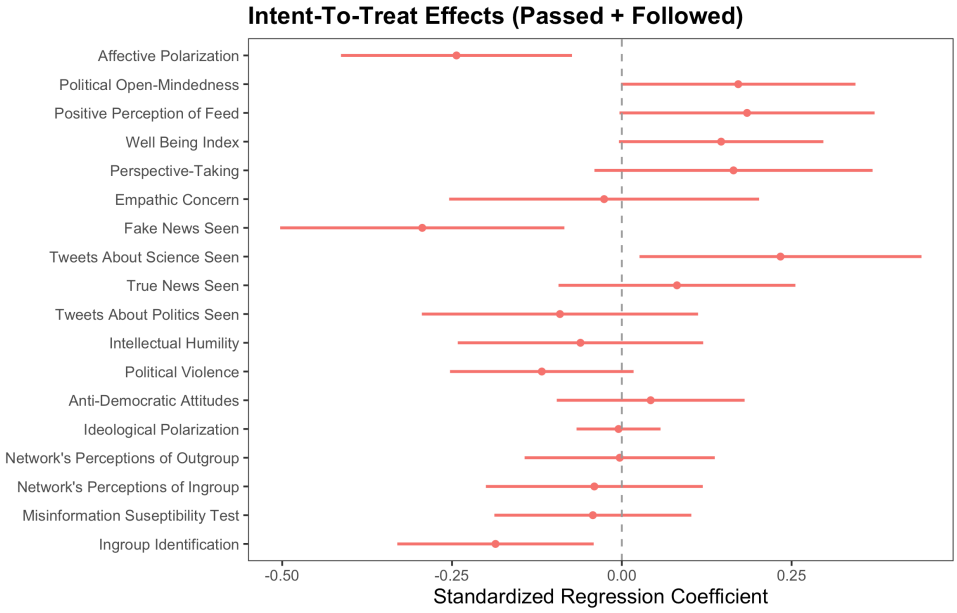
	estimate
(Intercept)	0.37 ***
	[0.15, 0.58]
PartisanAccounts	0.09 *
	[0.00, 0.17]
dislikeOfDemocrats	0.15 **
	[0.04, 0.25]
dislikeOfRepublicans	0.14 *
	[0.03, 0.25]
conspiracy	0.04
	[-0.05, 0.13]
vaccineLikely.x	-0.04
	[-0.14, 0.06]
mentalHealth.x	0.05
	[-0.07, 0.17]
lifeSatisfaction.x	-0.09
	[-0.20, 0.02]
Politics.x	-0.06
	[-0.18, 0.06]
GenderRecode	-0.08
	[-0.25, 0.10]
Age.x	0.01
	[-0.08, 0.10]
Bachelors	-0.41 ***
	[-0.64, -0.18]
N	527
R2	0.08

S3: Results for All Outcome Variables (Across Multiple Specifications)






Intent-To-Treat Effects (Passed Attention Check)
Affective Polarization
Political Open-Mindedness
Positive Perception of Feed
Well Being Index
Perspective-Taking
Empathic Concern
Fake News Seen
Tweets About Science Seen
True News Seen
Tweets About Politics Seen
Intellectual Humility
Political Violence
Anti-Democratic Attitudes
Ideological Polarization
Network's Perceptions of Outgroup
Network's Perceptions of Ingroup
Misinformation Suseptibility Test
Ingroup Identification
-0.4
-0.2
0.0
0.2
Standardized Regression Coefficient


Intent-To-Treat Effects (Passed + Followed)
Affective Polarization
Political Open-Mindedness
Positive Perception of Feed
Well Being Index
Perspective-Taking
Empathic Concern
Fake News Seen
Tweets About Science Seen
True News Seen
Tweets About Politics Seen
Intellectual Humility
Political Violence
Anti-Democratic Attitudes
Ideological Polarization
Network's Perceptions of Outgroup
Network's Perceptions of Ingroup
Misinformation Suseptibility Test
Ingroup Identification
-0.50
-0.25
0.00
0.25
Standardized Regression Coefficient

S4: Regression Tables for All Outcome Variables

Below are simple regressions for all outcome variables with four different exclusion criteria, the full sample with valid Twitter handles, the full sample that followed minimal instructions, the full sample that passed the attention check, and the full sample that passed the attention check and followed minimal instructions.

