Cognitive Processes Shaping Individual and Collective Belief Systems

Madalina Oana Vlasceanu

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Departments of
Psychology and Neuroscience
Adviser: Professor Alin Coman

May 2021

Abstract

Misinformation spread is among the top threats facing the world today. In my dissertation I strive to provide a deeper understanding of the socio-cognitive factors that can impact beliefs and of the potential for translating and implementing this understanding into policies aimed at reducing misinformation at a societal level. In the first Part, I introduce a theoretical generative framework for investigating factors influencing beliefs (Chapter 1). I then provide empirical support for this framework (N=7068) at the individual (Part 2, Chapters 2-5) and collective (Part 3, Chapters 6-8) levels of investigation. Finally, in Part 4 I suggest applications and future trajectories (Chapter 9). Starting with the individual level of investigation, in a series of online and lab experimental studies I show how psychological processes such as memory accessibility (Chapter 2), emotional arousal (Chapter 3), predictions errors (Chapter 4), and social norms (Chapter 5) can be leveraged to change people's beliefs. For instance, in a series of laboratory experiments, I find that strengthening the memory of a statement increases its believability, while weakening its memory in a targeted fashion decreases its believability (Chapter 2). Given the interdependence between memory and emotions, in a series of online experiments, I also show that pairing emotionally arousing images with statements subsequently increases the believability of these statements compared to statements that had been associated with neutral or no images (Chapter 3). In another series of online experiments I explore the effect of prediction errors on belief update. I find that people update their beliefs as a function of the size of the errors they make when evaluating relevant evidence, and that making large errors leads to more belief update than not engaging in prediction, while controlling for the evidence available. Importantly, I find that these effects hold across ideological boundaries (Democrats and Republicans, evaluating Neutral, Democratic, and Republican beliefs; Chapter 4). In the last series of experiments at the individual level, I show that people change their beliefs more

in line with evidence portrayed as normative (e.g., shared by many on social media platforms) compared to evidence portrayed as non-normative (Chapter 5.1) and that normativity cues signaled by large groups of people are the most effective at changing beliefs (Chapter 5.2). While the studies outlined in Part 2 focus on investigating phenomena at an individual level, extensive research shows that such cognitive processes are highly sensitive to the social context in which they manifest. Therefore, in Part 3, at the collective level of investigation, I uncover how macro-level societal outcomes can emerge from micro-level psychological processes. I first show how conversational interactions trigger belief change at a dyadic level, as individuals talking to one another change their beliefs to match those of their conversational partners (Chapter 6). Then, I find that collective level outcomes (i.e., collective beliefs) are influenced by both the conversational network structure that characterizes the community (Chapter 7) as well as by individual level mechanisms (Chapter 8). Finally, in Part 4 (Chapter 9) I propose future avenues of investigation, such as evaluating the connections between beliefs and behaviors, focusing on translating these controlled experimental strategies into applied settings. The goal of this translational work is to encourage a more active use of science in everyday life, for example, in developing actionable recommendations for policy makers and communicators to dispel misinformation at a societal level.

Acknowledgements

I would like to thank my parents and my sister, whose countless sacrifices allowed me to take on the path of scientific inquiry I had always aspired to. I would also like to thank my husband for his selfless and unconditional support, and my little Julie for brightening my world. I am also incredibly grateful to all the mentors I've had along the way who believed in me - Rich Handler, Anda Ritter, Mugurel Stefan - and my lifelong friends who kept me grounded - Miki Doman, Romina Iancu, Ajua Duker, Karina Tachihara. I would also like to thank the departments of Psychology and Neuroscience for creating the optimal environment for my academic and personal growth. Finally, I would like to thank my advisor, Alin Coman, for his unbounded enthusiasm, as well as my dissertation committee - Ken Norman, Betsy Levy Paluck, Eldar Shafir, Casey Lew Williams, Adele Goldberg - my research collaborators, the Socio-Conitive Processes lab, and my cohort for their continued support, help, and inspiration. I am also grateful for the funding sources that allowed me to conduct my graduate research: the Princeton University Cognitive Science Graduate Student Research Funding, the NSF Grant (BCS-1748285, BCS-2027225).

To my parents.

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Foundations

"Beliefs define how we see the world and act within it; without them, there would be no plots to behead soldiers, no war, no economic crises and no racism. There would also be no cathedrals, no nature reserves, no science and no art. Whatever beliefs you hold, it's hard to imagine life without them. Beliefs, more than anything else, are what make us human." — Graham Lawton

1.1 Beliefs

Beliefs are a mental construct that has sparked the interest of scientists for over a century (Lindsay, 1910), given their clear impact on all aspects of human societies - "The advertiser, salesman, journalist, author, politician, minister, and lawyer, cannot afford to be ignorant of the nature and conditions of belief" (Lund, 1925). And indeed, empirical work confirmed that people's beliefs can meaningfully impact their behavior (Shariff & Rhemtulla, 2012; Mangels, Butterfield, Lamb, Good, Dweck, 2006; Ajzen, 1991), although not always (Paluck, 2009), pointing to this relation's complexity. The factors influencing beliefs have also puzzled scholars for a long time, prompting research programs that identified various elements impacting beliefs, from evidence (Lund, 1925) to motivations (James, 1895). These early steps in the understanding of

both beliefs themselves and of the factors that impact them set in motion a burgeoning body of research of substantial theoretical and pragmatic developments. These developments are of particular importance today, considering technological advancements facilitating unprecedented exposure to false information (Gottfried & Shearer 2016; Vosoughi, Roy, Aral, 2018) with detrimental individual and societal consequences (Lewandowsky et al, 2012). For instance, believing false information has been linked to decreased vaccination rates (Jolley & Douglas, 2014), increased climate change denial (Lewandowsky, Gignac, Oberauer, 2015), and increased intergroup prejudice (Jolley, Meleady, Douglas, 2020). Misinformation occurs when people confidently hold false beliefs (Kuklinski et al, 2000). For example, a third of Americans believe global warming is a conspiracy (Swift, 2013), a third of American parents believe vaccines cause autism (National Consumers League, 2014), and 6.5 million Americans believe the Earth is flat (YouGov, 2018). The definition of the word 'post-truth' (declared the word of the year in 2016 by the Oxford Dictionaries) sheds some light on how such beliefs become widespread. When objective facts are less influential in shaping public opinion than appeals to emotion and personal beliefs, societies find themselves in a post-truth era (Oxford Dictionaries, 2016). Hence, in this dissertation, I strive to provide a deeper understanding of the socio-cognitive factors that can impact beliefs and the potential of translating and implementing this understanding into policies aimed at reducing misinformation at a societal level.

A belief has been defined as the mental acceptance of the truth of a statement (Schwitzgebel, 2010). Beliefs are thought to provide the 'mental scaffolding' for appraising one's environment (Halligan, 2007), thus constructing the "mental architecture" for interpreting the world (Jha, 2005). Beliefs are different from knowledge in the conviction they are held with (Fishbein & Ajzen, 1975; Jervis, 2006), and the self-referential element they embed (Connors & Halligan, 2014). Beliefs are also different from attitudes in that they lack the evaluative (e.g., good/bad) component of

attitudes, being centered instead on the accuracy component (i.e., true/false) (Eagly & Chaiken, 1993). The belief-dependent realism theory explains how beliefs are likely formed (Shermer, 2011). The mind, bombarded with sensory information, searches for patterns and infuses them with meaning. These meaningful patterns are our beliefs, which, once formed, are reinforced by selective incorporation of confirmatory evidence. Although a central feature of beliefs is their dynamic nature, as beliefs are constantly subject to change (Bendixen, 2002), certain beliefs can also be extremely persistent even in the face of contradictory evidence (Lewandowsky et al, 2012). And this persistence can become problematic when the belief in question is false. For instance, the now widespread belief that vaccines cause autism, originated in a scientific article published in 1998 in the Journal The Lancet. The belief has persisted for decades, even though it was revealed that the author had falsified the data, was found guilty of misconduct, and lost his medical license (Colgrove & Bayer, 2005; Larson et al, 2011). Nevertheless, this belief deterred many parents from vaccinating their children, which lead to an increase in preventable hospitalizations, deaths, and spending (Poland & Spier, 2010; Ratzan, 2010; Larson et al, 2011). And indeed, research on retractions has consistently found that retractions are rarely effective at changing false beliefs, even when people remember and believe the retraction (Ecker, Lewandowsky, Apai, 2011; Ecker, Lewandowsky, Swire, Chang, 2011; Ecker, Lewandowsky, Tang, 2010; Fein, McCloskey, Tomlinson, 1997; Gilbert, Krull, Malone, 1990; Gilbert, Tafarodi, Malone, 1993; Schul & Mazursky, 1990; van Oostendorp, 1996; van Oostendorp & Bonebakker, 1999; Wilkes & Leatherbarrow, 1988; Wilkes & Reynolds, 1999). Further research pointed to three categories of top down processing that could lead to the updating of beliefs: logical reasoning, motivation, and cognitive biases (Connors & Halligan, 2014). Within these, several processes have been empirically identified as playing a role in shaping beliefs. For example, beliefs can be impacted by processing fluency (e.g., phrases that rhyme are more believable; McGlone & Tofighbakhsh,

2000), language (e.g., phrases communicated in a native accent are more believable; Lev-Ari & Keysar, 2010), source credibility (e.g., statements from credible sources are more believable; Eagly & Chaiken, 1993), or repetition (e.g., statements that are encountered multiple times become more believable; Allport & Lepkin, 1945; Hasher, Goldstein, Toppino, 1977; Begg, Anas, Farinacci, 1992).

Even though beliefs have been the subject of scientific study of a large body of research, the cognitive architecture of belief systems lacks a clear framework to guide additional empirical advances (Bell et al., 2006a; Connors & Halligan, 2014).

1.2 A Theoretical Framework for Investigating Belief Change

I propose a theoretical framework of belief systems (Figure 1.1), which explores the internal scaffold of belief structures, facilitating a more systematic investigation of the construct of interest. Construed as a cognitive structure, the hereby introduced theoretical framework is comprised of three levels: evidence, beliefs, and social norms. On the first level, the evidence (i.e., fact-like information either favoring or contradicting beliefs) constitutes the knowledge base of the cognitive structure (e.g., "A 2014 analysis of 10 studies involving more than 1.2 million children reaffirms that vaccines don't cause autism"). This knowledge base can be objectively accurate or inaccurate, yet as long as it is subjectively endorsed as accurate it acts as viable evidence. The pieces of evidence are subjectively connected with each other according to a whole host of characteristics, including degree of similarity (i.e., topic correspondence) and compatibility (contradictory facts may be similar in topic but incompatible). Throughout, I will refer to the updating of beliefs as a function of the available evidence as rational belief updating, regardless of the objective accuracy of the evidence. The beliefs themselves (e.g., "Vaccines don't cause autism") account for the second level within

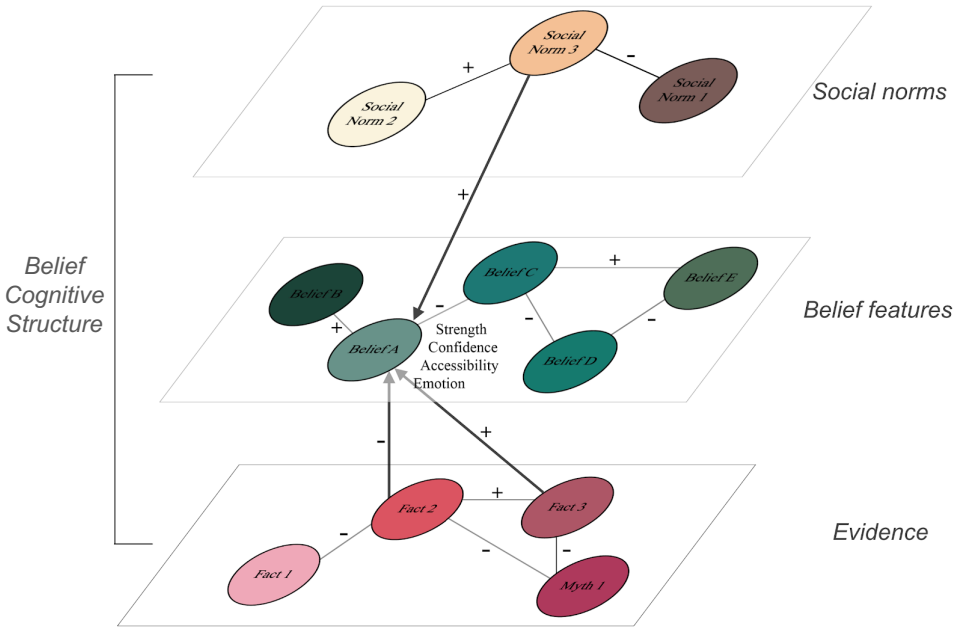


Figure 1.1: Individual level belief system framework, as a cognitive structure with 3 levels: evidence (facts and/or myths that support or contradict the belief), belief features (strength, confidence, accessibility, etc.), and beliefs about others' beliefs (social norms that reinforce or reject the belief). Compatible items (i.e., in agreement) are represented by the "+" symbol, incompatible items (i.e., in disagreement) are represented by the "-" symbol. The shading intensity of each color represents the strength of each belief.

the framework. Beliefs are more complex than factual information given their self-referential element, and the conviction they are held with (Fishbein & Ajzen, 1975). At this level, features such as strength, confidence, or accessibility play an essential role in belief endorsement and change (Babad, Ariav, Rosen, Salomon, 1987; Swann, Pelham, & Chidester, 1988; Brinol & Petty, 2009; Ozubko & Fugelsang, 2011). Similar to evidence, beliefs are dynamically connected with each other according to factors such as similarity and compatibility, in an Ising model fashion, initially used in physics to describe ferromagnetism (Ising, 1925) and later adapted in psychology to describe attitudes (Dalege et al., 2016) and beliefs (Brandt & Sleegers, 2021). Lastly, the third level of the cognitive structure around a belief is comprised of perceived social norms, or beliefs about others' beliefs (e.g., "Most Americans reject the theory that

vaccines cause autism"). Social norms have been long theorized to influence people's beliefs (Festinger, 1954), and indeed empirical explorations have showed that norms influence beliefs (Kim, 2018; Alhabash et. al., 2015). Importantly, these perceived norms do not always reflect reality. For instance, 74% of Democrats and independents would be comfortable with a woman president, yet only 33% believed others would be (Ipsos, 2019). As before, norms are also connected with each other according to factors such as similarity and compatibility.

This framework and opens and guides further directions of empirical exploration, of particular interest here being inquiries into processes effective at changing beliefs. For instance, do belief features such as memory accessibility and emotional arousal lead to belief change? Can making predictions about belief related evidence cause belief update? Are beliefs affected when a social norm is made salient? And what happens when the belief systems of different individuals come in contact with each other? Organized by this theoretical framework I strive to illuminate the internal scaffold of belief structures.

1.3 Collective Beliefs

I posed several research questions about belief change using the proposed theoretical framework. These questions were made mostly at an individual level, that is, how beliefs change within individual minds. However, the effect of cognition is not limited to individuals (Smith & Semin, 2007). The operations of the cognitive system influence (Kashima, 2014) and are influenced by (He, Lever, Humphreys, 2011; Smith & Semin, 2007) the social contexts in which they are manifested (Vlasceanu, Enz, & Coman, 2018). Indeed, a burgeoning literature has reported on the interaction between cognitive and social processes. These advances have led to a surge of interest in exploring the psychological processes involved in the

emergence of large-scale phenomena. Recent research has shown, for instance, that communities of individuals are characterized by shared attention (Risko, Richardson, Kingstone, 2016), collective memory (Hirst, Yamashiro, Coman, 2018), collective emotions (Yzerbyt, Kuppens, Mathieu, 2016), and collective action (Sebanz, Bekkering, Knoblich, 2006). Employing a mechanistic investigation of how the micro-level psychological phenomena of interest gives rise to macro-level social outcomes has the potential to bridge psychology and other social sciences, such as sociology (DiMaggio, 1997), anthropology (Sperber & Hirschfeld, 2004), and political science (Balcells & Justino, 2014). Although psychological research rarely deals with large groups characterized by networked interactions, sociologists have long established that a community's network structure affects a whole host of outcomes, from information propagation (Bakshy, Hofman, Mason, Watts, 2011) to adoption of health behaviors (Centola, 2011). Given that false beliefs are especially dangerous when endorsed by a large proportion of people, as they can lead to resistance to true information (Chua & Banerjee, 2017; Nagler, 2014), shift attention and resources away from real threats (Kuklinski et al, 2000), and cause suboptimal collective decisions (Lewandowsky, Ecker, and Cook 2017), in my exploration of belief change I am also interested in revealing meaningful properties of collective beliefs. Defined as the joint commitment of a group to accept a statement as true (Gilbert, 1987, 1994), collective beliefs are also malleable, subject to influences from social interactions within communities. Such influences can cause the convergence or divergence of individuals' beliefs, processes of central interest in my dissertation. Thus, in the collective processes chapters I address questions such as: what are the effects of conversational interactions on belief change? What role does the network structure play in determining a community's collective beliefs? And how do individual level belief change effects manifest in larger social networks of interacting individuals?

Understanding the individual and collective belief structures and dynamics is a first critical step in tackling one of the top threats faced by contemporary society: the misinformation epidemic (Farkas & Schou, 2019; Lewandowsky et al, 2012). Consequently, I propose targeting inaccurate beliefs with easy to implement interventions using the insights derived from carefully controlled studies inspired by the proposed cognitive framework. In the first chapter, I consider memory accessibility, and its potential to alter beliefs at the individual level.

Part II

Individual Beliefs

Chapter 2

Memory and Beliefs

2.1 Study 1: Memory Accessibility Triggers Belief Change

This Chapter is based on the paper "Mnemonic accessibility affects statement believability: The effect of listening to others selectively practicing beliefs" published in the Journal *Cognition* in 2018. The co-author of this publication is Alin Coman.

Abstract

Belief endorsement is rarely a fully deliberative process. Oftentimes, one's beliefs are influenced by superficial characteristics of the belief evaluation experience. Here, we show that by manipulating the mnemonic accessibility of particular beliefs we can alter their believability. We use a well-established socio-cognitive paradigm (i.e., the social version of the selective practice paradigm) to increase the mnemonic accessibility of some beliefs and induce forgetting in others. We find that listening to a speaker selectively practicing beliefs results in changes in believability. Beliefs that are mentioned become mnemonically accessible and exhibit an increase in believability, while beliefs that are related to those mentioned experience mnemonic suppression, which

results in decreased believability. Importantly, the latter effect occurs regardless of whether the belief is scientifically accurate or inaccurate. Furthermore, beliefs that are endorsed with moderate-strength are particularly susceptible to mnemonically-induced believability changes. These findings, we argue, have the potential to guide interventions aimed at correcting misinformation in vulnerable communities.

Introduction

Does ingesting sugar cause hyperactivity in children? The belief that it does is widespread in the population, despite scientific evidence to the contrary. On the one hand, answering the question in the affirmative could be because one has information that is supportive of the belief. On the other hand, belief endorsement could be due to superficial characteristics of the belief evaluation experience. Among these superficial characteristics, the ease with which information comes to mind has been found to influence one's judgments (Tversky & Kahneman, 1973). This ease of retrieval is taken as an internal cue as to whether one endorses it: high endorsement if the belief comes to mind easily, low endorsement otherwise.

Most of the experimental studies aimed at exploring the relation between memory and belief focuses on the up-regulation of memory. That is, increasing a belief's mnemonic accessibility has been shown to result in its increased believability (Ozubko & Fugelsang, 2011). No research to date has explored how the down-regulation of memory (i.e., mnemonic suppression) can lead to corresponding changes in belief endorsement. This latter investigation is important for both theoretical and practical reasons. On the theoretical side, the argument that mnemonic accessibility causally influences believability has to necessarily explore both sides of the mnemonic accessibility continuum: up-regulation and down-regulation. On the practical side, at a societal level decreasing the believability of inaccurate beliefs in the population might be as important as increasing the believability of accurate beliefs.