Intent-To-Treat Full Sample

term	estimate	std.error	statistic	p.value	conf.low	conf.high
Affective Polarization	-0.118	0.058	-2.031	0.043	-0.233	-0.004
Political Open-Mindedness	0.126	0.061	2.073	0.039	0.007	0.245
Positive Perception of Feed	0.172	0.062	2.778	0.006	0.050	0.293
Well Being Index	0.142	0.057	2.479	0.013	0.029	0.254
Perspective-Taking	0.143	0.070	2.051	0.041	0.006	0.280
Empathic Concern	0.054	0.079	0.688	0.491	-0.101	0.209
Fake News Seen	-0.185	0.078	-2.366	0.018	-0.338	-0.031
Tweets About Science Seen	0.155	0.069	2.261	0.024	0.020	0.290
True News Seen	0.015	0.061	0.242	0.809	-0.105	0.135
Tweets About Politics Seen	-0.044	0.069	-0.638	0.524	-0.179	0.091
Intellectual Humility	0.018	0.063	0.279	0.780	-0.106	0.141
Political Violence	-0.045	0.058	-0.780	0.436	-0.160	0.069
Anti-Democratic Attitudes	0.017	0.058	0.302	0.763	-0.096	0.131
Ideological Polarization	-0.009	0.024	-0.360	0.719	-0.056	0.039
Network's Perceptions of Outgroup	-0.027	0.050	-0.541	0.589	-0.126	0.072
Network's Perceptions of Ingroup	-0.112	0.057	-1.972	0.049	-0.224	0.000
Misinformation Suseptibility Test	-0.069	0.053	-1.294	0.196	-0.174	0.036
Ingroup Identification	-0.050	0.059	-0.862	0.389	-0.165	0.065

Effects for Participants Who Followed Minimal Instructions

term	estimate	std.error	statistic	p.value	conf.low	conf.high
Affective Polarization	-0.174329	0.0768049	-2.26976	0.023823	-0.3253799	-0.023278
Political Open-Mindedness	0.1884169	0.0787666	2.392092	0.017274	0.0335076	0.3433261
Positive Perception of Feed	0.2581645	0.0831615	3.104375	0.00206	0.0946149	0.4217141
Well Being Index	0.1622791	0.0668066	2.42909	0.015634	0.0308915	0.2936667
Perspective-Taking	0.1059115	0.0929477	1.139473	0.255276	-0.0768877	0.2887106
Empathic Concern	-0.027618	0.1022581	-0.27008	0.787254	-0.228728	0.1734916
Fake News Seen	-0.278536	0.0948764	-2.93578	0.003544	-0.4651266	-0.091946
Tweets About Science Seen	0.3146671	0.0926663	3.395702	0.000762	0.1324232	0.496911
True News Seen	0.1059627	0.0763628	1.387623	0.166122	-0.0442176	0.256143

Tweets About Politics Seen	-0.078141	0.0928518	-0.84157	0.400598	-0.2607497	0.1044679
Intellectual Humility	-0.059021	0.0817454	-0.72201	0.470764	-0.2197888	0.1017465
Political Violence	-0.123111	0.0626393	-1.96539	0.050152	-0.2463042	8.221E-05
Anti-Democratic Attitudes	0.00853	0.0678339	0.125748	0.900003	-0.1248795	0.1419395
Ideological Polarization	-0.018343	0.0277327	-0.66142	0.508778	-0.072885	0.0361993
Network's Perceptions of Outgroup	-0.013874	0.0680427	-0.20391	0.838544	-0.147693	0.1199444
Network's Perceptions of Ingroup	-0.142891	0.0761164	-1.87727	0.061303	-0.2925886	0.0068059
Misinformation Suseptibility Test	-0.030332	0.0672071	-0.45133	0.652031	-0.1625091	0.1018442
Ingroup Identification	-0.116312	0.0691998	-1.68082	0.09368	-0.2524065	0.0197822