To explore the relation between mnemonic down-regulation and believability, I build on a well-established literature that shows that selective practice of previously encoded information can result in better memory for practiced information – a rehearsal effect - and can also induce forgetting in unmentioned, but related to the mentioned information – a retrieval-induced forgetting effect (Anderson, Bjork, & Bjork, 1994). In a typical selective practice paradigm, participants first study category-exemplar pairs (e.g., the "Nutrition" category contains the "Carrots are rich in vitamins" and "Broccoli is rich in iron" exemplars; the "Hydration" category contains the "Milk is rich in calcium" and "Coconut water is rich in potassium" exemplars) and then receive selective practice for half of the exemplars from half of the categories by way of a stem completion task (e.g., "Carrots are rich in v_"). Analyses of a final cued-recall test show that practiced items (Rp+ items: Nutrition- Carrots/Vitamins) are remembered better than unpracticed unrelated items (Nrp items: the exemplars in the Hydration category)—a rehearsal effect. Unpracticed items related to those practiced (Rp- items: Nutrition-Broccoli/Iron) are remembered worse than Nrp items—a retrieval-induced forgetting effect (RIF). The rehearsal effect has been explained by trace strengthening (Karpicke & Roediger, 2008), whereas RIF is thought to arise because of inhibitory processes triggered by response competition during the practice phase (Kuhl, Dudukovic, Kahn, & Wagner, 2007; but see Mensink & Raaijmakers, 1988). Of note, RIF is a well-established phenomenon that is reliably obtained with various stimulus materials and delay intervals (Murayama, Mityatsu, Buchli, & Storm, 2014, for a meta-analysis). It has also been consistently found when the selective practice of information occurs in a conversational setting (Coman, Manier, & Hirst, 2009). That is, when listeners monitor the speaker selectively practicing previously encoded information they experience what Cuc, Koppel, and Hirst (2007) call socially-shared retrieval-induced forgetting. This phenomenon, they showed, is due to the fact that under certain circumstances, listeners concur-

rently retrieve the information along with the speaker, which, like in the case of RIF, triggers response competition from related memories.

In the current study I reasoned that the easiness with which a belief comes to mind should affect its believability. The two cognitive processes triggered by selective retrieval practice (i.e., strengthening and suppression) should lead to corresponding effects on believability. Because repeated exposure to a belief leads to increased mnemonic accessibility, one would expect an increase in its believability, a prediction consistent with research on the illusory truth effect (Fazio, Brashier, Payne, & Marsh, 2015). At the same time, beliefs related to those practiced should experience suppression of their mnemonic representations, which should in turn result in decreased believability.

But not all information can be suppressed. Recent research has found that moderately activated memories are most susceptible to forgetting (Newman & Norman, 2010; Poppenk & Norman, 2014). This is due to the fact that weakly activated memories do not have the strength to trigger competition among memory traces, while highly activated memories are too strong to experience suppression. During the selective practice phase, therefore, weakly activated Rp- memories are unlikely to compete for activation, while strongly activated Rp- memories will exceed the activation threshold. For these reasons, neither should experience suppression following selective practice. Transferring this reasoning in the domain of beliefs, it follows that only moderately held beliefs should experience suppression following selective practice. In other words, if one strongly endorses or strongly opposes the belief that "sugar makes kids hyperactive," than this endorsement/opposition might make the belief chronically accessible, and, therefore, less susceptible to suppression.

Several findings in the retrieval-induced forgetting literature are consistent with this prediction. Evidence for a relation between belief strength and probability of retrieval comes from research on memory for stereotypes. Dunn and Spellman (2003)

found that the more strongly participants endorsed a stereotype, the less suppression of stereotype-relevant information they exhibited. Similarly, Coman and Hirst (2012) found that the participants who held extreme views on a topic (i.e., legalization of euthanasia) were less likely to experience retrieval-induced forgetting in topic-relevant information compared to participants who held moderate views. Based on this research I hypothesize that only moderately-held beliefs will be susceptible to forgetting and its hypothesized believability decrement. To test these hypotheses, I conducted two studies. After collecting data for the main study between October 2017 and January 2018, I conducted an exact replication study between March and May (2018) with a separate sample of participants recruited from the same population (i.e., Princeton students).

Method

Open science practices. The data and stimulus materials can be found on my open science framework page: https://osf.io/xe6nd/

Participants. To detect a moderate effect size of 0.30 for paired-sample comparisons with 0.80 power, I collected data from 80 participants. Pilot testing the procedure indicated that finishing the task in less that 15 min constituted inadequate study engagement. I therefore used this pre-established criterion to discard participants. The final sample was comprised of 58 participants affiliated with Princeton University (66% women; Mean-Age = 21.76). For the replication study, I recruited 100 participants, with similar calculations for the projected sample size. Eighty-eight participants affiliated with Princeton University (56% women; Mean-Age=20.58) completed the study and passed the pre-established exclusion criterion. The study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. A set of 24 statements distributed in four categories (i.e., nutrition, allergies, vision, health) was selected to be used in the main study. Each

category was comprised of 2 myths and 4 correct pieces of information. The myths were comprised of statements commonly endorsed by individuals as true, but in fact are false, whereas the facts were scientifically accurate statements. For example, a myth was that "reading in dim light can damage children's eyes," while an accurate statement was that "children who spend less time outdoors are at greater risk to develop myopia." Based on a separate pilot study with 112 Amazon Mechanical Turk participants (46% women; Mean-Age = 35.18; SD-Age = 10.85), the myths and correct pieces of information were selected such that they were not different on believability, perceived scientific support, and personal relevance. In addition, I selected the categories for which the items were correctly categorized as being part of a category by more than 75% of the sample.

Design and Procedure. The study included 4 phases. In the study/evaluation phase, participants were presented with 24 statements in a category-blocked fashion. The order of presentation of the categories, as well as the order of statements within the category was random. Participants were instructed to carefully read these statements that are, supposedly, "frequently encountered on the internet." In this phase, they rated the degree to which they believe each statement is accurate (from 1-Not at all to 7-Very much so) and has scientific support (from 1-Definitely not to 7-Definitely yes). Next, participants went through a selective practice phase. They listened to an audio of a participant who, supposedly, remembered the information he/she was exposed to during the experiment in a previous session. In the audio, the speaker (a confederate) recalled the statements with minor hesitations to indicate a naturalistic recall. The participants (listeners) were asked to carefully monitor the speaker's utterances for whether the speakers were accurately remembering the initially studied statements. Each participant listened to a gender-matched audio containing half of the correct statements (i.e., 2 statements) from half of the categories (i.e., 2 categories), for a total of 4 statements. Counterbalancing the

selectively practiced stimuli ensured that every correct statement was equally likely to be RP+ (beliefs mentioned in the audio), RP- (beliefs in the same category as those mentioned in the audio, but not mentioned themselves), or NRP (beliefs that were not mentioned in the audio and were unrelated to those mentioned). RP+ beliefs were always correct. In each category, RP- beliefs were either correct (2 beliefs) or incorrect (2 beliefs). Similarly, NRP items were either correct (4 beliefs) or incorrect (2 beliefs). Participants were then asked to recall the information in a cued-recall task. They were given the category name (e.g., Nutrition) and were instructed to remember the initially studied statements. Finally, in a belief post-test phase, participants were randomly presented with the initially read statements and were asked to rate them on the same two scales as before (i.e., accuracy and scientific support).

Analyses and Coding. Each statement was coded as successfully remembered if the recall captured the gist of the original statement. For instance, if for the studied item "Crying helps babies' lungs develop," participants remembered "Crying is good for the lungs," their recall was coded as accurate, since it captures the gist of the original statement. Ten percent of the data were double-coded for reliability (Main study kappa = 0.87; Replication study kappa = 0.87) and all disagreements were resolved through discussion. All reported statistical analyses are computed following the guidelines and corrections described in Lakens (2013).

Results

First, I wanted to establish whether there is any difference in believability between the accurate statements and the myths in the initial belief evaluation phase. Because the correlation between a statement's believability and its perceived scientific evidence was high (r = 0.80) I averaged the two scores to compute a believability index. A paired-sample t-test comparing the believability indices of facts (M = 4.03, SD = 0.69)

and myths (M = 3.88, SD = 0.79) revealed no difference between them, t(57)=1.28, d=0.17, p=0.203. In subsequent analyses, I combined the facts and myths, but note that I conducted analyses separately for facts/myths and the pattern of results shows no significant differences between the two. This is consistent with the fact that from the perspective of the participants they were indistinguishable.

To explore whether the selective practice phase had an effect on the participants' memories of the statements, I conducted a Repeated-Measures ANOVA with Retrieval-Type (Rp+ vs. Rp- vs. Nrp) as a within-subjects variable, and recall proportion as the dependent variable. I found a significant main effect for Retrieval-Type, F(2, 56)=23.63, p < .001, η_p^2 = 0.46. Post-hoc analyses revealed a rehearsal effect, with the recall proportion of Rp+ items (M = 0.66, SD = 0.31) significantly higher than the recall proportion of Nrp items (M = 0.40, SD = 0.16), t(57) = 6.47, p < .001, d = 0.86, CI[0.18, 0.34]. I also found a retrieval-induced forgetting effect, with the recall proportion of Rp- items (M = 0.34, SD = 0.20) significantly lower than the recall proportion of Nrp items (M = 0.40, SD = 0.16), t(57) = 2.15, p < .036, d = 0.28, CI[0.00, 0.12]. A similar effect size was found for retrieval-induced forgetting when I restricted the analyses to the misinformation items, with Rp- items (M = 0.31, SD = 0.23) remembered significantly lower than Nrp items (M = 0.39, SD = 0.23), t(57) = 2.22, p < .030, d = 0.29, CI[0.00, 0.15]. I thus replicated previous research showing that selectively practicing information leads to the strengthening of the retrieved information and the retrieval-induced forgetting of unmentioned, but related to the mentioned information.

The replication study exhibited a more pronounced pattern in the hypothesized direction, given that I had a more adequate sample size to detect an effect: in a Repeated-Measures ANOVA with Retrieval-Type (Rp+ vs. Rp- vs. Nrp) as a within-subjects variable, and recall proportion as the dependent variable, I found a significant main effect for Retrieval-Type, F(2, 86)=76.90, p<.001, η_p^2=.64. Post-

hoc analyses revealed a rehearsal effect, with the recall proportion of Rp+ items (M=.69, SD=.26) significantly higher than the recall proportion of Nrp items (M=.36, SD=.23), t(87)=9.61, p< .001, d=1.38, CI[.26, .39]. I also found a retrieval-induced forgetting effect, with the recall proportion of Rp- items (M=.27, SD=.23) significantly lower than the recall proportion of Nrp items (M=.36, SD=.23), t(87)=3.02, p<.003, d=.39, CI[.03, .16]. A similar effect size was found for retrieval-induced forgetting when I restricted the analyses to the misinformation items, with Rp- items (M=.28, SD=.23) remembered significantly lower than Nrp items (M=.36, SD=.23), t(87)=3.12, p<.002, d=.35, CI[.03, .13].

I further hypothesized that moderate beliefs will be more susceptible to forgetting than either low-strength or high-strength beliefs. To investigate this hypothesis, I categorized, for each participant, the 24 beliefs depending on their strength. To do so, I used the belief evaluation from the study/evaluation phase to compute within-participant z-scores. That is, for each participant, the set of 24 belief scores from the encoding phase constituted the pool of scores used for standardization. I used cutoff scores to split the beliefs in participant-specific High-Strength beliefs (z-scores > 0.5; MRawScore = 5.65, SD = 0.80), Moderate-Strength beliefs (z-scores between 0.5 and -0.5; MRawScore = 4.03, SD = 0.81), and Low-Strength beliefs (z-scores $<$ -0.5; MRawScore = 2.23, SD = 0.83), such that, on average, across participants, 33% of a participant's beliefs fell in each of the three belief strength categories.

I next conducted paired-sampled t-tests, separately for the rehearsal and RIF effects for each belief-strength. I found a reliable rehearsal effect for low-strength beliefs, $t(41) = 2.40$, $p < 0.021$, $d = 0.38$, CI[0.03, 0.37], moderate-strength beliefs, $t(45) = 3.26$, $p < 0.002$, $d = 0.49$, CI[0.05, 0.37]. Regardless of belief strength, if one listens to another person repeat the belief, then one is likely to remember it subsequently. As predicted, I only found a RIF effect for Moderate-Strength beliefs,

t(53) = 2.24, p < .029, d = 0.31, CI[0.01, 0.25], and not for Low-Strength (p = .377) or for High-Strength (p = .324) beliefs.

The same results emerged in the replication study, with the magnitude of the effect size slightly larger than that in the main study: I found a reliable rehearsal effect for low-strength beliefs, t(62)=5.49, p<.001, d=.68, CI[.21, .44], moderate-strength beliefs, t(72)=6.03, p<.001, d=.71, CI[.21, .43], and high-strength beliefs, t(69)=5.84, p<.001, d=.69, CI[.22, .46]. As predicted, I only found a RIF effect for Moderate-Strength beliefs, t(81)=2.73, p<.008, d=.33, CI[.03, .18], and not for Low-Strength (p=.09) or for High-Strength (p=.08) beliefs.

To explore whether rehearsal and RIF effects impact statement believability, I computed the change in belief endorsement as a function of whether the item was an Rp+, Rp-, or Nrp item during the selective practice phase. I maintained the designation of High-Strength/Moderate-Strength/Low-strength beliefs based on the pre-test phase and I computed post-test belief z-scores for each participant. That is, for each participant, the 24 beliefs that they evaluated in the post-test constituted the set of belief scores that were used for within-individual standardization. For a measure of belief strength change, I subtracted each belief's pre-test z-score from its post-test z-score. The more positive this belief change score - the more believable the belief became after the selective practice phase than before; the more negative the belief change score - the less believable the belief became after the selective practice phase than before. It is important to note that I used standardized scores for this analysis to avoid fatigue or habituation effects from pre to post-evaluation.

I conducted a Repeated-Measures ANOVA with Retrieval-Type (Rp+ vs. Rp- vs. Nrp) and Belief-Strength (High vs. Moderate vs. Low) as within-subjects variable, and belief change as the dependent variable. I found a significant main effect for Retrieval-Type, F(2, 18) = 3.81, p < .042, $\eta_p^2 = 0.30$ and for Belief-Strength, F(2, 18) = 15.80, p < .001, $\eta_p^2 = 0.64$. I also found an interaction between Retrieval-Type and

Belief-Strength, F(4, 16) = 3.59, p < .028, η_p^2 = 0.47. In exploring the interaction, I focused on comparisons involving Rp+ beliefs and Nrp beliefs (belief rehearsal effect) and Rp- beliefs and Nrp beliefs (belief suppression effect). For the belief rehearsal effect, the Rp+ High-Strength beliefs decreased in believability from pre to post-evaluation (M = -0.08, SD = 0.36) to a smaller extent than the Nrp High-Strength beliefs (M = -0.36, SD = 0.39), t (46) = 3.61, p < .001, d = 0.53, CI[0.12, 0.44]. Similarly, the Rp+ Moderate-Strength beliefs increased in believability from pre to post-evaluation (M = 0.21, SD = 0.58) more than the Nrp Moderate-Strength beliefs (M = -0.12, SD = 0.44), t(46) = 2.98, p < .004, d = 0.44, CI[0.11, 0.55]. The pattern of results for High and Moderate-Strength beliefs showcases, thus, a belief rehearsal effect. I found no significant belief rehearsal effect for the Low-Strength beliefs, with the belief change in Rp+ beliefs (M = 0.55, SD = 0.82) larger than that for Nrp beliefs (M = 0.33, SD = 0.39), as predicted, but not significantly so, t(43) = 1.51, p = .138, d = 0.23, CI[-0.07, 0.51]. These results suggest that monitoring a speaker repeating beliefs increases their believability, but only if one believes them at least moderately (Figure 2.1).

Finally, I explored whether retrieval-induced forgetting resulted in diminished endorsement of unmentioned, but related to the mentioned beliefs. As predicted, I found a belief suppression effect only for Moderate-Strength beliefs, with the Rp- beliefs decreasing in believability from pre to post-evaluation (M = -0.29, SD = 0.40) significantly more than the Nrp beliefs (M = -0.11, SD = 0.46), t (53) = 2.05, p < .045, d = 0.28, CI[0.00, 0.36]. The Low-Strength beliefs and the High-Strength beliefs did not show diminished endorsement following selective practice. This is consistent with the pattern of recall that I explored above, suggesting that the suppression of mnemonic representations associated with particular beliefs diminishes their believability (Figure 2.1).

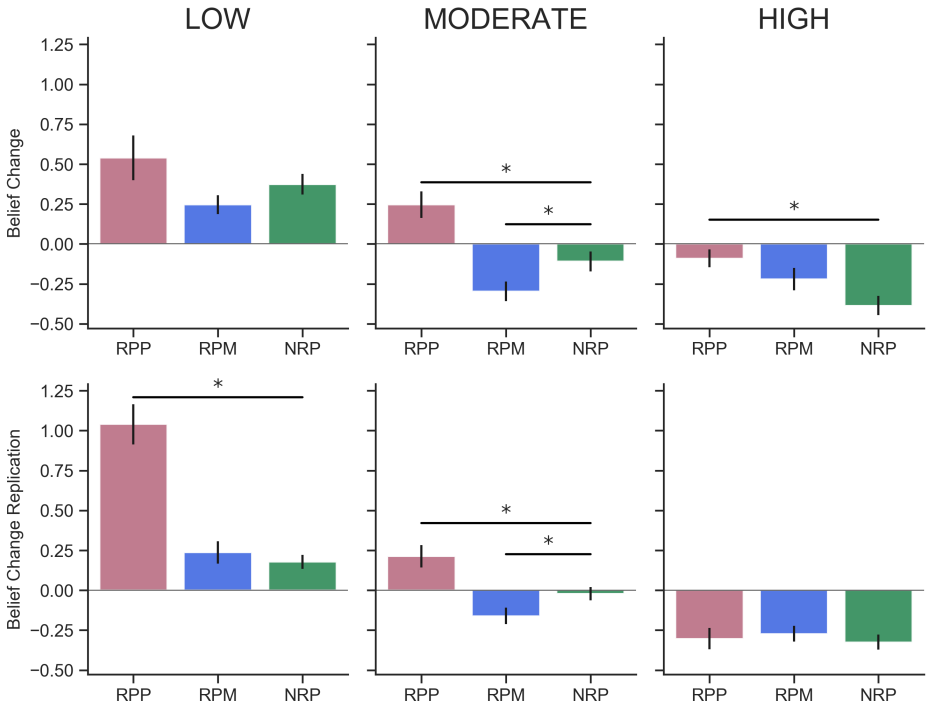


Figure 2.1: Belief change scores by retrieval type (Retrieval Practice Plus or RP+, Retrieval Practice Minus or RP-, and No-Retrieval Practice or NRP), separate for initially Low-Strength, Moderate-Strength, and High-Strength beliefs in the main study (top panel) and the replication study (bottom panel). Error bars represent ±1 standard errors of the mean.