Intent-To-Treat Passed Attention Check

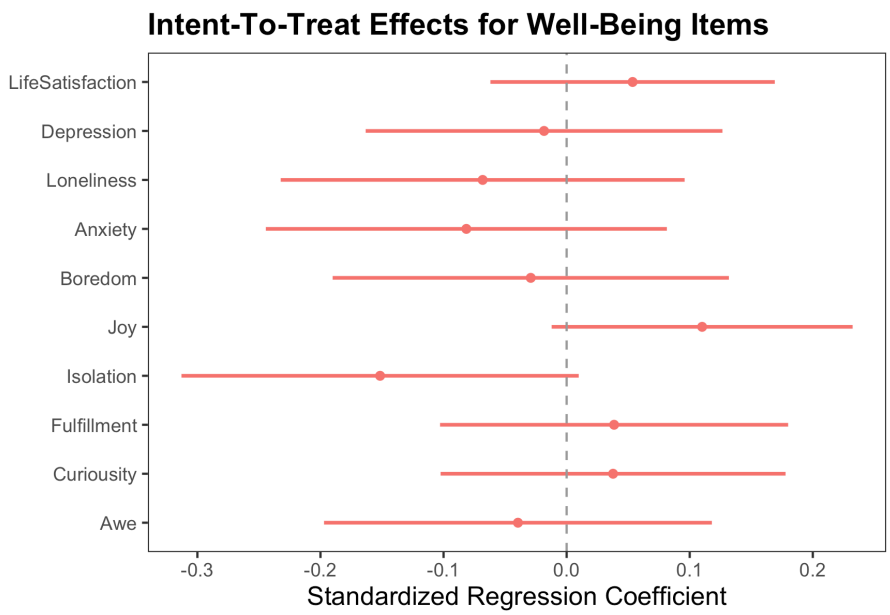
term	estimate	std.error	statistic	p.value	conf.low	conf.high
Affective Polarization	-0.153095	0.0651064	-2.35146	0.019097	-0.2810184	-0.025171
Political Open-Mindedness	0.0970431	0.0684115	1.418522	0.156677	-0.0373742	0.2314605
Positive Perception of Feed	0.1176751	0.0701294	1.677971	0.093989	-0.0201156	0.2554658
Well Being Index	0.116284	0.0644325	1.80474	0.071732	-0.0103155	0.2428834
Perspective-Taking	0.1785294	0.0780305	2.287943	0.022569	0.0252114	0.3318474
Empathic Concern	0.1266902	0.0880167	1.439388	0.150683	-0.0462492	0.2996295
Fake News Seen	-0.2012	0.0897162	-2.24263	0.025368	-0.3774771	-0.024924
Tweets About Science Seen	0.1084538	0.0770917	1.406814	0.160118	-0.0430182	0.2599257
True News Seen	0.0206568	0.0700371	0.294941	0.768164	-0.116954	0.1582676
Tweets About Politics Seen	-0.040507	0.0774257	-0.52317	0.60109	-0.1926352	0.111621
Intellectual Humility	0.0002952	0.0702691	0.004201	0.99665	-0.1377721	0.1383625
Political Violence	-0.041826	0.0634396	-0.65931	0.510007	-0.1664754	0.0828226
Anti-Democratic Attitudes	0.0128238	0.0651395	0.196867	0.844014	-0.1151653	0.1408129
Ideological Polarization	-0.014194	0.0275985	-0.51429	0.607279	-0.0684205	0.040033
Network's Perceptions of Outgroup	-0.01383	0.0535182	-0.25842	0.796195	-0.1189844	0.0913246
Network's Perceptions of Ingroup	-0.030489	0.061843	-0.49301	0.62223	-0.1520004	0.0910224
Misinformation Suseptibility Test	-0.092844	0.060737	-1.52863	0.127006	-0.2121831	0.0264946
Ingroup Identification	-0.085842	0.0617186	-1.39086	0.164902	-0.2071088	0.0354251

Intent-To-Treat Passed Attention Check & Minimally Complied

term	estimate	std.error	statistic	p.value	conf.low	conf.high
Affective Polarization	-0.243441	0.086396	-2.81773	0.005182	-0.413512	-0.07337
Political Open-Mindedness	0.1714118	0.087707	1.954369	0.051656	-0.0012396	0.3440633
Positive Perception of Feed	0.1843849	0.0954134	1.932484	0.054305	-0.0034309	0.3722007