The replication study exhibited a similar pattern for moderately held beliefs: I found a belief suppression effect only for Moderate-Strength beliefs, with the belief change in Rp- items (M=-.15, SD=.42) significantly smaller than that of the Nrp items (M=-.02, SD=.33), t(81)=2.09, p<.04, d=.23, CI[.01, .25]. The Low-Strength beliefs and the High-Strength beliefs did not experience diminished endorsement following selective practice. Notably, the belief rehearsal effect for low-strength beliefs that only showed a statistical trend in the main study became highly significant in the replication study ($p < .001$), while the significant rehearsal effect for high-strength beliefs that was statistically significant in the main study was no longer present in

the replication study (Figure 2). I speculate that these differences might have been due to the fact that the low-strength beliefs in the replication study were endorsed significantly less (MRawScore = 1.58, SD = 0.49) than the low-strength beliefs in the main study (MRawScore = 2.23, SD = 0.83) ($p < .001$), which made the selective practice of these beliefs more consequential in the replication study. Similarly, the high-strength beliefs were endorsed much more in the replication study (MRawScore = 6.43, SD = 0.50) than in the main study (MRawScore = 5.65, SD = 0.80), ($p < .001$), which made the selective practice of these beliefs non-consequential in the replication study, since they were already highly endorsed.

Discussion

People constantly communicate with one another about their beliefs. Here I show that monitoring others mentioning beliefs affects listeners in meaningful ways: listeners' mnemonic accessibility for the mentioned beliefs increases, and their accessibility for unmentioned, related beliefs decreases. This change in mnemonic accessibility of beliefs, in turn, affects their believability, such that rehearsed beliefs become more believable and suppressed beliefs become less believable. Two important qualifications of this conclusion are in order. First, I obtained these results with participants who were engaged in the experimental task, given the pre-established exclusion criteria. This limits, to some degree, the generalization of this conclusion to the population at large. I believe that it is fair to conclude, though, that given at least a moderate level of engagement, one could reasonably expect these results to hold in the general population. And second, the belief suppression effect only happened for moderately-held beliefs, a finding consistent with previous work that shows forgetting for moderately activated memories (Newman & Norman, 2010). Beyond its theoretical importance, this finding suggests that in communities in which inaccurate beliefs are widely cir-

culated, one would be well-served to know which beliefs could be most susceptible to change through mnemonic accessibility strategies.

The research presented here opens intriguing research trajectories. First, in the current study selective practice is implemented with the use of an implied ingroup social source (i.e., speaker in the audio thought to be another Princeton-affiliated participant). Extensive social psychological research has shown that the characteristics of the source have a critical impact when it comes to belief endorsement (Pornpitakpan, 2004). I conjecture that programmatically manipulating the profile of the source – e.g., political ideology, trustworthiness, expertise – will likely affect both the memories of the mentioned and unmentioned information (Coman & Hirst, 2015) as well as the believability of this information in predictable ways.

Finally, the current findings are relevant in the context of interventions aimed at countering the spread of misinformation in vulnerable communities (Coman & Berry, 2015; Schwarz, Sanna, Skurnik, & Yoon, 2007). Existing interventions involve information campaigns that typically attempt to counter false information by refuting it (Wegner, Wenzlaff, Kerker, & Beattie, 1981). Even though these campaigns are generally successful (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Wood & Porter, 2018), under certain conditions they have been found to reinforce misconceptions (Nyhan & Reifler, 2010; Cook, Ecker, & Lewandowsky, 2015). Supplementing existing interventions with alternative strategies (i.e., suppressing false beliefs through retrieval-induced forgetting), will add to the tools that policy makers and communicators could use to dispel misinformation at a societal level.

In Chapter 2, I showed that memory accessibility can change beliefs, one implication being that strategies aimed at increasing the mnemonic accessibility of beliefs results in belief enhancement. A well-established way to enhance memory is through emotionally charged images (Cahill et al, 1994). Thus, the next belief feature I

investigated as to whether it can lead to belief change was emotion, specifically emotion induced by visual imageries.

Chapter 3

Emotions and Beliefs

3.1 Study 2.1: Emotional Arousal Triggers Belief Change

This Chapter is based on the paper "The Emotion-Induced Belief Amplification Effect" published in the *Proceedings of the Annual Meeting of the Cognitive Science Society* in 2020. Sections 2.2, 2.3, and 2.4 are new follow-up experiments conducted subsequently. The co-authors of the publication are Jacob Goebel and Alin Coman.

Results in this Chapter have also been presented at the 32nd APS Annual Convention.

Abstract

Exposure to images constitutes a ubiquitous day-to-day experience for most individuals. From mass-media exposure, to engagement with social-networking sites, to educational contexts, we are bombarded with images. Here, we explore the effect that emotional images have on belief endorsement. To investigate this effect, we test whether statements accompanied by emotionally arousing images become more or less believable than the same statements when they are accompanied by neutral

images or by no images. In three online samples, we find that emotional images increase statement believability (Experiment 2.1a, replicated in preregistered Experiment 2.1b), effect likely driven by an increased mnemonic accessibility mechanism (Experiment 2.2). This effect, however, failed to replicate in a Princeton University student sample (Experiment 2.3). When testing a binding mechanism (again on Princeton students) to explain this replication failure, we failed again to observe the original effect (Experiment 2.4). A potential explanation why the effect of emotional arousal on belief change is only observable in online samples but not in Princeton student samples may be that Princeton students process information more rationally, in a way that makes them immune to such subtle biasing effects. This speculative explanation for the boundary case of the emotion-induced amplification effect needs empirical evidence.

Introduction

Humans are a highly visual species. They can process an image in only 13 milliseconds (Potter et al, 2014), they can remember for days 2000 images they've been minimally exposed to (Grady et al, 1998), they are more persuaded if an argument contains visual aids (Vogel, Dickson, & Lehman, 1986), and they judge statements as true more often when these statements are accompanied by an image (Newman et al, 2012). Thus, images can have a powerful influence on people, especially if they elicit emotions. A classic example of an emotional image having a powerful impact on human society is Sam Shere's 1937 photograph of the Hindenburg airship traveling from Frankfurt and arriving in New Jersey in flames (Figure 3.1). This image captured the areal tragedy and thereby instilled the belief that Hydrogen fueled passenger air travel is highly dangerous. It put an end to this means of travel, even though this hadn't been the first, nor the deadliest such incident (Lowndes, 2019).

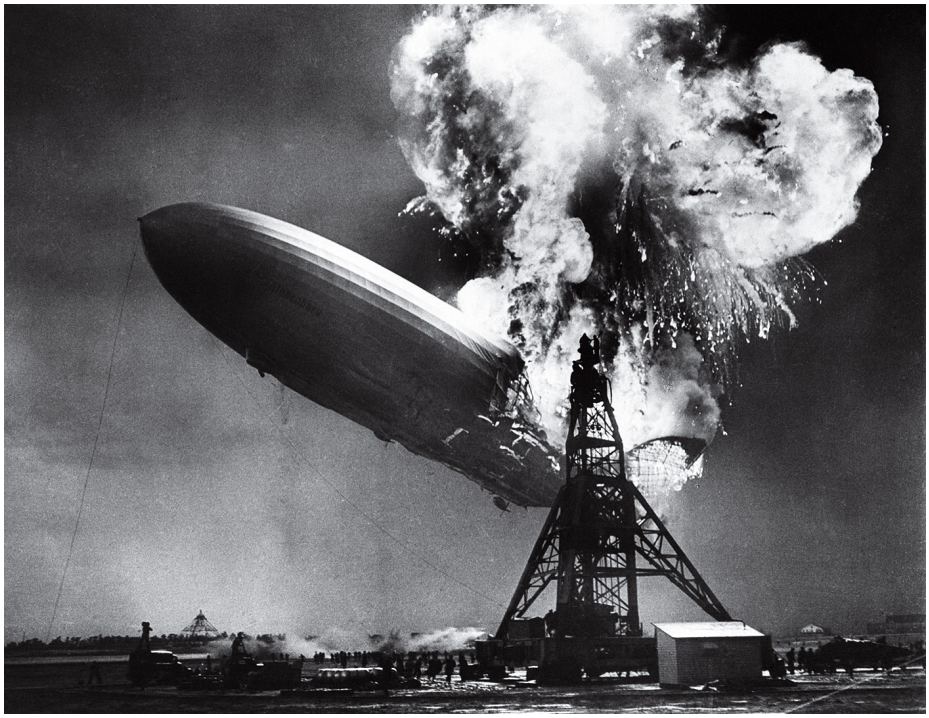


Figure 3.1: Sam Shere's famous picture of the Hindenburg airship as it exploded while landing in Lakehurst, New Jersey, on May 6, 1937.

The power of impactful, emotional images hasn't been lost in journalistic reporting. Thus, being a consumer of news media guarantees your exposure to emotionally arousing imagery, from natural disasters and wars, to tragic accidents, and grieving victims (Newhagen, 1998; Allan & Zelizer, 2004; Miller, 2006; Shoshani & Slone, 2008; Höijer, 2010). And reporters' use of emotional visual content to capture the attention of their audiences is not just based on lay intuitions on how to win the competition for sensationalism. Empirical research provides strong evidence for the hypothesis that viewers are more engaged with information presented by news reports that feature emotionally arousing content. For example, viewers are more likely to attend to news articles if these are accompanied by visual depictions of victimization, (Zillmann, Knobloch, & Yu, 2001), threatening images (Knobloch et al, 2003), or alarming images (Gibson & Zillmann, 2000).

Furthermore, images presented alongside information as supporting material have also been found to influence the believability of the information conveyed. For example, including brain images in neuroscience articles increases the believability of the articles' conclusions (McCabe & Castel, 2008; but see Schweitzer, Baker, & Risko, 2013). In another study, smoking warnings accompanied by images illustrating smoking hazards have been found to be more believable than written warnings alone (Shi et al, 2016). Despite this burgeoning literature, the impact of emotional images on message believability remains unexplored. Here, I are investigating whether statements eliciting negative emotional arousal by having been previously associated with arousing visual stimuli become more or less believable than those same statements eliciting neutral arousal by having been previously associated with either neutral or no visual stimuli.

The belief formation literature provides compelling indications that emotional arousal could influence the believability of information. On one hand, emotional images may enhance believability of associated statements by making these statements more memorable. Extensive research shows that emotionally charged events experience enhanced encoding and subsequent recall (Cahill et al, 1994). For example, in a naturalistic experiment, Miller (2006) found that television news viewers were more likely to recall information when the news report was more emotional in nature, especially when it elicited feelings of disgust (Miller, 2006). And better memory for an experienced event has been shown, in a different context, to increase believability. Repeated presentation of trivia facts leads to increased believability of the facts, a phenomenon known as the illusory truth effect (Begg, Anas, & Farinacci, 1992; Ozubko & Fugelsang, 2011; Vlasceanu & Coman, 2018). This effect is so robust it has been shown to hold even in the presence of countervailing knowledge (Fazio, Brashier, Payne, & Marsh, 2015). Based on these findings, one would expect

that emotionally arousing images would increase believability, an effect potentially mediated by memory accessibility.

On the other hand, emotional arousal could negatively affect belief endorsement. This prediction is supported by at least two potential mechanisms: a heuristic judgement mechanism (Murphy & Zajonc, 1993; Tversky & Kahneman, 1973; Petty & Briñol, 2015), and an attention mechanism (Loftus, Loftus, & Messo; 1987; Loftus, 1979; Hope & Wright, 2007). The former mechanism would posit that since emotions can be used as heuristics for judgements, negative feelings during statement evaluations could be misinterpreted as disagreement with the statement (Petty & Briñol, 2015). The latter mechanism would predict that negative emotional images may capture attention in a weapon focus effect manner (Loftus, Loftus, & Messo; 1987; Loftus, 1979; Hope & Wright, 2007), leading to increased memory for the image but suppressed memory for the statement, which can then lead to a decrease in the statement's believability (Vlasceanu & Coman, 2018). Thus, both mechanisms would predict negative emotional arousal to decrease statement believability.

Here, I investigate the impact of emotional images on believability in a main study and a preregistered direct replication study. Participants first rated the believability of a set of statements. They were then exposed to the statements again – this time alongside images that were conceptually relevant and negatively-valenced. The images were either highly arousing, neutrally arousing, or blank screens. I was interested to assess the degree to which the believability of the statement is affected by the arousing image that accompanied it, as measured in a subsequent believability task.

Study 2.1: Main Effect

Method

Open science practices. The data and stimulus materials can be found on my open science framework page: https://osf.io/6f4hk/

Participants. To detect a moderate effect size of 0.30 for within-subject comparisons with 0.80 statistical power I estimated a sample of 90 participants. Based on previous studies conducted in the lab which result in approximately 10% of the sample discarded due to pre-established criteria, I collected data from a total of 107 participants. Participants were recruited on Cloud Research (Litman, Robinson, & Abberbock, 2016), an Internet-based research platform similar to Amazon Mechanical Turk (MTurk) and were compensated at the platform's standard rate. The study protocol was approved by the Princeton University Institutional Review Board. Of the 107, 5 participants failed the attention checks and were therefore discarded from further analyses, following the pre-established discarding criteria. The attention checks were: two open-ended questions asking participants what their favorite food is (1) and how their day is going (2). An additional exclusion criterion was embedded in the task and involved indoor/outdoor and animate/inanimate judgements for each image (3). Participants who provided no answers to (1) and (2) or incorrectly answered more than 30% of the image judgments were discarded from analyses. After the exclusions, I performed statistical analyses on the final set of 102 participants (Mage=49.21, SDage=17.26; 61% women).

In a direct replication, I collected data from a total of 104 participants, also recruited on Cloud Research, and compensated at the platform's standard rate. Of the 104, 4 participants failed the attention checks and were therefore discarded from further analyses as stated in the preregistration. After the exclusions, I performed statistical analyses on the final set of 100 participants (Mage=55.8, SDage=14.98; 63% women).

Stimulus materials. I undertook preliminary studies to develop a set of 42 state-

ments and their associated images. These statements were equally split into: 21 correct pieces of information (e.g., "There are more jails than colleges in the U.S."), and 21 incorrect pieces of information. (e.g., "Using a phone while pumping gas can ignite a fire."). The 42-statement set was selected from a larger initial set of 70 statements that I pretested using the Qualtrics platform on Cloud Research (N=153; Mage=35.06, SDage=10.55; 39% women). I matched the correct and incorrect statements on three dimensions: perceived believability (i.e., as measured by the composite value of the questions "How accurate do you think this statement is?" and "Do you think there is evidence that supports this statement?"), perceived relevance (i.e., "How relevant is this statement to you?"), and perceived emotionality (i.e., "How emotionally charged do you think this statement is?"). All questions involved 0-100-point scales. The final set of chosen statements was selected such that the 21 true statements (Facts) did not differ significantly from the 21 false statements (Myths) on these dimensions.

I also developed a set of 42 pairs of images, each pair being representative of a statement. Of the two images in a pair, one was intended to be more emotionally arousing than the other. Otherwise, the two images were intended to be equivalent on several dimensions. To construct the intended stimuli set, I pretested a total of 288 images using the Qualtrics platform on Cloud Research (N=203; Mage=36.70, SDage=28.99; 40% women) on four dimensions: emotional arousal (i.e., as measured by the question "How emotionally arousing do you find this image?"), emotional valence (i.e., "How positive or negative do you find this image?"), relevance for the statement (i.e., "How closely do you think this image represents the statement?"), and visual complexity (i.e., "How visually complex do you find this image?"). All questions involved 1-7 Likert scales, except where indicated otherwise. The final set of images contained 42 emotional images rated significantly more emotionally arousing than the 42 neutral images (Mean-Emotional images rating=4.94, SD=0.60; Mean-Neutral im-

ages rating=3.59, SD=0.53, p<0.001). To ensure both image types were categorized as having negative emotional valence, participants rated the question "How emotionally positive or negative do you find this image?" on a 9-point scale from 1="Extremely positive", to 9="Extremely negative", with the midpoint being marked at 5="Neutral"). Both the Emotional images (M=7.76, SD=0.41, p<0.001) and the Neutral images (M=6.59, SD=0.67, p<0.001) were rated significantly more negative than the Neutral midpoint of 5. Crucially, the emotional images did not significantly differ from the neutral images on how representative of the statement they were (Mean-Emotional images rating=5.25, SD=0.58, Mean-Neutral images rating=5.18, SD=0.48, p=0.57) and they also did not differ on how visually complex they were (Mean-Emotional images rating=4.24, SD=0.79, Mean-Neutral images rating=4.28, SD=0.74, p=0.82). These controls within the stimulus set serve to disambiguate a potential effect of emotional arousal on believability from confounding explanations. For instance, they ensure the emotional images do not add additional evidence in support of their corresponding statements compared to the neutral images.

Design and procedure. Participants were told they would participate in an experiment about people's opinions concerning information frequently encountered on the Internet and were directed to the survey on the Qualtrics platform. After completing the informed consent form, participants were warned about the graphic content of the experiment, and told they can end their participation at any point should they experience any discomfort. After the warning, participants were asked to complete a series of demographic measures. Then, they were instructed to rate a set of 42 statements (one on each page) by indicating the degree to which they believed each statement (i.e., "How accurate do you think this statement is?" from 1-Extremely inaccurate to 100-Extremely accurate). This phase acted as both the believability pre-test and the encoding phase. Next, participants were asked to answer a distracter question (e.g., "Please describe your favorite food"), that also

served as an attention check. During the statement-image association phase of the experiment, participants were shown the initial 42 statements again (also one on each page), this time alongside an image, and were asked two irrelevant questions that also served as attention checks (e.g., "Does this image capture an indoor or an outdoor scene?" and "Does this image capture any people?"). The pseudo-random pairing of images to statements was assigned such that of the 42 total statements, 6 of them were randomly assigned to the emotional image association condition, 18 statements were assigned to the neutral image association condition, and 18 statements were assigned to the blank image association condition. The decision to present fewer highly emotional images was made in order to minimize the psychological discomfort of the participants and to more closely match real-world conditions, where typically highly emotional events happen with reduced frequency, relative to neutral events (Walker, Skowronski, & Thompson, 2003). Moreover, varying the proportion of critical items has been found in a meta-analysis to have no impact on a similar effect, the illusory truth effect (Dechêne, Stahl, Hansen, & Wänke, 2010). These assignments were counterbalanced such that across the entire sample each statement was equally likely to be displayed with either an emotional, a neutral, or a blank image. After another distracter task (i.e., "Please provide a brief description of how your day is going so far"), participants were instructed to rate the believability of the initial 42 statements again (believability post-test phase). No images were presented in this phase. Finally, participants were debriefed and asked to review the false statements they were exposed to during this experiment to acknowledge their inaccuracy.

Results

The belief change score for every statement for each participant was computed by subtracting the belief score in the pre-test phase from the belief score in the post-test

phase, and then averaging these differences across all statements within each condition (Emotional, Neutral, and Blank). I note that in this study some of the statements were accurate, while some of the statements were inaccurate. To test whether the accurate and inaccurate statements elicited different results, I ran a Repeated Measures ANOVA with Accuracy and Image Type as independent variables and found no main effect of accuracy (p=0.462), and no interaction between image type and accuracy (p=0.115). Therefore, I decided to conduct the analyses combining the correct and incorrect statements. A Repeated Measures ANOVA with Item Type as the within-subjects independent variable and belief change as the dependent variable revealed a main effect of Item Type, F (2, 202)=3.491, p<0.032, η_p^2=0.033. Posthoc analyses revealed that statements in the Emotional condition showed a significant increase in believability (M=8.50, SD=15.92) compared to statements in both the Neutral condition (M=6.10, SD=12.50), t(101)=2.11, p<0.038, Cohen's d=0.17, CI[1.14, 4.66], and Blank condition (M=5.17, SD=14.55), t(101)=2.18, p<0.032, Cohen's d=0.22, CI[1.52, 6.35] (Figure 3.2A).

In the replication study, to test again whether there is a difference in the effect between the accurate and inaccurate statements, I conducted a Repeated Measures ANOVA with Accuracy and Image Type as independent variables, and found no main effect of accuracy (p=0.751), and no interaction between image type and accuracy (p=0.466). Thus, as before, I conducted the analyses combining the correct and incorrect statements. A Repeated Measures ANOVA with Item type as the within-subjects independent variable and belief change as the dependent variable revealed a statistically significant main effect of Item Type, F(2,402)=9.937, p<0.001, η_p^2=0.076. Posthoc analyses revealed that statements in the Emotional condition showed a significant increase in believability (M=7.57, SD=14.91) compared to statements in both the Neutral condition (M=5.21, SD=10.39), t(99)=2.38, Cohen's d=0.18, p<0.019, CI[0.39, 4.32], and Blank condition (M=4.01, SD=10.55),

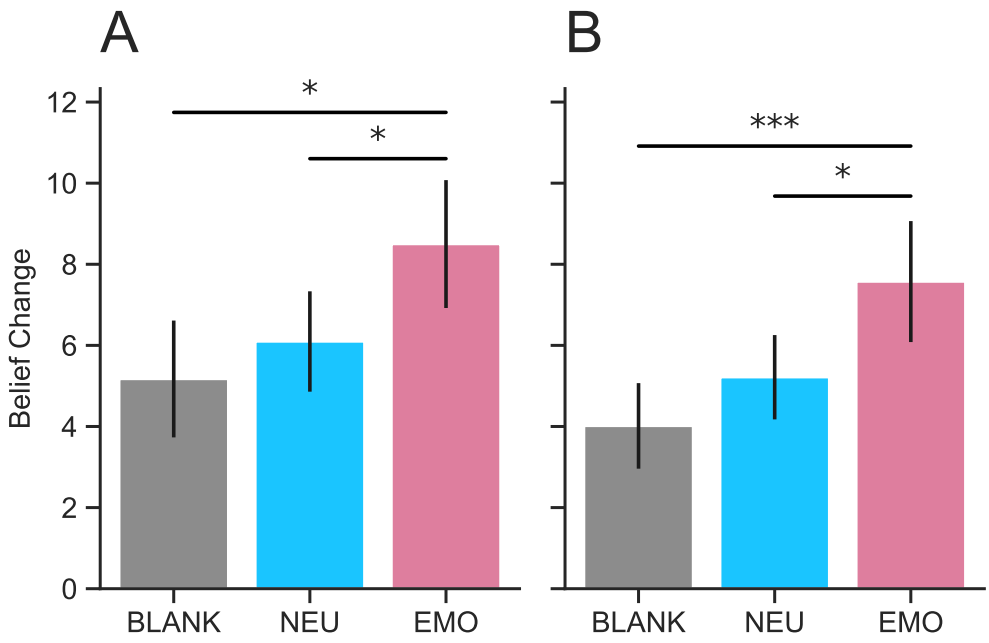


Figure 3.2: Belief change in Main Study (Panel A) and Replication Study (Panel B). Blank condition represented in grey, Neutral condition represented in blue, and Emotional condition represented in red. Error bars represent ±1 standard error of the mean.

t(99)=3.55, Cohen's d=0.28, p<0.001, CI[1.56, 5.55] (Figure 3.2B).

Discussion

In two independent online samples I find all statements increase in believability upon re-exposure. Furthermore, the results show that statements in the Emotional condition increased in believability even more, compared to those in both the Neutral and Blank conditions. This finding is consistent with the hypothesis that emotional images may enhance believability of associated statements by making these statements more memorable, according to literature on the illusory truth effect (Begg et al, 1992; Ozubko & Fugelsang, 2011). This finding also provides evidence against the competing hypothesis that emotional images may hinder statement believability, as predicted by a heuristic judgement mechanism or an attention mechanism.

The hypothesis that emotional images may enhance the statements' believability built on previous research showing that increasing the mnemonic accessibility of statements should result in an increase in believability (Vlasceanu & Coman, 2018). The paradigm I used in Study 2.1 cannot adequately address whether this is indeed the mechanism responsible for the effect I observed. This is because I did not measure whether the statements associated with emotional images were remembered better than the statements in the neutral and the blank conditions. To gather evidence regarding the underlying mechanism, I conducted another experiment, expecting that if mnemonic accessibility is indeed involved in producing the effect (1) the emotional images would be remembered better than neutral images and that (2) the statements associated with Emotional images would be remembered better than the statements paired with the neutral and blank images.

3.2 Study 2.2: Mechanism

Method

Participants. To detect moderate effect sizes of 0.35 for within-subject comparisons with 0.80 statistical power in two independent memory tests I calculated I need 134 participants. I collected data from a total of 147 participants, again expecting I would need to discard some participants due to attention check failures. However, I only had to discard one participant based on the same pre-established exclusion criteria described in Study 2.1. A potential reason for why most participants in this sample passed the attention checks is that in this experiment participants were instructed they could gain a bonus based on their performance, which may have led them to pay more attention to the experimental task, including the attention checks. I conducted the analyses on the final set of 146 participants (Mage=34.95, SDage=10.70; 40%

women), randomly assigned to either the memory for the statements condition (74 participants) or the memory for the images condition (72 participants). Participants were recruited on Amazon Mechanical Turk because the Turk-Prime platform I used in Study 2.1 has a no-bonus compensation policy. I compensated participants at the platform's standard rate. Additionally, I distributed a bonus of 5 cents for each correctly recalled statement or each correctly recognized image, depending on condition. All participants provided informed consent, and the study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. I used the same stimulus materials as in Experiment 2.1, with the addition of 84 new lure images used in the image recognition task. These new images were selected to be as similar as possible to the initial set of images used in the statement-image association phase, such that the two forced-choice recognition task does not reach ceiling recognition rates.

Design and procedure. The procedure was similar to the one used in Experiment 2.1, with one difference. In this experiment, instead of the post-test measure of the statements' believability, I tested participants' memory for the statements (N=74) or for the images associated with the statements (N=72). The statement memory task was an incidental, incentivized, free recall test, in which participants were asked to write as many of the initial statements as they could remember. They were compensated proportionally to their performance (i.e., 5 cents for each correct statement recalled). Each statement was coded as successfully remembered if the recall captured the gist of the original statement. For instance, if for the statement "Pneumonia is the prime cause of death in children" participants remembered "Pneumonia is the reason for most children's deaths", their recall was coded as accurate as it captured the gist of the original statement. Ten percent of the data were double coded for reliability (kappa = 0.87) and all disagreements were resolved through discussion between coders. The image memory task was also incidental and

incentivized, but in this case the task was a two forced-choice recognition test, also compensated proportionally to their performance (i.e., 5 cents for each correct image identification). The hits (i.e., images previously displayed during the statement-image association phase) appeared alongside lures (i.e., images as similar as possible to the hits), and participants were asked to choose the image they recognized from the statement-image presentation phase.

Results

To investigate whether the Emotional images are recognized to a larger degree than the Neutral images I first computed the average of the accuracy scores (proportion of correct hits) in the recognition task, for each image type, for each participant. I then ran a paired sample t-test comparing the proportion of correctly recognized Emotional images with the proportion of correctly recognized Neutral images, and found that, as expected, the Emotional images were remembered better (M=0.979, SD=0.062) than the Neutral images (M=0.959, SD=0.080), t(71)= 2.274, p<0.025, Cohen's d= 0.266, CI[-0.038, -0.002] (Figure 3.3A).

To investigate whether the statements in the Emotional condition were recalled more than statements in the Neutral or Blank condition, I computed the recall rates for each participant as the average of the recall proportion in each of the three conditions. A Repeated Measures ANOVA with Item type as the within-subjects independent variable and statement recall as the dependent variable confirmed the main effect of Item Type, F(2,146)=3.596, p<0.030, η_p^2=0.047. Posthoc analyses revealed that statements in the Emotional condition were remembered to a greater extent (M=0.121, SD=0.183) than statements in the Neutral condition (M=0.084, SD=0.128), t(73)= 2.39, p<0.018, Cohen's d=0.278, CI[0.066, 0.068] and marginally more than statements in the Blank condition (M=0.089, SD=0.105), t(73)= 1.792,

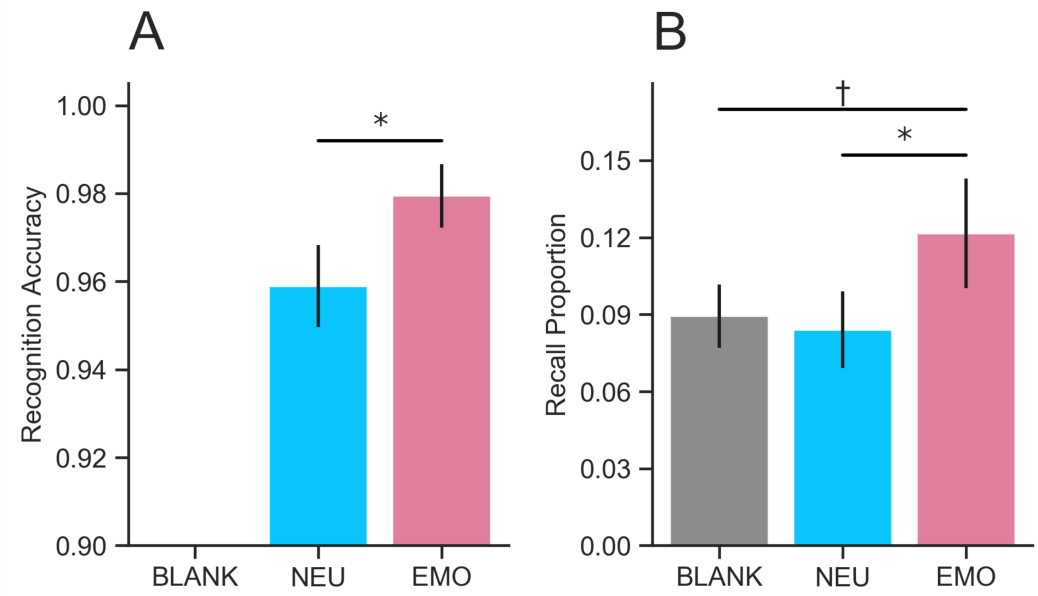


Figure 3.3: Panel A. Recognition accuracy for images (proportion correct) as a function of condition: Blank condition not represented here, since those are simply a blank screen, Neutral condition represented in blue, and Emotional condition represented in red. Error bars represent ±1 standard error of the mean. Panel B. Recall memory for statements (proportion recalled) as a function of condition: Blank condition represented in grey, Neutral condition represented in blue, and Emotional condition represented in red. Error bars represent ±1 standard error of the mean.

p=0.077, Cohen's d=0.214, CI[-0.003, 0.068] (Figure 3.3B).

Discussion

In this experiment I tested whether mnemonic accessibility could explain the effect I observed in Experiment 2.1, according to which emotional images increase believability. I found better recall for emotional images compared to neutral images, and better recall for statements associated with emotional images compared to statements associated with neutral images. This pattern of results is consistent with a memory enhancement mechanism underlying the emotion-induced belief amplification effect. However, further investigations should provide additional support to supplement the current evidence in favor of the memory mechanism.

Experiment 2.1 showed that emotional arousal increases statement believability, an effect likely driven by an increased mnemonic accessibility mechanism as supported by evidence presented in Experiment 2.2. While these findings seem robust in online samples, I wanted to know whether the effect replicates in Princeton University students, given the burgeoning literature advocating for replicating the effects reported in psychological studies in different contexts to explore their boundary conditions (Crump, McDonnell, & Gureckis, 2013; Amir & Sharon, 1990).

3.3 Study 2.3: Replication in a Princeton Sample

Method

Participants. I collected data from a total of 105 participants, recruited in the Princeton University student center. Participants were compensated at the standard Princeton University research participation rate. I performed statistical analyses on the entire set of 105 participants, given that no participants failed the attention checks (Mage=21.53, SDage=3.46; 59% women). All participants provided informed consent and the study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. I used the same stimuli as in Experiment 2.1.

Design and procedure. I used the same design and procedure as in Experiment 2.1, with the exception that participants were administered the survey on laptops set up in the Princeton student center, instead of online.

Results

To test whether there is a difference between the accurate and inaccurate statements in this college student sample, I conducted a Repeated Measures ANOVA with Ac-

curacy and Image Type as independent variables, and found no main effect of accuracy (p=0.214), and no interaction between image type and accuracy (p=0.395). I therefore combined the accurate and inaccurate statements for the main analysis. A Repeated Measures ANOVA with Condition as the within-subjects independent variable and belief change as the dependent variable revealed no main effect of Condition, F (2, 208)=0.91, p=0.404 (Fig.3a). Thus, the main effect did not replicate in this sample.

To explore potential reasons why the effect found in the online samples did not replicate in the college student sample, I conducted exploratory analyses in search for meaningful differences between the two samples that may have been responsible for the differences in effects. When looking at the duration of the experiment in the two samples, I noticed the college student participants spent significantly less time on the task (Mmin=22.2, SDmin=16.19) compared to both the Experiment 2.1 sample of participants (Mmin=33.96, SDmin=32.59), t(203)= 3.275, p<0.001 and the Experiment 2.1-replication sample of participants (Mmin=30.68, SDmin=33.33) t(205)= 2.324, p<0.021. This potentially indicates more shallow processing of the materials in the college student sample, which could have led to the disappearance of the effect. I derived more evidence for this conjecture though a complementary analysis I conducted on the combined data from Experiments 2.1 and its online replication. A median split by experiment duration revealed that the effect holds strongly in the half of the participants who spent longer time on the task (Mmin=46.94, SDmin=42.24): statements in the Emotional condition increased in believability (M=9.35, SD=16.08) compared to statements in both the Neutral condition (M=5.69, SD=11.01), t(100)=3.51, Cohen's d=0.266, p<0.001, and Blank condition (M=4.64, SD=13.16), t(100)=3.17, Cohen's d=0.32, p<0.002. For the participants who spent less amount of time on the task (Mmin=18.22, SDmin=3.64) according to the median split, there was no difference between statements in the

Emotional condition (M=6.72, SD=14.64) and statements in the Neutral condition (M=5.61, SD=12.01), t(100)=2.38, Cohen's d=0.08, p=0.293. A small difference was detected when comparing the Emotional and the Blank conditions (M=4.55, SD=12.30), t(100)=2.13, Cohen's d=0.16, p<0.035.

Discussion

The effect of emotional arousal on statement believability found in Experiment 2.1 and its replication (both online samples), did not replicate in a college student sample. Exploratory analyses hint at why this may have been the case: participants in the student sample completed the experiment in approximately half the time the online participants did, thus the image-statement associations may have not been solidified strongly enough for the effect to occur. Indeed, this hypothesis is supported by exploratory analyses in which I separated participants in the online samples into two equal groups according to experiment duration. I found that the group of participants who completed the task very quickly (at an equivalent pace to the college student sample) displayed a much weaker effect compared to those who completed the task more slowly, who displayed a very strong effect. This difference in time spent completing the task could have resulted in a superficial processing of information, such that there was insufficient binding between the image and its associated statement (Pacton & Perruchet, 2008). If so, an experimental manipulation aimed at facilitating the image-statement binding should result in increased believability for emotional images, while one aimed at diminishing image-statement binding should eliminate the emotion-induced believability amplification. To test this hypothesis, I manipulated the degree of binding between images and their associated statements.

3.4 Study 2.4: Binding Manipulation

Based on exploratory analyses I hypothesized that the difference in the effect I observed is due to an additional variable: image-statement binding. This hypothesis is supported by classic associative learning literature showing that two stimuli (e.g., a statement and an image) need to be processed together in order for an association between the two to be learned and subsequently remembered (Mackintosh, 1975; Hoffman & Sebald, 2005; Jimenez & Mendez, 1999). Therefore, I would expect image-statement binding to modulate the effect by determining whether or not a memory for the association forms. However, the literature on the Elaboration Likelihood Model of Persuasion (ELM; Petty & Cacioppo, 1983) suggests that under information processing conditions that are too high, subtle cues (e.g., music, humor, visuals) that affect persuasion under lower processing conditions, are no longer successful persuasion tools. Therefore, I would expect very high image-statement processing conditions to also modulate the effect.

To test these predictions, I conducted another study in which participants were randomly assigned to a Low Binding Condition, a Moderate Binding Condition, and a High Binding Condition. In the Low Binding Condition, during the image-statement association phase participants were asked to indicate the image quality of each photograph on a scale from 0="very bad quality" to 100="very good quality". To ensure low processing levels, participants were only allowed to spend 3 seconds rating each image. In the Moderate Binding Condition participants were asked to indicate whether each image captured an indoor or an outdoor scene, and whether each image captured any people. Participants were allowed to spend as long as 15 seconds on each image. Finally, in the High Binding Condition participants were instructed to rate how representative each image is of the statement and how much evidence for the statement each image provides. For each image, participants were allowed to spend as long as 30 seconds and as little as 15 seconds. Like before, in

each condition, to assess the degree to which believability changed as a function of the emotional arousal of the image and of condition, I collected believability ratings of all statements before and after the manipulation. I predicted the emotion-induced belief amplification effect to hold in the Moderate Binding Condition, but not in the Low or High Binding Conditions.

Method

Participants. I collected data from a total of 329 Princeton University students. Participants were compensated at the standard Princeton University research participation rate. I performed statistical analyses on the entire set of participants, given that no participants failed the attention checks (Mage=19.39, SDage=1.31; 65% women). All participants provided informed consent and the study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. I used the same stimuli as in Experiment 2.1.

Design and procedure. I used the same design and procedure as in Experiment 2.1, with the exception that I conducted the experiment in the lab, instead of online, and with the addition of the 3 between-subject conditions (Low, Moderate, and High Binding) during the image-statement binding phase. Participants were randomly assigned to one of the three conditions, which amounted to 109 participants in the Low, 111 participants in the Moderate, and 109 participants in the High Condition.

Results

I conducted a Repeated-Measures ANOVA with Item Type (Emotional, Neutral, and Blank) as the within-subjects variable, and Condition (Low, Moderate, and High Binding) as the between-subject variable. Belief change was the dependent variable. I found a significant main effect of Condition, $F(2, 326)=6.716$, $p<0.001$, $\eta_p^2=0.04$ but not of Type $F(2, 652)=2.154$, $p=0.11$, $\eta_p^2=0.007$, or interaction, $F(4, 652)=46.193$,

p=0.15, η_p^2=0.01. Post-hoc analyses revealed that, as hypothesized, there was no difference between the emotion (M=0.40, SD=8.19) and neutral (M=1.44, SD=5.49, p=0.24) or blank items (M=0.24, SD=4.78, p=0.81) in the Low Binding Condition, and also no difference between the emotion items (M=0.97, SD=8.37) and neutral (M=-0.19, SD=5.41, p=0.15) or blank items (M=0.17, SD=4.40, p=0.29) in the High Binding Condition. In the Moderate Binding however, I did find that emotional items (M=3.50, SD=8.52) increased in believability significantly more than blank items (M=1.95, SD=6.46, p=0.04) but not significantly more than neutral items (M=2.61, SD=7.14, p=0.23) (Figure 3.4).