Well Being Index	0.1462653	0.0764677	1.912771	0.056801	-0.0042617	0.2967923
Perspective-Taking	0.1644437	0.1040228	1.580843	0.115047	-0.0403255	0.369213
Empathic Concern	-0.026044	0.1159216	-0.22467	0.822398	-0.2542365	0.2021476
Fake News Seen	-0.293834	0.1063074	-2.764	0.006089	-0.5030968	-0.08457
Tweets About Science Seen	0.2336391	0.105472	2.215176	0.027553	0.0260204	0.4412578
True News Seen	0.0811943	0.0885922	0.916495	0.360196	-0.093197	0.2555856
Tweets About Politics Seen	-0.091081	0.1033167	-0.88157	0.378767	-0.2944568	0.1122954
Intellectual Humility	-0.060873	0.0917938	-0.66315	0.507784	-0.2415691	0.1198236
Political Violence	-0.117807	0.0686615	-1.71576	0.087315	-0.2529673	0.0173535
Anti-Democratic Attitudes	0.0424973	0.0702558	0.604894	0.545741	-0.0958014	0.180796
Ideological Polarization	-0.004804	0.0314119	-0.15293	0.878565	-0.0666383	0.0570307
Network's Perceptions of Outgroup	-0.003093	0.0711252	-0.04348	0.965346	-0.1431029	0.1369173
Network's Perceptions of Ingroup	-0.040453	0.0811385	-0.49856	0.618481	-0.200174	0.1192689
Misinformation Suseptibility Test	-0.042633	0.0736667	-0.57873	0.563236	-0.1876465	0.1023798
Ingroup Identification	-0.185963	0.0734579	-2.53156	0.011905	-0.3305654	-0.041361

S5: Effects for the Individual Well-Being Items



Intent-To-Treat Well-Being Items

term	estimate	std.error	statistic	p.value
LifeSatisfaction	0.054	0.059	0.909	0.364
Depression	-0.018	0.074	-0.248	0.804
Loneliness	-0.068	0.084	-0.815	0.415
Anxiety	-0.081	0.083	-0.980	0.327
Boredom	-0.029	0.082	-0.355	0.723
Joy	0.110	0.062	1.765	0.078
Isolation	-0.152	0.082	-1.841	0.066
Fulfillment	0.039	0.072	0.534	0.593
Curiousity	0.038	0.072	0.527	0.598
Awe	-0.040	0.080	-0.492	0.623

S6: Intervention Instructions and Wording for Main Outcome Variables

Introduction Text (Delivered to Both Participants in the Experimental and Control Condition).

In this study, we are interested in seeing whether certain ways of using Twitter change your experience with Twitter.

We will ask you to change minor aspects of your Twitter behavior by, for example, following and unfollowing certain Twitter accounts for one month in exchange for an Amazon gift card payment of up to $20.

Then, we will follow up with you in a survey one month later and ask you about your experience with Twitter, your emotions, and your beliefs and attitudes.

Your time participating in this experiment will help contribute to scientific research that will inform people about how to use social media in more positive ways.

[page break]

First, please share your Twitter handle (@) with us.

You must have a valid Twitter handle that is not set to “protected” to participate in this study. This must be your own Twitter handle. We will check if this is your Twitter handle by seeing if you follow/unfollow certain Twitter accounts as instructed.

Make sure you are sharing your handle as opposed to your Twitter name. For example, while Tom Hanks' Twitter name is "Tom Hanks," his Twitter handle is @tomhanks.

Your Twitter handle will be kept strictly confidential, and we will only use your handle to measure anonymized data from your Twitter account. If you are not comfortable sharing your Twitter handle or changing your Twitter behavior, you may exit the survey now.

Please enter your Twitter handle below:

[page break]

Experimental Condition:

We want to learn more about how you experience Twitter when you follow or unfollow certain accounts.

To do this, we have provided you with a list of accounts to follow. The list can be accessed with the link below. Note: when you click the below link, it will open in a new tab so that you don't exit the survey.

List: https://twitter.com/i/lists/1527700772653346817

Please **follow every account in this list** by clicking on the "members" button and then clicking “follow” on each account you are not already following. Once you have clicked on the "follow" button, it should say "following." If you are already following an account, please do not accidentally unfollow it.

Here is a video demonstrating specifically how to follow all accounts in a list: https://www.youtube.com/watch?v=oCEYyNIMMUw&feature=emb_logo

At several points throughout the four-week study, we will use the Twitter handle you provided us to check whether you are still following all accounts in this list.

In order to receive $8.5 of the $20 total payment for this study, you must be following all accounts in this list for a total of four weeks. We will use your Twitter handle that you have provided us to confirm whether you have followed all participants we have requested you to follow.

In order to receive an extra $2.5 in bonus payments, you will have to correctly identify some tweets from accounts we asked you to follow.

[page break]

We have also provided you with a list of Twitter accounts to "unfollow."

The list can be accessed with the link below. Note: when you click on the below link, it will open in a new tab so that you don't exit the survey.