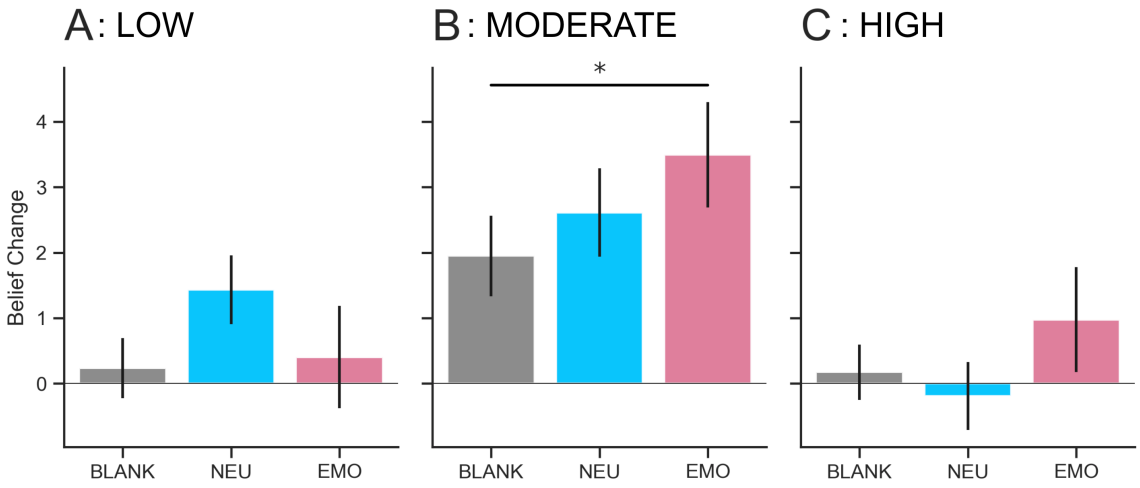


Figure 3.4: Belief change in the Low (Panel A), Moderate (Panel B), and High (Panel C) Condition. Blank condition represented in grey, Neutral condition represented in blue, and Emotional condition represented in red. Error bars represent ±1 standard error of the mean.

Discussion

The emotion-induced belief amplification effect found in Experiment 2.1 in two independent online samples but not a Princeton student sample in Experiment 2.3, failed again to replicate in another Princeton student sample in Experiment 2.4. Here, I tested whether the degree (low, moderate, high) of image-statement binding may

modulate the effect by experimentally manipulating this potentially modulatory variable. However, I failed to find the initially observed effect in any of these three conditions.

A potential explanation why the effect of emotional arousal on belief change is only observable in online samples but not in Princeton student samples may be that Princeton students process information more rationally, in a way that makes them immune to such subtle biasing effects. This speculative explanation for the boundary case of the emotion-induced amplification effect needs empirical evidence. For instance attempting to replicate the effect at another higher education institution comparable to Princeton University versus in a nationally representative sample could shed some light on this puzzle.

General Discussion

In two online studies (main study and preregistered replication) I found that emotional arousal increases statement believability (Study 2.1). This finding is consistent with previous research showing that increasing mnemonic accessibility increases believability (Vlasceanu & Coman, 2018; Vlasceanu, Morais, Duker & Coman, 2020). In another online study, I found evidence supporting a memory mechanism being responsible for this effect: the statements associated with emotionally arousing images became more memorable compared to the statements associated with neutral or no images (Study 2.2). In an attempt to extend the investigation to another context, I found a boundary condition of the uncovered effect, which failed to replicate in a sample of Princeton students (Study 2.3). When testing a binding explanation for the replication failure, again in a sample of Princeton students, I failed again to replicate the effect (Study 2.4).

I speculate the difference in effects between the online samples and Princeton student samples may be due to a more fundamental difference between the two pop-

ulations in the way they are used to process and assimilate information, and suggest this interpretation be empirically investigated in future work. This is however not the first effect in psychological research to render different outcomes in online versus college student samples. Previous work has shown that online participants (Amazon Mechanical Turk workers) pay more attention during psychology experiments compared to undergraduate students in the subject pools of North American universities (Hauser & Schwarz, 2016; Behrend et al, 2011; Goodman et al, 2013). For example, Klein and colleagues (2014) found this difference in a study showing that online participants passed significantly more attention checks than undergraduate students at 15 out of 19 colleges, and only 3 college sites were comparable to the online samples in attention check passing rates. In addition to paying more attention to the experiments they participate in, another benefit of online samples is that participants (e.g., MTurk workers), have been found to be more representative of the US population than college students (Berinsky et al, 2012).

Moreover, there might be other factors that could also impact the manifestation of this effect. For instance, in the current studies, I purposefully employed a highly controlled paradigm in which participants received no details about the source of the information. In real-world contexts, the source of information was found to reliably impact believability (Pornpitakpan, 2004; Jennings, 2018). In mass-mediated communication, for example, news stations are more or less aligned with the ideology of their viewers, as viewers spend more time-consuming attitude-consistent media (Knobloch-Westerwick & Meng, 2009). MSNBC viewers might be more convinced of the accuracy of a statement if associated with an emotional image, as would Fox viewers be upon exposure to emotionally charged images. Such ideological commitments could constitute factors that would amplify the effect observed in the current investigation. This might be due to the normative component implied by the ideological nature of the information source (Brady et al, 2017). Future research could

programmatically investigate variables such as ideological commitment, attitudinal extremity, or perhaps vary the alignment between a participant's ideology and that of the source of information.

At the same time, when it comes to beliefs, individuals oftentimes communicate with one another, shaping each other's beliefs in the process. Future research building on the effect investigated herein could explore how communicative contexts impact its manifestation (Vlasceanu, Enz, & Coman, 2018; Vlasceanu, Morais, Duker & Coman, 2020). How does the effect propagate in a social network after a central speaker broadcasts messages to individuals? It is often the case that factual statements are broadcasted by newscasters while images or videos supporting the conveyed information are concurrently displayed. Does the belief amplification effect propagate from the original source? If so, how deep in the network does it propagate? Fowler and Christakis (2010) show that even though there are 6 degrees of separation in real-world networks, influence only spreads three degrees away from the originating source. Empirical findings from the literature on the propagation of memory effects in social networks also support this conclusion (Coman et al, 2016). I do not know whether other cognitive effects, such as the one investigated here, have similar characteristics in communicative settings.

Finally, the current research has broad, direct implications for understanding how images could affect both accurate and inaccurate beliefs, which, in my study were equally subject to the belief amplification effect. This is important as this effect could be instrumental to the investigation of strategies aimed at decreasing the believability of inaccurate information in the population. The urgency of research uncovering misinformation reduction tactics has been acknowledged by many (Pennycook et al, 2018; Vosoughi, Roy, & Aral, 2018; Vlasceanu & Coman, 2018), and is of immediate importance given the potential for fake news to spread at increasingly fast speeds on social media platforms. For example, a recent Twitter study reported that false

news diffuse "farther, faster, deeper, and more broadly than the truth" (Vosoughi, Roy, & Aral, 2018). Understanding which cognitive processes are successful in amplifying trust in accurate information is a crucial first step that can inform future investigations into techniques aimed at reducing misinformation spread in vulnerable communities. For instance, future such work could consider pairing classical strategies of debunking of inaccurate beliefs (Lewandowsky et al, 2012) with associating negatively arousing images with accurate information. This pairing of tactics could prove more effective than a simple debunking intervention. If proven effective, such interventions will add to the tools that policy makers employ in the battle against one of the top threats faced by the world today, the misinformation epidemic (Farkas & Schou, 2019; Lewandowsky et al, 2012).

In Chapters 2 and 3 I showed that mnemonic accessibility and emotional arousal can be manipulated to change belief endorsement. These effects, in the proposed framework, occur at the belief level. What happens when manipulations occur at the level of evidence? The prediction is that changes at the evidence level should result in changes at the belief level.

Chapter 4

Predictions and Beliefs

4.1 Study 3.1: Prediction Errors Trigger Belief Change

This Chapter is based on the paper "The Effect of Prediction Error on Belief Update Across the Political Spectrum" in press at the Journal *Psychological Science*. The co-authors of this publication are Michael J. Morais and Alin Coman.

Abstract

Making predictions is an adaptive feature of the cognitive system, as prediction errors are used to adjust the knowledge they stemmed from. Here, we investigate the effect of prediction errors on belief update in an ideological context. In Study 3.1, 704 Cloud Research participants first evaluated a set of beliefs, then either made predictions about evidence associated with the beliefs and received feedback or were just presented with the evidence. Finally, they re-evaluated the initial beliefs. Study 3.2, which involved a US census matched sample of 1073 Cloud Research participants, replicated Study 3.1. We find that the size of the prediction errors linearly predicts belief update and that making large errors leads to more belief update than

not engaging in prediction. Importantly, the effects hold for both Democrats and Republicans across all belief types (Democratic, Republican, Neutral). We discuss these findings in the context of the misinformation epidemic.

Introduction

"The political ignorance of the American voter is one of the best documented data in political science" (Bartels, 1996). A burgeoning literature across the social sciences has been dedicated to developing strategies to address this notorious limitation. Kuklinski and colleagues (2000) identify two conditions that are needed to assuage this problem: increased access to objective facts and their incorporation in individuals' mental models. The first condition is difficult to satisfy given that nearly half of Americans get their news from Facebook (Pew Research Center, 2017), a social media platform known for providing access to a vast volume of misinformation (Shu et al., 2017). However, even if such organizations successfully implement strategies to diminish misinformation, a more daunting challenge arises: persuading people to incorporate these facts into their belief systems. Findings from social psychology hint that new facts are easily dismissed if they increase cognitive dissonance (Festinger & Carlsmith, 1959), reduce coherence among already held beliefs (Lord, Ross, Lepper, 1979), or counter one's political allegiance (Nyhan & Reifler, 2010). Here, I am interested in exploring cognitive processes that could facilitate the incorporation of facts into people's belief systems.

A central feature of beliefs – defined as statements that individuals hold to be true (Schwitzgebel, 2010) - is their dynamic nature, as beliefs are susceptible to change (Bendixen, 2002). Prior work has identified several strategies that proved effective at changing beliefs, such as using fictional narratives (Wheeler, Green, Brock, 1999), manipulating memory accessibility (Vlasceanu & Coman, 2018), associating beliefs with emotionally arousing images (Vlasceanu, Goebel, Coman, 2020), or nudging ac-

curacy goals (Pennycook et al., 2020). Here, I propose that one powerful strategy to facilitate belief change might involve updating mental models through prediction errors. This conjecture builds on the seminal finding that learning is proportional to prediction error (PE), where PE is the difference between the prediction one makes about a state of the world and the actual outcome (Rescorla & Wagner, 1972). Since expectations are based on the agent's model of the world, when predictions are validated, they reinforce the model of the world they stemmed from, and when they are invalidated, the model gets updated accordingly (Den Ouden, Kok, & De Lange, 2012). Generating predictions is, arguably, a ubiquitous process implemented by the cognitive system that has adaptive consequences for the organism (Bar, 2009). In the current investigation, I am interested in whether prediction errors may have a similar effect on belief change, and whether this effect might be modulated by motivational factors that involve political ideology. Is the influence of prediction errors on belief update a general process, or are there partisan biases in the way prediction errors impact beliefs? On the one hand, in support of prediction errors as a general mechanism, they have been found to impact a wide range of cognitive processes, including perception (deLange, Heilbron, Kok, 2018), action (Bestmann et al., 2008), memory (Erickson & Desimone, 1999), language (Kutas & Hillyard, 1980), cognitive control (Alexander & Brown, 2011), and decision-making (Greve et al., 2017). On the other hand, past research shows that motivations to reach particular conclusions affect information processing (Nyhan & Reifler, 2010). This suggests that there might be meaningful motivational differences between liberals and conservatives that could affect the relation between prediction errors and belief update (Ditto et al., 2019; Haidt, Graham, & Joseph, 2009). One way in which ideological biases could influence the belief updating process may involve a reduced susceptibility to changing ideologically consistent beliefs as a function of prediction errors (as opposed to ideologically inconsistent or neutral beliefs). That is, people might be entrenched with respect

to their party's beliefs, but flexibly updating other beliefs (Toner, Leary, Asher, & Jongman-Sereno, 2013). Another way in which ideological biases might impact belief update could involve a differentiation between liberals and conservatives, such that conservatives might be more resistant to change than liberals, as has been already shown (Jost, Glaser, Kruglanski, & Sulloway, 2003; White et al., 2020). This would predict that conservatives might be less likely than liberals to change their beliefs according to the prediction errors they make, regardless of the ideological nature of those beliefs. Yet another possibility is that belief updating could be dynamically dependent on environmental factors involving uncertainty and political identity threat (Haas & Cunningham, 2014). The more uncertain and threatened, the more resistant to changing one's beliefs.

I explore the relationship between prediction error and belief updating in an experiment that exposes participants to belief-associated statistical evidence by passive viewing versus active predicting. I hypothesized a positive linear relationship between prediction error size and belief update. I also hypothesized that making large prediction errors would lead to more evidence incorporation and belief change than not engaging in prediction. I did not have a priori hypotheses regarding how participant and belief ideology would interact with the effect of PE on belief update.

Method

Open science practices. I preregistered the study's experimental design and hypotheses on an open science platform (https://aspredicted.org/blind.php?x=h4fm5n). In addition, the stimuli, pilot study results, and the data for the main study can be found on the study's open science framework page (https://osf.io/aur2t). The data analysis (in Python) can be accessed here: https://github.com/mvlasceanu/PredictionBelief.

Participants. To detect a moderate effect size of 0.3 for two between-subject

comparisons with 0.80 statistical power I estimated a sample of 704 participants. Participants were recruited on Cloud Research, an Internet-based research platform similar to Amazon Mechanical Turk (MTurk) but with more intensive participant pool checks, and were compensated at the platform's standard rate (Litman, Robinson, & Abberbock, 2017). In total I recruited 945 participants, of which 241 were excluded from the analysis based on preregistered criteria (i.e., attention checks, political party affiliation). I stopped data collection as soon as I reached the pre-registered sample size of 704 valid participants (Mage=50.32, SDage=16.51; 67.7% women). Of these, the 352 participants self-identified as Democrats were randomly assigned to the Experimental Condition (N=176) or the Control Condition (N=176), and the 352 self-identified as Republicans were also evenly split between the two conditions. The study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. I undertook preliminary studies to develop a set of 36 statements. These statements were equally split into 12 Neutral statements (e.g., "Shark attack rates are similar for men and women."), 12 Democratic statements (e.g., "The US has loose gun laws."), and 12 Republican statements (e.g., "A large proportion of immigrants in the US is not in the workforce."). The 36-statement set was selected from a larger initial set of 48 statements that I pretested on an independent sample of Cloud Research participants (N=50; Mage=41.94, SDage=15.83; 62% women). In the pilot study I first measured the believability of each statement with the question "How accurate do you think this statement is?" on a 0-100-point scale ranging from "Extremely inaccurate" to "Extremely accurate". I selected the final set of 36 statements such that each Neutral statement was equally believed by the Democratic and Republican participants, but each Democratic statement was believed significantly more by Democrats than by Republicans, and each Republican statement was believed more by Republicans than by Democrats. Overall the Neutral statements were

equally endorsed by Democrats (M=62.83, SD=9.35) and Republicans (M=59.58, SD=15.76), p=0.546; the Democratic statements were endorsed significantly more by Democrats (M=70.78, SD=7.89) than by Republicans (M=49.68, SD=16.15), p<0.001; and the Republican statements were endorsed significantly more by Republicans (M=61.74, SD=12.37) than by Democrats (M=48.36, SD=10.85), p<0.01. I also developed a set of 36 facts that provide evidence either in support or against the 36 statements. For example, for the statement "Very few Americans identify are vegetarian", the evidence in support was "5% of Americans are vegetarian", and for the statement "Many American adults exercise on a daily basis", the evidence against was "5% of Americans participate in 30 minutes of physical activity every day". These factual statistics were selected from a larger set of 48 accurate facts I found in scientific papers or official polls, and pretested on the same sample of participants as the pilot study, to match on how strongly each piece of evidence would influence each associated statement (e.g., "How likely is this piece of evidence to influence your support for this statement?" on a 1="Not at all" to 5="A great deal" scale). The 36 facts were selected such that for Democrats, the Neutral (M=3.05, SD=0.54), Democratic (M=3.24, SD=0.54), and Republican facts (M=3.05, SD=0.25) did not significantly differ on how strongly Democrats thought they influence the statements; likewise, for Republicans the Neutral (M=3.56, SD=0.41), Democratic (M=3.57, SD=0.21), and Republican facts (M=3.43, SD=0.42) did not significantly differ from each other on the evidence strength dimension.

In addition, a set of 36 scale-based estimation questions were constructed to be used as part of the evaluation phase. These questions were constructed by rephrasing the facts constructed as evidence. For example, for the fact "In the US, 3 of the 50 states require a permit to purchase a rifle", the corresponding question was "How many of the 50 states require a permit to purchase a rifle?". Each question had 12 potential answers, linearly increasing on the 12-item scale, one of which was the

correct one. Across the 36 questions, the correct answer had an equal chance of being in any of the 12 scale positions from 1 to 12. This prevented forming probability estimates for the most likely positions on the scale to contain the correct answer.

I measured resistance to change, a construct that has been found to differentiate between liberals and conservatives (Jost et al, 2003). The measure was adapted from the Willingness to Compromise Scale (Wee, 2013), and was computed as the average response to the 3-item scale ("I would stick to my beliefs even when others might think that they are not reasonable.", "Reality constraints should not stand in the way of one's beliefs.", and "Once I believe in something, no piece of evidence would change my mind."). All questions were rated on a scale from 1-"Strongly disagree" to 5-"Strongly agree". Thus, higher scores indicate more resistance to change and lower scores indicate less resistance to change. Finally, I measured participants' strength of identification with their selected political party, with the question "How strongly do you identify with the party you just selected?" on a scale from 1-"Not at all" to 5-"A great deal", as well as their support for the current president, with the question "How would you qualify president Donald Trump's performance in office for the past 3 years?" from 1-"Awful" to 7-"Excellent". I used both of these questions as measures of political polarity. I note that 92.18% of the sample indicated they are registered to vote for the party they identified with.

Design and procedure. The data for this study was collected between October 10, 2019 and October 14, 2019. Participants were told they would participate in an experiment about how people evaluate information encountered on the Internet and were directed to the survey on the Qualtrics platform. After completing the informed consent form, participants were directed to the pre-test phase, where they were instructed to answer questions about information encountered on the Internet, which meant rating a set of 36 statements (one on each page) by indicating the degree to which they believed each statement was accurate (i.e., "How accurate do you think

this statement is?" from 1-Extremely inaccurate to 100-Extremely accurate). Then, in the evidence phase, participants were randomly assigned to one of two between-subject conditions: Prediction Condition and Control Condition. Participants in the Control Condition were shown a series of 36 facts that provided direct evidence either in favor or against the set of 36 beliefs. Instead of simply being exposed to the facts, participants in the Prediction Condition were asked to predict the correct answers to questions equivalent in content to the 36 facts used in the Control Conditions. After choosing an answer, participants were immediately given feedback (i.e., the correct answer). In both conditions, the evidence was presented one on each page and in a random order. Then, in a post-test believability phase, participants were instructed to rate the believability of the initial 36 statements again. Finally, participants were asked to complete the Resistance to Change Scale, a series of demographic measures including their strength of party affiliation and support for President Trump, after which they were debriefed.

Analysis and coding. I operationalize rational belief update as a belief change from pre-test to post-test in the direction corresponding to incorporating the available evidence. Critically, whether this update corresponds to increasing or decreasing beliefs depends on (i) counterbalanced features of the stimuli and (ii) observed features of participants' predictions. The first counterbalanced feature of the stimuli is that one half of the presented evidence supports half of the beliefs – in this case the rational update is to increase in believability from pre-test to post-test. The other half of the evidence refutes the other half of the beliefs – in this case the rational update is to decrease in believability from pre-test to post-test. Using this setup, I ensure that participants cannot trivially infer that "correct" updates must necessarily occur in one direction. This is the only variable necessary to compute rational belief update in the Control Condition. For example, for the belief "Very few Americans identify as vegetarian" with the corresponding supporting piece of evidence "5% of Americans

identify as vegetarian", the rational update is to increase believability from pre-test to post-test. Conversely, for the belief "Many American adults exercise on a daily basis" with the corresponding piece of evidence arguing against it "5% of Americans exercise on a daily basis", the rational update is to decrease believability from pre-test to post-test.

Rational belief update in the Prediction Condition has two additional variables that determine in which direction update is rational for each belief: the magnitude of the correct answer (high or low on the scale) and the sign of the prediction error (positive or negative relative to the correct answer). The magnitude of the correct answer was counterbalanced across the stimuli such that the "surprise" was possible in both directions. For example, the answer to the question "How many child deaths worldwide is pneumonia responsible for every year?" is of high magnitude (i.e., 1 million deaths; the alternative answers of lower magnitudes are situated below the correct answer on the scale), whereas the answer to the question "What percentage of people who collapse on the street fully recover from receiving CPR?" is of low magnitude (i.e., 2%; the alternative answers of higher magnitudes are situated above the correct answer on the scale). The last variable needed to determine the direction of rational update is the prediction error sign. Prediction error (PE) is defined as the difference between the selected answer by the participant and the correct answer for that question. If a participant selects an answer of higher magnitude than the correct answer then their PE for that item will have a positive sign, whereas if they select a lower answer than the correct one their PE for that item will have a negative sign. For example, if on the question "How many Americans identify as vegetarian?" the participant selects 20%, given that the correct answer is 5% their PE will be positive. If, however, the participant selects 2%, their PE will be negative. These three variables (evidence in support/evidence against, high correct answer/low correct answer, and positive PE/negative PE) determine the direction of rational update in

the Prediction Condition. An example of a trial in which the evidence is in support and the correct answer is high is the belief "Pneumonia is dangerous for children" with the associated evidence "Worldwide, 1 million children die of pneumonia each year." In this case, if the participant chooses any value lower than 1 million (PE is negative) they will realize the number of deaths is higher than they thought, which should increase their belief in the danger of pneumonia, so the rational update is to increase the belief that "pneumonia is dangerous for children". If however, the participant chooses a value higher than 1 million (PE is positive) they will realize number of deaths is lower than they thought, which should decrease their belief in the danger of pneumonia, so the rational update is to decrease the belief that "pneumonia is dangerous for children". Analogous logic applies to each of the other eight combinations of the variables determining rational update in the Prediction Condition.

To calculate prediction error size, I took the absolute value of the prediction errors to obtain an index of prediction error size (from 0 to 11) with higher scores indicating larger errors and lower scores indicating lower errors. I then binned the 11 prediction error sizes into 3 bins: no PE (PE=0), small PE (PE =1-5), and large PE (PE=6-11). I decided to bin the data in this manner given the higher degree of interpretability of the binned PE sizes, as well as increased statistical power. For the sake of transparency, I will present both the binned and the unbinned results, although they are equivalent.

Results

Does PE linearly predict rational update? To test the first hypothesis, I analyzed the rational belief update as a function of the prediction errors in the Prediction Condition. I fitted a linear regression of Prediction Error size against Rational Belief Update and found that, as hypothesized, Prediction Error size linearly and posi-

tively predicts Rational Belief Update (β=2.87, SE=0.11, t(3699)=24.26, R^2=0.137, p<0.001; Figure 4.1A). I verified and expanded this analysis with a more rigorous linear mixed model with rational belief update as the dependent variable, prediction error size and belief at pre-test as fixed effects, as well as by-participant random intercepts and by-item random intercepts. I included belief at pre-test as a fixed effect in the model to control for the effect of the baseline level of belief on the degree of belief update, while observing the independent effect of prediction error size on rational belief update. Again, I found that prediction error size linearly predicts rational belief update (β=1.58, SE=0.09, t(12490)=16.68, p<0.001).

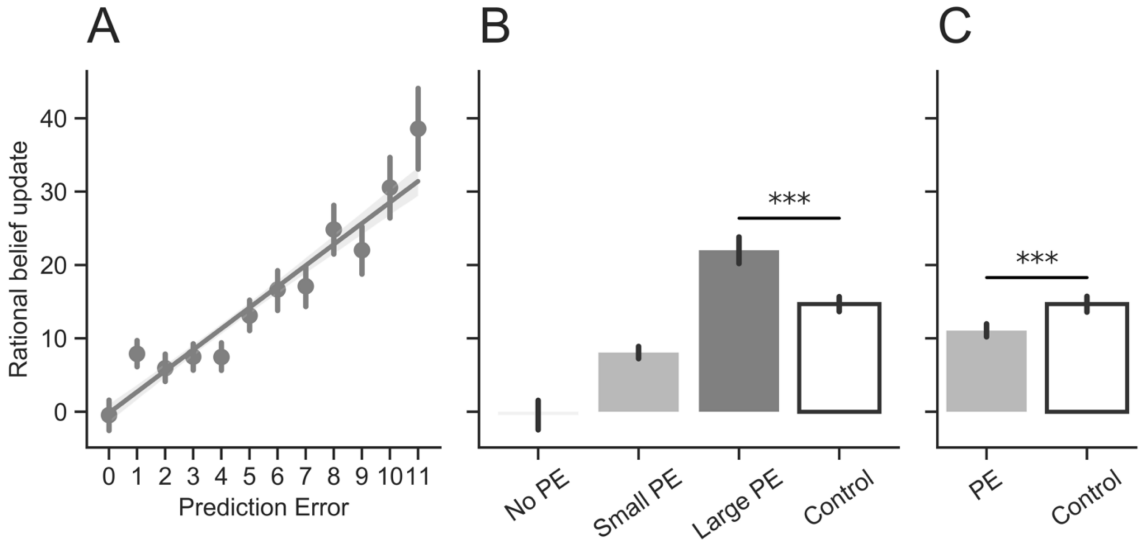


Figure 4.1: Panel A: Rational belief update (post-test raw scores minus pre-test raw scores) as a function of prediction error size. Error bars represent 95% confidence intervals. Panel B: Rational belief update as a function of prediction error size (absolute value of prediction error, binned by no error, small error, and large error). The Prediction Condition is displayed in grey bars (increasing in color intensity as the prediction error size increases), and the Control Condition is displayed in white. Panel C: Rational belief update in the Prediction Condition (grey) and the Control Condition (white). Error bars represent ±1 standard error of the mean.

Do large PEs lead to more rational update than the Control Condition? I ran an independent sample t-test comparing rational belief update in the Prediction and Control Conditions and found that items in the Prediction Condition were rationally

updated (M=11.05, SD=8.23) to a lower degree than items in the Control Condition (M=14.67, SD=10.28), t(670)=-5.14, p<0.001, Cohen's d=0.38, CI [-4.99, -2.23] (Figure 4.1C). This indicates that, on average, beliefs were changed more when participants were provided with passive evidence than when they were asked to make predictions and then provided with the correct answer. To further explore this pattern, I assessed whether this conclusion applies independently of the size of the prediction error. The preregistered hypothesis was that large prediction errors would lead to a larger belief update compared to the control condition. We, thus, compared the degree of rational belief update in the large PE bin of the Prediction Condition with the degree of rational belief update in the Control Condition. An independent sample t-test established that, consistent with the preregistered hypothesis, items in the large PE bin were rationally updated (M=22.00, SD=17.40) to a higher degree than items in the Control Condition (M=14.67, SD=10.28), t(569)=6.802, p<0.001, Cohen's d=0.51, CI [5.21, 9.44] (Figure 4.1B). Of note, the proportion of items that ended up in the 3 bins of the prediction condition was: 9.45% in No PE, 61.48% in Small PE, and 29.05% in Large PE.

Is there a partisan bias in how PE linearly predicts rational belief update? First, to investigate whether there is a difference between Republicans and Democrats in how they update their beliefs, I turned to the Prediction Condition. I ran a linear mixed model with rational belief update as the dependent variable, prediction error size, participant ideology, and belief at pre-test as fixed effects, and by-participant and by-item random intercepts. There was no interaction between PE and participant ideology (p=0.18), suggesting that overall, Republicans and Democrats do not update their beliefs differently as a function of prediction errors. This result was surprising for two reasons. First, that participants' self-reported resistance to change was found to significantly moderate the effect of PE on update (i.e., participants who self-reported as more resistant to change were less likely to update beliefs as a function of prediction

errors [β=-0.22, SE=0.1, t(12550)=-2.12, p=0.03]). Second, Republican participants self-reported to be significantly more resistant to change (M=3.57, SD=0.78) than Democrats (M=3.31, SD=0.83), t(700)=4.179, p<0.001, Cohen's d=0.315, CI [0.13, 0.37].

Furthermore, I tested for a potential ideological modulation of the effect of PE on rational update, this time while also taking into account the item ideology. I ran a linear mixed model testing the interaction of prediction error size with participant ideology (Democratic and Republican) and item ideology (Democratic, Republican, Neutral). The dependent variable was again, rational belief update. I fitted prediction error size, participant ideology, item ideology, and belief at pre-test as fixed effects, and included by-participant random intercepts and by-item random intercepts. The results show that prediction error size linearly predicts rational belief update in all of the six ideological conditions crossing participant ideology and item ideology (i.e., Democrats on Neutral, Democratic, and Republican beliefs, as well as Republicans on Neutral, Democratic, and Republican beliefs; summarized in Table 4.1, plotted in Figure 4.2).

		β	*SE*	*df*	*t*	*p*
(intercept)		4.59	1.416	45.24	3.24	=0.002
Belief at pre-test		16.8	0.255	12510	66.00	<0.001
Democratic Participants	Neutral Items	2.12	0.188	6904	11.24	<0.001
Democratic Participants	Democratic Items	1.80	0.182	8156	9.86	<0.001
Democratic Participants	Republican Items	1.29	0.181	8693	7.13	<0.001
Republican Participants	Neutral Items	2.00	0.185	6870	10.76	<0.001
Republican Participants	Democratic Items	1.08	0.180	7964	6.00	<0.001
Republican Participants	Republican Items	1.29	0.185	8672	6.96	<0.001

Table 4.1: Rational belief update predicted by a linear mixed model testing the interaction of prediction error size with participant ideology (Democratic and Republican) and item ideology (Democratic, Republican, Neutral), while controlling for the belief at pre-test.

I established that prediction errors linearly predict rational belief update in all ideological subsamples. However, a partisan bias could still exist in how strongly

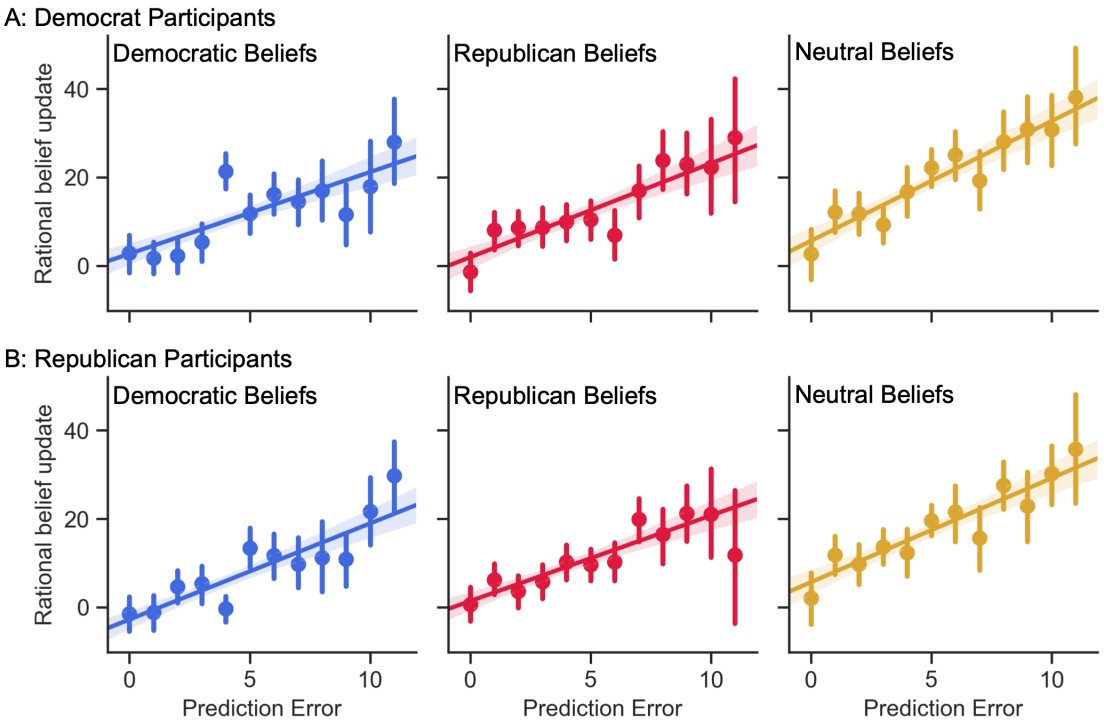


Figure 4.2: Rational belief update of Democrats (Panel A) and Republicans (Panel B) as a function of prediction error size split by belief ideology. Democratic beliefs are represented in blue, Republican beliefs are represented in red, and Neutral beliefs are represented in yellow. Error bars represent 95% confidence intervals.

this effect manifests in these ideological subsamples. To test this possibility, I ran a linear mixed model on the Prediction Condition, with rational belief update as the dependent variable, prediction error size, belief at pre-test, item ideology, and participant ideology as fixed effects, as well as by-participant random intercepts and by-item random intercepts. I did not find a significant 3-way interaction between prediction error size, item ideology (Democratic vs. Republican), and participant ideology (Democratic vs. Republican), suggesting that rational update as a function of prediction error size is a similar process regardless of participant ideology and item ideology. Finally, when considering the measures of political polarization, neither political party affiliation strength (β=0.16, SE=0.2, t(12474)=0.809, p=0.4186) nor support for President Trump (β=-0.16, SE=0.15, t(12490)=-1.055, p=0.2913) significantly moderated the effect of PE on rational update.

Is there a partisan bias in how beliefs are rationally updated in the large PE Condition compared to the Control Condition? To further investigate a potential ideological modulation of the uncovered effect of PE on rational update I also tested whether rational update is higher in the large PE bin of the Prediction Condition compared to the Control Condition in each of the six subsamples of the data (i.e., Democrats on Democratic, Republican, and Neutral items, and Republicans on Democratic, Republican, and Neutral items; summarized in Table 4.2, and displayed in Figure 4.3). I found that all of the independent sample t-tests comparing the large PE bin of the Prediction Condition to the Control Condition in each of the six ideological subsamples of the data were statistically significant (statistics reported in Table 4.2, plotted in Figure 4.3).

Participant	*Item*	*Condition*	*M*	*SD*	*df*	*t*	*d*	*CI*	*p*
Democrats	Neutral	Large PE	28.59	38.34	258	4.45	0.47	[5.62, 14.51]	<0.001
		Control	21.64	35.03					
Democrats	Democratic	Large PE	16.01	31.89	278	1.98	0.21	[0.04, 7.00]	=0.048
		Control	12.62	31.78					
Democrats	Republican	Large PE	19.01	36.94	240	4.36	0.46	[4.96, 13.07]	<0.001
		Control	11.01	34.06					
Republicans	Neutral	Large PE	26.73	38.47	275	3.81	0.41	[3.96, 12.36]	<0.001
		Control	20.17	35.01					
Republicans	Democratic	Large PE	17.55	34.01	274	4.31	0.45	[4.51, 12.11]	<0.001
		Control	11.88	35.39					
Republicans	Republican	Large PE	16.22	32.86	265	2.88	0.31	[1.70, 9.05]	=0.004
		Control	10.69	32.35					

Table 4.2: Rational belief update difference between large PE and Control, in all participant ideologies (Democratic and Republican) and item ideologies (Democratic, Republican, Neutral)

Second, to test for a partisan bias in the differences in update between the two conditions, I ran a mixed ANOVA with rational belief update as the dependent variable, condition (large-PE Prediction vs Control) as a between-subject variable, and participant-item ideology congruence as a within-subjects variable. I found a significant main effect of condition, $F(1, 1398)=47.38$, $p<0.001$, $\eta_p^2=0.03$, a significant main effect of participant-item ideology congruence, $F(1, 1398)=3.91$, $p=0.0482$, and

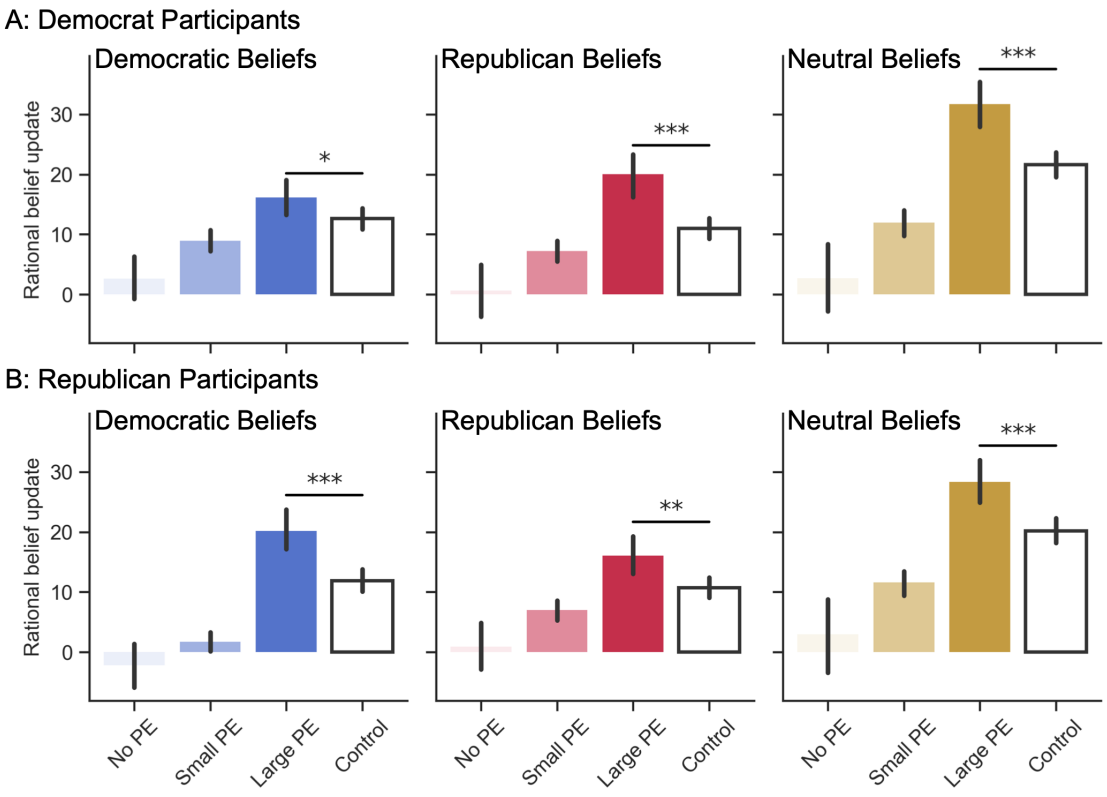


Figure 4.3: Rational belief update of Democrats (Panel A) and Republicans (Panel B) as a function of prediction error size split by belief ideology. Democratic beliefs in the Prediction Condition are represented in blue, Republican beliefs in the Prediction Condition are represented in red, and Neutral belief in the Prediction Condition are represented in yellow. The Control Condition is represented in white. Error bars represent ±1 standard error of the mean.

a significant interaction of condition with participant-item ideology congruence F(1, 1398)=4.89, p=0.02717, η_p^2=0.003, showing that overall, participants updated ideologically consistent items in large PE compared to ideologically consistent items in the Control condition less than ideologically inconsistent items in large PE compared to ideologically inconsistent items in the Control condition. This suggests a partisan bias in rational belief update, manifested as higher rigidity for large PE beliefs in one's own ideology. To explore whether this result further interacts with participant ideology I conducted another mixed ANOVA with rational belief update as the dependent variable, condition (large-PE Prediction vs. Control) and participant ideology (Democrats vs. Republicans) as between-subject variables, and

participant-item ideology congruence as a within-subjects variable, and I did not find a significant 3-way interaction F(1, 143)=0.45, p=0.501.