List: https://twitter.com/i/lists/1444489439120482308

Please **unfollow every account in this list that you are currently following**. You must make sure you are already following an account before clicking "unfollow" so that you do not accidentally follow any accounts.

[page break]

Here is a video demonstrating how to unfollow all accounts in a list that you are currently following: https://www.youtube.com/watch?v=rZfNLtvB3PA

Once again, we will use your Twitter handle to confirm that you are not following these accounts for the next four weeks.

To receive $8.5 of the $20 payment for this study, you must not be following any of the accounts in this list for the next four weeks.

Control Condition:

We want to learn more about how you experience Twitter when you follow or unfollow certain accounts.

To do this, we have provided you with a list of accounts to follow. The list can be accessed with the link below. Note: when you click the below link, it will open in a new tab so that you don't exit the survey.

List: *[This list included 2 accounts controlled by the research team that tweeted out pictures of animals to measure compliance. These accounts were also followed by people in the control condition. We have redacted this list of accounts because they were followed by study participants and thus might reveal the participants in the intervention.]*

Please **follow every account in this list** by clicking on the "members" button and then clicking “follow” on each account you are not already following. Once you have clicked on the "follow" button, it should say "following." If you are already following an account, please do not accidentally unfollow it.

Here is a video demonstrating specifically how to follow all accounts in a list: https://www.youtube.com/watch?v=WgH9WLb_QzM

At several points throughout the four-week study, we will use the Twitter handle you provided us to check whether you are still following all accounts in this list.

In order to receive $8.5 of the $20 total payment for this study, you must be following all accounts in this list for a total of four weeks. We will use your Twitter handle that you have provided us to confirm whether you have followed all participants we have requested you to follow.

In order to receive an extra $2.5 in bonus payments, you will have to correctly identify some tweets from accounts we asked you to follow.

[page break].

We have also provided you with a list of Twitter accounts to “unfollow.”

The list can be accessed with the link below. Note: when you click on the below link, it will open in a new tab so that you don't exit the survey.

List: https://twitter.com/i/lists/1520196732897808384

Please unfollow every account in this list that you are currently following. You must make sure you are already following an account before clicking "unfollow" so that you do not accidentally follow any accounts.

Here is a video demonstrating how to unfollow all accounts in a list that you are currently following:

Once again, we will use your Twitter handle to confirm that you are not following these accounts for the next four weeks.

To receive $8.5 of the $20 payment for this study, you must not be following any of the accounts in this list for the next four weeks.

…

Algorithm Condition:

Finally, we ask you to set your Twitter timeline to the "latest tweets" setting, as opposed to the default "home" setting.

Note: we recommend that you once again have Twitter opened in a different tab so you do not exit the survey.

On both your desktop **and** your phone, please click on a button that looks like the below image in the upper right-hand corner of your news feed:

Then, if you do not have the latest tweets option already selected, select the option "see latest tweets first."

Here is a video demonstrating how to turn on the "latest tweets" setting on a desktop computer: https://www.youtube.com/watch?v=aDMRnOG4N_M

The process for changing the setting to latest tweets is very similar on the phone: you select the button in the upper-right hand corner of your news feed and then select the "latest tweets" option, as shown in the video.

You must keep the "latest tweets" feature on (for both your phone *and* your desktop) for the next four weeks of the experiment in order to receive $8.5 of your total $20 payment. We will check at the end of the experiment whether you kept this setting on, for both your desktop and phone.

Main Outcome Variables

Affective Polarization

Over the past four weeks, how warm and favorable did you feel toward Democrats?
Over the past four weeks, how warm and favorable did you feel toward Republicans?

[Questions were asked on a scale of 1-100. Affective polarization was then measured by subtracting warmth toward the opposing party from warmth toward one's own party.]

Over the past four weeks, how often has your Twitter feed been...

Educational
Informative
Intriguing
Inspiring
Positive
Uplifting
Polarizing
Divisive
Angry
Sad
Depressing
Upsetting

[Questions were asked on a 5-point scale from "Never" to "Always." A composite score of positive feelings toward one's Twitter feed was calculated by averaging the first six items and the reverse score of the final six items.]

Subjective Well-Being:

Please indicate how much you agree with the following statements:

Never, Rarely, Sometimes, Often, Always

How often did you feel the following over the last month:

Satisfaction with life,
Depression,
Loneliness,
Anxiety,
Boredom,
Joy,
Isolation,
Fulfillment,
Curiousity,
Awe