Discussion

In this study, I found that prediction error size linearly predicts rational belief update and that making large prediction errors leads to larger belief updates than being passively exposed to evidence. These effects hold for both Democrats and Republicans and for all belief types (Neutral, Democratic, Republican). Despite the fact that self-reported resistance to change significantly moderates the effect of PE on update and Republicans self-reported to be more resistant to change, I did not find differences between Democrats and Republicans in how they updated beliefs. Finally, I found a partisan bias manifested as higher rigidity for updating large PE beliefs in one's own ideology. To assess the replicability and generalizability of these findings, I conducted a high-powered replication with a US census matched sample.

4.2 Study 3.2: Replication in a US Census Matched Sample

Method

Open science practices. I preregistered the study's experimental design and hypotheses on an open science platform (https://aspredicted.org/blind.php?x=4qr2bt). The data for the replication study can be found on the study's open science framework page (https://osf.io/aur2t). The data analysis (in Python) can be accessed as a jupyter notebook here: https://github.com/mvlasceanu/PredictionBelief.

Participants. For the replication, I aimed for a US census matched sample of 1000 participants. I recruited 1387 Americans, using the Cloud Research platform, of which

313 were excluded based on preregistered criteria (i.e., failed attention checks). I conducted statistical analyses on the final US census matched sample of 1073 participants (57% female; Mage=48.32, SDage=16.92), that matched the census age, gender, race, and ethnicity quotas (see Table 4.3). The total sample contains 552 participants self-identified as Democrats, who were randomly assigned to the Experimental Condition (N=324) or the Control Condition (N=228), and 521 self-identified as Republicans who were randomly assigned to the Experimental Condition (N=296) or the Control Condition (N=225). The study protocol was approved by the Princeton University Institutional Review Board.

I used the same stimulus materials, procedure, and coding as in Study 3.1. The data for Study 3.2 was collected between May 26, 2020 and June 4, 2020.

		Census Proportion	Data Proportion
Gender	Male	49.4%	42.8%
	Female	50.6%	57.2%
Age	18-29	22.6%	19.9%
	30-39	16.8%	16.4%
	40-49	16.2%	14.8%
	50-59	17.8%	21.6%
	60-69	14.0%	16.0%
	70-99	12.4%	11.2%
Race	Caucasian	78.8%	75.6%
	African American	13.0%	9.9%
	Native American	1.2%	3.4%
	Asian	4.8%	4.9%
	Other	2.2%	5.8%
Ethnicity	Hispanic	16.0%	13.7%
	Not Hispanic	84.0%	86.1%

Table 4.3: Sample's demographic distribution compared to the US census.

Results

Does PE linearly predict rational update? I fitted a linear regression of Predic-

tion Error size against Rational Belief Update and replicated the result that Prediction Error size linearly positively predicts Rational Belief Update (β=2.36, SE=0.09, t(6466)=26.12, R^2=0.095, p<0.001; Figure 4.4A). As in Study 3.1, I ran a linear mixed model with rational belief update as the dependent variable, prediction error size and belief at pre-test as fixed effects, as well as by-participant random intercepts and by-item random intercepts, which included belief at pre-test as a fixed effect to control for the effect of the baseline level of belief on the degree of belief update, while observing the independent effect of prediction error size on rational belief update. I replicated the finding that prediction error size linearly predicts rational belief update (β=1.40, SE=0.07, t(22140)=19.50, p<0.001).

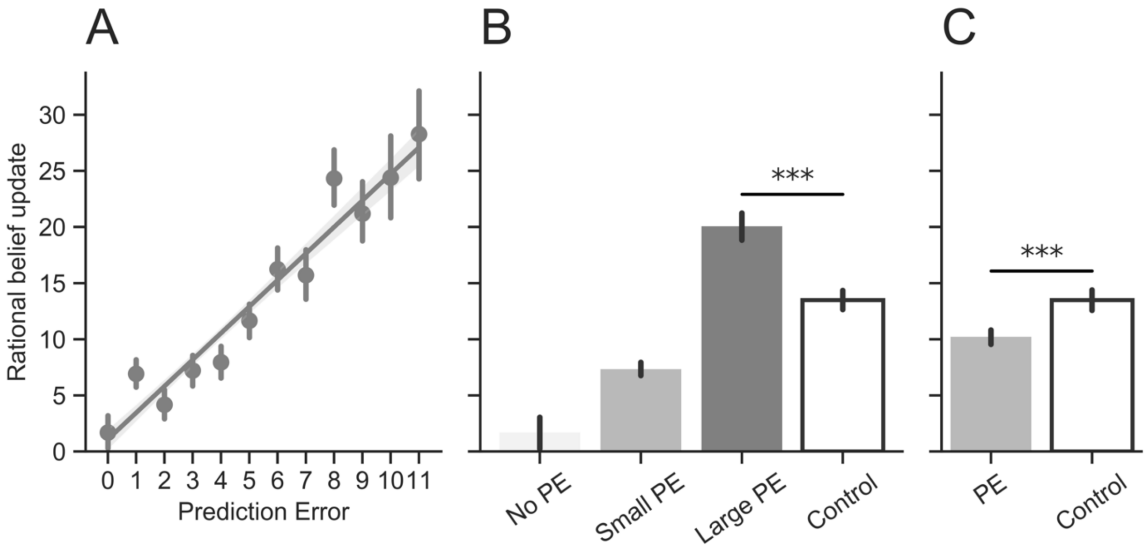


Figure 4.4: Rational belief update (post-test raw scores minus pre-test raw scores) as a function of prediction error size. Error bars represent 95% confidence intervals. Panel B: Rational belief update as a function of prediction error size (absolute value of prediction error, binned by no error, small error, and large error). The Prediction Condition is displayed in grey bars (increasing in color intensity as the prediction error size increases), and the Control Condition is displayed in white. Panel C: Rational belief update in the Prediction Condition (grey) and the Control Condition (white). Error bars represent ±1 standard error of the mean.

Do large PEs lead to more rational update than the Control Condition? I ran an independent sample t-test comparing rational belief update in the Prediction and

Control Conditions, and replicated the result that items in the Prediction Condition were rationally updated (M=9.81, SD=16.52) to a lower degree than items in the Control Condition (M=13.49, SD=9.79), t(825)=-5.92, p<0.001, Cohen's d=0.38, CI [-4.32, -2.23] (Figure 4.4C). To further explore this pattern as preregistered, I again assessed whether this conclusion applies independently of the size of the prediction error. With an independent sample t-test I also replicated the result that items in the large PE bin were rationally updated (M=20.06, SD=16.01) to a higher degree than items in the Control Condition (M=13.49, SD=9.79), t(1041)=8.301, p<0.001, Cohen's d=0.47, CI [4.90, 8.23] (Figure 4.4B). Of note, the proportion of items that ended up in the 3 bins of the prediction condition was: 9.2% in No PE, 62.7% in Small PE, and 28.1% in large PE.

Is there a partisan bias in how PE linearly predicts rational belief update? As in Study 3.1, to investigate whether there is a difference between Republicans and Democrats in how they update their beliefs I turned to the Prediction Condition. I ran a linear mixed model with rational belief update as the dependent variable, prediction error size, participant ideology, and belief at pre-test as fixed effects, and by-participant and by-item random intercepts. Contrary to Study 3.1 (where I found a non-significant interaction between PE and participant ideology) in Study 3.2 the interaction between PE and participant ideology reached statistical significance (β=-0.57, SE=0.12, t(22231)=-4.47, p<0.001), likely given the increased sample size. This interaction indicates that Republicans updated their beliefs (β=1.11, SE=0.16) less than Democrats (β=1.68, SE=0.09) as a function of prediction errors. This result is now consistent with the replicated findings that (1) self-reported resistance to change moderates the effect of PE on belief update (i.e., the interaction between PE and resistance to change [β=-0.50, SE=0.08, t(22190)=-6.18, p<0.001] shows that participants who self-reported as more resistant to change were less likely to update beliefs as a function of prediction errors), and (2) the finding that Republicans self-reported to be

more resistant to change (M=3.35, SD=0.87) than Democrats (M=3.03, SD=0.88), t(1062)=5.908, p<0.001, Cohen's d=0.361, CI [0.21, 0.42]. Furthermore, when also including item ideology (Democratic, Republican, Neutral) in the model, I replicated the finding that prediction error size linearly predicts rational belief update in all of the six ideological conditions crossing participant ideology and item ideology (i.e., Democrats on Neutral, Democratic, and Republican beliefs, as well as Republicans on Neutral, Democratic, and Republican beliefs; summarized in Table 4.4, plotted in Figure 4.5).

		β	SE	df	t	p
(intercept)		4.57	1.293	40.37	3.53	=0.001
Belief at pre-test		16.4	0.191	22170	85.97	<0.001
Democratic Participants	Neutral Items	2.13	0.141	14070	15.17	<0.001
Democratic Participants	Democratic Items	1.72	0.139	13920	12.36	<0.001
Democratic Participants	Republican Items	1.35	0.138	16240	9.76	<0.001
Republican Participants	Neutral Items	1.44	0.141	14310	10.26	<0.001
Republican Participants	Democratic Items	0.75	0.139	14200	5.42	<0.001
Republican Participants	Republican Items	0.97	0.140	16200	7.01	<0.001

Table 4.4: Rational belief update predicted by a linear mixed model testing the interaction of prediction error size with participant ideology (Democratic and Republican) and item ideology (Democratic, Republican, Neutral), while controlling for the belief at pre-test.

As before, now that I established that prediction errors linearly predict rational belief update in all ideological subsamples of the data, I tested how strongly this effect manifests in these ideological subsamples. I ran a linear mixed model on the Prediction Condition, with rational belief update as the dependent variable, prediction error size, participant ideology, item ideology, and belief at pre-test as fixed effects, as well as by-participant random intercepts and by-item random intercepts, and replicated the lack of significance of the 3-way interaction between prediction error size, item ideology (Democratic vs. Republican), and participant ideology (Democratic vs. Republican).

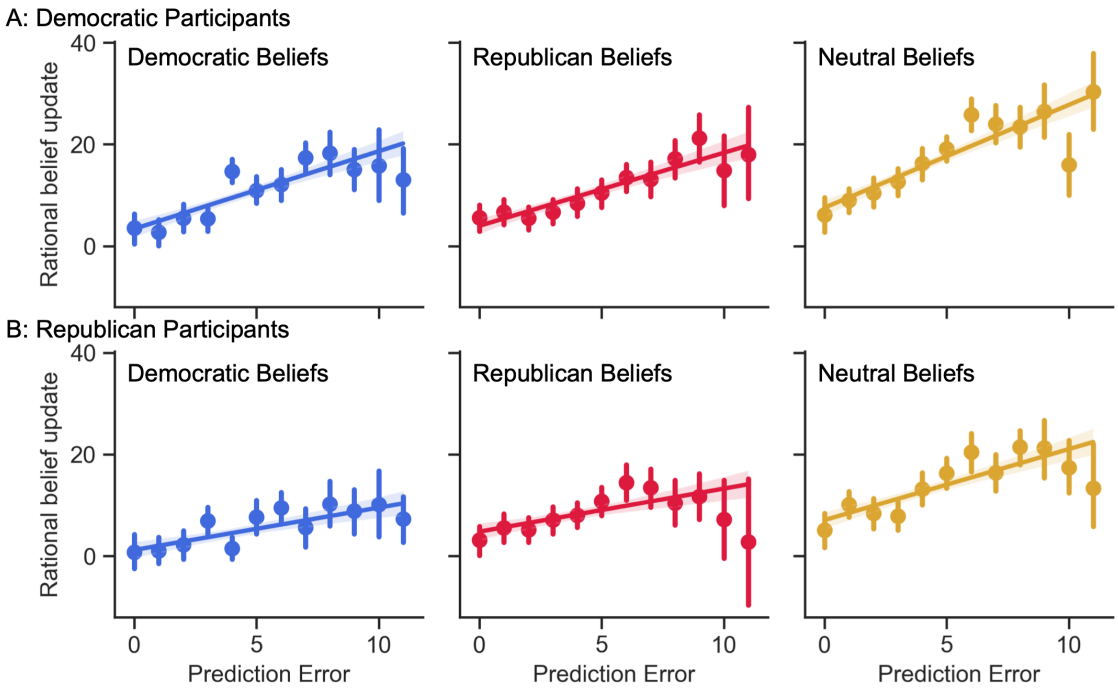


Figure 4.5: Rational belief update of Democrats (Panel A) and Republicans (Panel B) as a function of prediction error size split by belief ideology. Democratic beliefs are represented in blue, Republican beliefs are represented in red, and Neutral beliefs are represented in yellow. Error bars represent 95% confidence intervals.

When considering the measures of political polarization, in contrast with Study 3.1, now both political party affiliation strength (β=-0.16, SE=0.06, t(12402)=-2.628, p=0.0086) and support for President Trump (β=-0.15, SE=0.02, t(21440)=-5.767, p<0.001) reached statistical significance in moderating the effect of PE on rational update. These moderation analyses suggest that the more extreme one is on the ideological spectrum and the stronger they support President Trump the less they update beliefs according to prediction errors.

Is there a partisan bias in how beliefs are rationally updated in the large PE Condition compared to the Control Condition? To further investigate a potential ideological modulation of the uncovered effect of PE on rational update I again tested whether rational update is higher in the large PE bin of the Prediction Condition compared to the Control Condition in each of the six subsamples of the data (i.e., Democrats

Participant	*Item*	*Condition*	*M*	*SD*	*df*	*t*	*d*	*CI*	*p*
Democrats	Neutral	Large PE	29.89	37.38	526	6.74	0.54	[7.81, 14.99]	<0.001
		Control	19.76	34.65					
Democrats	Democratic	Large PE	16.62	31.58	539	5.11	0.41	[3.88, 9.35]	<0.001
		Control	11.09	30.77					
Democrats	Republican	Large PE	17.06	34.56	478	4.95	0.39	[4.07, 10.41]	<0.001
		Control	10.24	32.04					
Republicans	Neutral	Large PE	25.14	36.26	464	3.94	0.32	[3.21, 10.54]	<0.001
		Control	19.00	35.37					
Republicans	Democratic	Large PE	12.57	34.15	447	2.18	0.17	[0.08, 6.28]	=0.029
		Control	10.63	33.62					
Republicans	Republican	Large PE	14.46	35.56	414	3.04	0.25	[1.43, 7.94]	=0.002
		Control	10.22	31.39					

Table 4.5: Rational belief update difference between large PE and Control, in all participant ideologies (Democratic and Republican) and item ideologies (Democratic, Republican, Neutral).

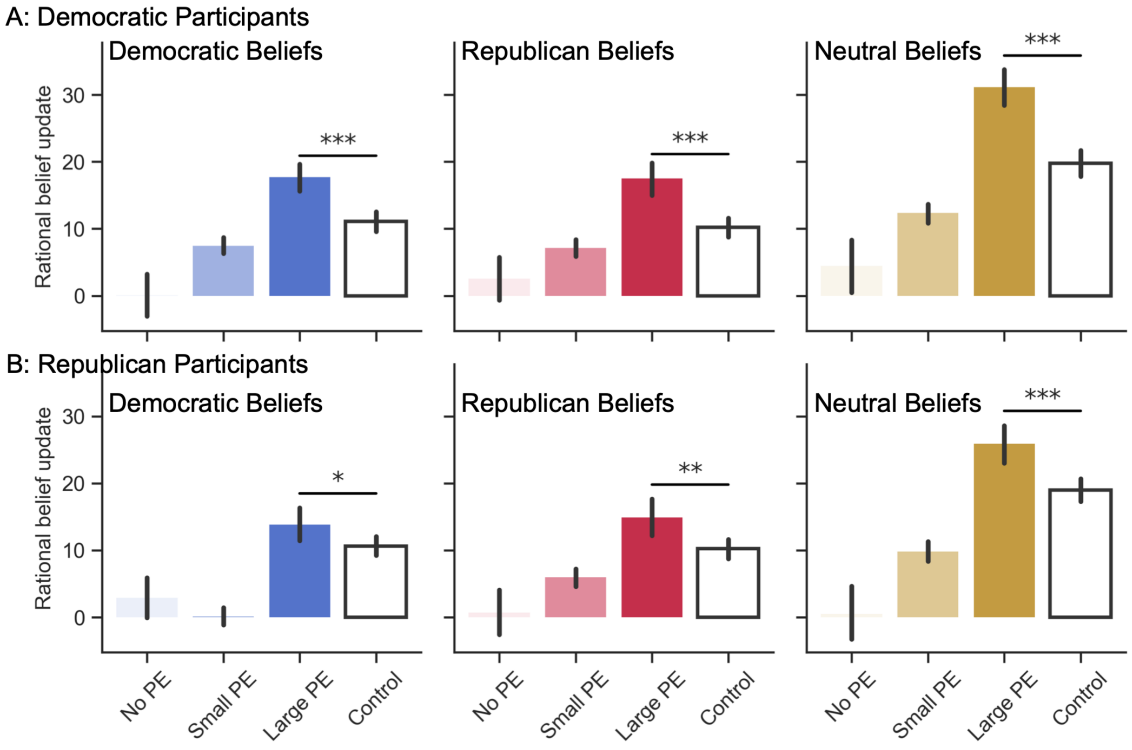


Figure 4.6: Rational belief update of Democrats (Panel A) and Republicans (Panel B) as a function of prediction error size split by belief ideology. Democratic beliefs in the Prediction Condition are represented in blue, Republican beliefs in the Prediction Condition are represented in red, and Neutral belief in the Prediction Condition are represented in yellow. The Control Condition is represented in white. Error bars represent ±1 standard error of the mean.

on Democratic, Republican, and Neutral items, and Republicans on Democratic, Republican, and Neutral items; summarized in Table 4.5, and displayed in Figure 4.6). I found that all of the independent sample t-tests comparing the large PE bin of the Prediction Condition to the Control Condition in each of the six ideological subsamples of the data were statistically significant (statistics reported in Table 4.5, plotted in Figure 4.6).

To test the partisan bias in belief update I obtained in Study 3.1 according to which participants were less likely to update their ideologically-consistent beliefs, I ran a mixed ANOVA with rational belief update as the dependent variable, condition (large-PE Prediction vs Control) as a between-subject variable, and participant-item ideology congruence as a within-subjects variable. In contrast to Study 3.1, I did not find an interaction between condition and congruence, $F(1, 28)=0.08$, $p=0.77938$, suggesting that participants were not more resistant to changing their party's ideological beliefs compared to the other party's beliefs.

Discussion

Study 3.2 replicated the main findings that prediction error size linearly predicts rational belief update and that making large prediction errors leads to a larger belief update than being passively exposed to evidence. It also replicated the result that these effects hold for both Democrats and Republicans and for all belief types (Neutral, Democratic, Republican). Moreover, I again found that self-reported resistance to change significantly moderates the effect of PE on belief update, and that Republicans self-reported to be significantly more resistant to change. Consistent with these effects (but in contrast to Study 3.1) I now found that Republicans updated all beliefs less than Democrats. Notably, I no longer found the higher rigidity in updating large error beliefs in one's party ideology.

General Discussion

Changing people's beliefs is notoriously difficult (Bendixen, 2012). Here, in two pre-registered studies - including one on a US census matched sample - I show that an intervention that builds on prediction errors could be successfully used, under specific circumstances, to change beliefs. The main finding that rational belief update is proportional to the magnitude of prediction error aligns with the associative learning principle that learning is proportional to prediction error (Rescorla & Wagner, 1972). I would claim that my findings do not constitute a simple extension of this prior work, given that beliefs are deemed as meaningfully different than knowledge due to their associated conviction and self-referential element (Connors & Halligan, 2014). The ideological dimension of both believers and their beliefs could have amplified, attenuated, or even eliminated the effect of prediction error on belief update. Yet the fact that they haven't points to its generalizability across the cognitive system.

These findings also align with prior work showing that surprising information can tune knowledge, attitudes, and beliefs (Ranney & Clark, 2016). The element of surprise employed to increase acceptance of climate change, for instance, likely operates similarly to the prediction error processes triggered in my paradigm (Ranney et. al., 2001). However, my findings supplement this work in several ways. Critically (1) I isolated the effect of prediction errors from that of evidence alone, (2) I quantified the magnitude of the surprise (i.e., prediction error size), which I then used to predict belief update, and (3) I incorporated beliefs as well as participants from both sides of the political-ideological spectrum, which allowed the comparison of the effect's magnitude both within and across ideological boundaries.

I note an important difference between the two studies in how the effects interact with ideology. In Study 3.1, I found a partisan bias in the form of higher rigidity in updating beliefs in one's own ideology (in large PE compared to Control). In Study 3.2, I found that Republicans updated all beliefs less than Democrats, suggesting a partisan

bias manifested as Republicans' resistance to changing all beliefs following prediction errors. At least two explanations could account for this difference in the manifestation of these partisan biases. First, the sample size and the national representativeness of Study 3.2 may have provided the statistical power and the necessary variation to observe the true effect – Republican participants' diminished belief-change based on prediction errors. Second, and perhaps more interestingly, the different socio-political contexts at the time of the data collection between the studies (October, 2019 versus May, 2020) might have shifted the ideological bias from a symmetric effect (for both Democrats and Republicans) to the Republicans' resistance to update all beliefs. This possibility is consistent with existing work on the impact of threat and uncertainty on political beliefs (Haas & Cunningham, 2014). While difficult to programmatically explore in a highly dynamical real-world situation (i.e., COVID-19 and nation-wide anti-racism protests), further research clarifying how consequential events affect belief change is certainly worthwhile pursuing.

A reliable finding across the two studies was that the belief update in the prediction condition was, on average, significantly lower than in the control condition. I speculate that having to remember both the correct and predicted answer could create interference (when the difference between them is small) resulting in a memory decrement for the correct answer in the prediction condition compared to the control condition. This memory decrement could, in turn, lead to less rational belief update. Alternatively, there might be a cost to changing one's mind (e.g., one may appear inconsistent) so people might only be willing to pay that cost when extraordinary evidence is presented. Regardless of the mechanism, there is a pragmatic implication of the difference between the prediction and the control conditions. When addressing an audience for which one has no baseline belief information one should simply provide accurate information. On the other hand, when one does have baseline belief

information about a community, one would be well-served to attack misinformation by narrowing the message to the most egregious belief violations.

Several important aspects of the belief updating process were omitted in this study. One such factor is the credibility of the source presenting the evidence (Chung, Fink, Kaplowitz, 2008). Future studies could explore how information source affects the incorporation of evidence into one's belief system and how this impact might interact with ideology. For example, a Democrat receiving evidence against a Democratic belief from CNN might update their belief accordingly, whereas evidence from Fox News might be completely discarded. Conversely, a Republican may be more open to evidence incorporation when watching Fox News compared to CNN (Haidt, Graham, & Joseph, 2009). Another important extension could involve investigating the effect of conversations on prediction-based belief update and how these conversations, when they occur in larger communities, could impact collective beliefs (Vlasceanu, Enz, Coman, 2018). One possibility is that when given the opportunity to discuss, people would display a novelty bias and mention the evidence most surprising to them. Conversely, people might instead display a confirmation bias, and mention the evidence they correctly predicted. Depending on what they choose to discuss, the community's collective beliefs would be shaped accordingly, as previous research found that conversations influence collective beliefs (Vlasceanu, Morais, Duker, Coman, 2020). Clarifying this process would be particularly meaningful for policy makers interested in impacting communities (Dovidio & Esses, 2007).

Beyond their theoretical importance, these findings might provide useful tools in the battle against misinformation, a prominent threat facing the world today (Lewandowsky et al, 2012. For example, a third of Americans believe global warming is a conspiracy (Swift, 2013), a third of American parents believe that vaccines cause autism (National Consumers League, 2014), and 30% of Americans believe COVID-19 was engineered in a lab (Pew Research Center, 2020). False beliefs are

dangerous when endorsed by a large proportion of people, as they can shift attention and resources away from real threats, dramatically impact normative behavior, and cause suboptimal collective decisions (Kuklinski et al, 2000). Crucial steps in the misinformation prevention battle are understanding the processes driving belief update and using that understanding to design misinformation-combating interventions. The present findings point to such interventions. For instance, as an alternative to refutation, which may backfire especially if beliefs are ideologically charged (Nyhan & Reifler, 2010), these findings point to a powerful strategy that could shortcut ideological biases. First, one needs to map the community's estimates on relevant statistics that can be used as surprising evidence. These statistics need to be carefully compiled given that people's predictions about everyday events are fairly accurate (Griffiths & Tenenbaum, 2006). After selecting the statistics eliciting the largest misestimates, these pieces of evidence need to be disseminated back to the community in a predictions-then-feedback format. This procedure is intensive but might have a stronger impact in diminishing misinformation than existing approaches. Conducting more empirical research to ensure the stability of these findings, their boundary conditions, and behavioral instantiation could offer policy makers a powerful tool to address this global epidemic.

So far, I established that memory accessibility and emotionally arousing images at the belief level of the cognitive framework, and predictions regarding surprising evidence at the evidence level of the cognitive framework can be used as strategies that lead to belief change. What happens when manipulations occur at the level of the social norms (Figure 1.1)? I predicted that changes at the social norms level would also be reflected in changes at a belief level.

Chapter 5

Social Norms and Beliefs

5.1 Study 4: Social Norms Trigger Belief Change

Abstract

People are constantly bombarded with information they could use to adjust their beliefs. Here, we are interested in exploring the impact of social norms on belief update. To investigate, we recruited a sample of 200 Princeton University students, who first rated the accuracy of a set of statements (pre-test). They were then provided with relevant evidence either in favor or against the initial statements, and they were asked to rate how convincing each piece of evidence was. The evidence was randomly assigned to appear as normative or non-normative, and also randomly assigned to appear as anecdotal or scientific. Finally, participants rated the accuracy of the initial set of statements again (post-test). The results show that participants changed their beliefs more in line with the evidence, when the evidence was scientific compared to when it was anecdotal. More importantly to our primary inquiry, the results show that participants changed their beliefs more in line with the evidence when the evidence was portrayed as normative compared to when the evidence was portrayed as non-normative, pointing to the impactful influence social norms have

on beliefs. Both effects were mediated by participants' subjective evaluation of the convincingness of the evidence, indicating the mechanism by which evidence is selectively incorporated into belief systems.

Introduction

False statements cluttering the informational landscape have always been a challenge for societies (Garrett, 2011), although their detrimental effects have increased over time with the advancements in technology that allows false information to spread faster than accurate information (Vosoughi, Roy, & Aral, 2018). And believing false information can have destructive outcomes in all aspects of society, as concluded by numerous empirical articles linking false beliefs to decreased vaccination rates (Jolley & Douglas, 2014), increased climate change denial (Lewandowsky, Gignac, Oberauer, 2015), and increased intergroup prejudice (Jolley, Meleady, Douglas, 2020). Encouragingly, beliefs are subject to change, given their dynamic nature (Bendixen, 2002). And indeed, prior work has identified several strategies that proved effective at changing beliefs, such as using fictional narratives (Wheeler, Green, Brock, 1999), nudging accuracy goals (Pennycook et al., 2020), manipulating memory accessibility (Vlasceanu & Coman, 2018; Vlasceanu, Morais, Duker, Coman, 2020), appending emotional arousing images (Vlasceanu, Goebel, Coman, 2020), and triggering prediction errors (Vlasceanu, Morais, Coman, 2020). However, changing one's beliefs is not a trivial task. Adjusting beliefs by incorporating new evidence has been shown particularly challenging if the evidence increases cognitive dissonance (Festinger & Carlsmith, 1959), reduces coherence among already held beliefs (Lord, Ross, Lepper, 1979), or counter one's political allegiance (Nyhan & Reifler, 2010). Here, I am interested in exploring strategies of making people more receptive to evidence in changing their beliefs.

Building on the vast literature on the impact of social norms in people's lives (Cialdini & Trost 1998; Cialdini & Goldstein, 2004; Paluck & Green, 2009), I propose that one such strategy involves portraying the evidence as normative. This hypothesis is supported by past research showing that information qualified by high virality metrics on social media platforms (i.e., high number of likes, shares on Twitter) appears as more believable than information qualified by low virality metrics (i.e., low number of likes, shares on Twitter), effect explained by differences in the perception of descriptive and injunctive norms around sharing information of high versus low virality (Kim, 2018). Thus, if normative information is perceived as more believable, is it also more likely to be incorporated in people's beliefs when encountered as evidence, triggering belief change?

Additionally, I am interested in what type of evidence is more effective at changing beliefs. Prior work has identified messages featuring narratives/anecdotes (Escalas, 2007; McQuiggan et al, 2008), and scientific information (Zhang et al, 2019) as having strong influence on people. Therefore, here, I am interested in whether anecdotal or scientific evidence is more likely to be incorporated into and change people's beliefs. In support of anecdotal evidence being more effective at changing beliefs than scientific evidence, prior work showed that messages featuring narratives were more effective at persuading people to sign up for organ donation than messages conveying statistics (Weber et al, 2006). In another study regarding health interventions, narratives were perceived as more believable than newsletter articles (Slater et al, 2003). In the domain of cancer prevention, the most shared messages on the Twitter platform were anecdotal experiences (Chung, 2017; So et al., 2016), which have been shown to have a boosted online presence because they enhance users' emotional involvement with the messages (Berger, 2014). In support of scientific evidence being more effective at changing beliefs than anecdotal evidence, prior work showed that informational tweets were shared more than personal experience tweets (Zhang et al, 2019), and public

health organizations typically post factual information on social media platforms (Zhang et al, 2019; Lyles et al, 2013).

To test belief change as a function of normative perceptions and evidence type, I designed an experiment composed of three phases: pre-test, evidence, and post-test. First, participants rated the accuracy of a set of statements (e.g., "Only-children have higher self-esteem"; pre-test phase). They were then exposed to a series of tweets, serving as pieces of evidence either in favor or against the initial statements, and were asked to rate how convincing each piece of evidence was. Half of the tweets were randomly assigned to appear as anecdotes, and the other half as scientific findings; also, half of them were randomly assigned to appear as having a large number of retweets, likes, and comments (normative) whereas the other half only a small number of retweets, likes, and comments (non-normative). Thus, I constructed a 2 by 2 experimental design with type (anecdotal vs. scientific) and normativity (normative vs. non-normative) as within-participant independent variables. Lastly, participants were asked to rate the accuracy of the initial statements again (post-test phase).

The first hypothesis was that the scientific evidence would influence beliefs more than anecdotal evidence (i.e., lead to more belief change). The second hypothesis was that normative evidence would influence beliefs more than non-normative evidence. The last hypothesis was that both effects would be mediated by the degree of evidence convincingness.

Method

Open science practices. The materials and data can be found on my open science framework page: https://osf.io/7nvcf/ The data analysis (in python and R) can be viewed as a jupyter notebook here: https://github.com/mvlasceanu/normativebeliefs

Participants. A total of 200 Princeton undergraduate students (Mage=19.49, SDage=1.39; 64% women) were recruited for the study. They participated in the

study for either monetary compensation or research credit. All participants passed preestablished attention checks. I aimed for a sample size of 200 participants to achieve a 0.8 power for an effect size of 0.2 in a two tailed paired comparison at an alpha level of 0.05. The study was approved by the Institutional Review Board at Princeton University.

Stimulus materials. I undertook preliminary studies to pretest a set of 32 statements of moderate believability (e.g., "Eating carrots will make eyesight sharper"). A pilot study was conducted on a Cloud Research sample (N=217; Mage=54.16, SDage=16.3; 82% women), in which I collected believability ratings (i.e., "How accurate or inaccurate do you think this statement is" on a scale from 0-Extremely Inaccurate to 100-Extremely Accurate). I conducted this pilot to make sure all the statements were moderately believable, to avoid any floor or ceiling effects in belief change. Even though each of these 32 statements was moderately believable by design, half of them were actually accurate, while the other half were actually inaccurate pieces of information, as determined by published scientific papers or other official sources.

I also constructed a set of 128 pieces of direct evidence (32 scientific & normative, 32 scientific & non-normative, 32 anecdotal & normative, 32 anecdotal & non-normative). The pieces of evidence were constructed such that they always argued in favor of the initial statement if the statement was accurate (e.g., "Children who spend less time outdoors are at greater risk to develop nearsightedness, study shows") and argued against the initial statement if the statement was inaccurate (e.g., "Eating carrots does not makes eyesight sharper, study shows"). To increase external validity, these pieces of evidence were constructed and displayed to participants in as if they were tweets collected from the Twitter platform. To construct the conditions of interest, I counterbalanced the phrasing suggesting the evidence posted is the result of a scientific study (scientific condition) with phrasing suggesting the

evidence posted is anecdotal (anecdotal condition), such that each statement was in either counterbalancing condition with equal probability randomly assigned across participants. I also counterbalanced the number of retweets, likes, and comments of each tweet, to have either a large (normative condition) or a small (non-normative condition) number of retweets, also assigned with equal probability across participants. Thus, I constructed a 2 by 2 experimental design with type (anecdotal vs. scientific) and normativity (normative vs. non-normative) within subjects variables. Moreover, for each piece of evidence, the sources of the tweets were constructed to be as similar as possible while allowing some variability, to maintain the credibility of the stimuli. In each case, the person tweeting was depicted as a white middle-aged male, with a common name and appearance. The dates of the tweets were randomly chosen from dates in the month on July 2019.

Design and procedure. Data collection occurred between October and December 2019. Participants were told they would participate in an experiment about people's evaluation of information and were directed to the survey on the Qualtrics platform. After completing the informed consent form, participants were directed to the first phase (pre-test), in which they rated a set of 32 statements (one on each page) by indicating the degree to which they believed each statement (i.e., "How accurate do you think this statement is," from 1-Extremely inaccurate to 100-Extremely accurate). Then, in the second phase (evidence phase), participants were exposed to a set of 32 pieces of evidence (in the form of tweets, one on each page), half of which argued in favor and the other half argued against the initial statements. At this stage, participants were instructed to rate each tweet on how convincing it appeared to them by answering the question "How convincing is this tweet?" from 1-Not at all to 100-Very much so. Finally, in the third phase (post-test) participants rated again the believability of the initial 32 statements, after which they were asked to complete a series of demographic information and were debriefed.

Analysis and coding. I operationalize rational belief update as the belief change from pre-test to post-test in the direction corresponding to incorporating the available evidence. For statements with supporting evidence – the rational update is to increase in believability from pre-test to post-test. For statements followed by refuting evidence –the rational update is to decrease in believability from pre-test to post-test. Through counterbalancing, I ensure that participants cannot trivially infer that "correct" updates must be in one direction.

Results

Initial Checks. Since there was no difference in the initial level of believability of accurate statements (M=50.96, SD=10.58) and inaccurate statements (M=52.14, SD=11.07), p=0.766, as intended, I combined them for the rest of the analyses conducted. I first ran a repeated measures ANOVA with rational belief update as the dependent variable, and evidence type (anecdotal-normative, anecdotal-non-normative, scientific-normative, scientific-non-normative) as a within-subject variable, and found a main effect of evidence type $F(3, 597)=6.54$, $p<0.001$, $\eta_p^2=0.03$ (Figure 5.1C). I also ran a repeated measures ANOVA with rational belief update as the dependent variable, and evidence type (anecdotal vs. scientific) and normativity (normative vs. non-normative) as a within-subject variables, and found a main effect of evidence type $F(1, 199)=6.81$, $p<0.0098$, $\eta_p^2=0.03$ and a main effect of normativity $F(1, 199)=10.27$, $p<0.0016$, $\eta_p^2=0.04$, but no type by normativity interaction $F(1, 199)=0.93$, $p=0.336$, $\eta_p^2=0.005$ (Figure 5.1C).

Scientific versus Anecdotal Evidence. To investigate the first hypothesis, that scientific evidence leads to more belief change than anecdotal evidence, I ran a paired sample t-test and found that indeed, when statements were followed by scientific evidence their endorsement changed more to align with that evidence (M=3.03,

SD=7.36) than when they were followed by anecdotal evidence (M=1.74, SD=6.96), t(199)=2.87, p<0.0043, d=0.18, CI=[0.29, 2.28] (figure 5.1A).

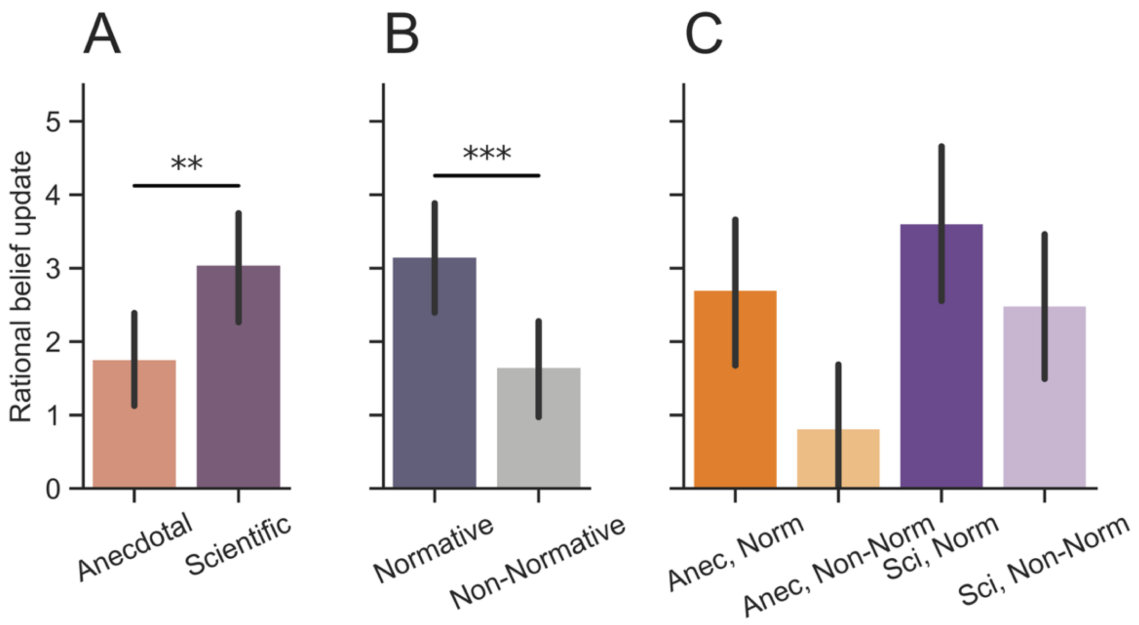


Figure 5.1: Rational belief update as a function of evidence type (Anecdotal versus Scientific in Panel A; Normative versus Non-Normative in Panel B; Anecdote-Normative, Anecdote-Non-Normative, Scientific-Normative, Scientific-Non-Normative in Panel C). Error bars represent ±1 standard errors of the mean.

Normative versus Non-Normative Evidence. To investigate the second hypothesis, that normative evidence leads to more belief change than non-normative evidence, I ran a paired sample t-test and found that indeed, when statements were followed by normative evidence their endorsement changed more to align with that evidence (M=3.14, SD=7.52) than when they were followed by non-normative evidence (M=1.64, SD=6.76), t(199)=3.45, p<0.001, d=0.21, CI=[0.5, 2.5] (figure 5.1B).

Mechanism: Mediation analyses. To asses the last hypothesis, that the main effects of evidence type (anecdotal versus scientific) and normativity (normative versus non-normative) on rational belief change would be mediated by how convincing the evidence was perceived by participants, I ran two mediations models, following guide-

lines and using the R mediation package published by Tingley and colleagues (2014). Evidence convincingness as a function of condition can be observed in Figure 5.2.

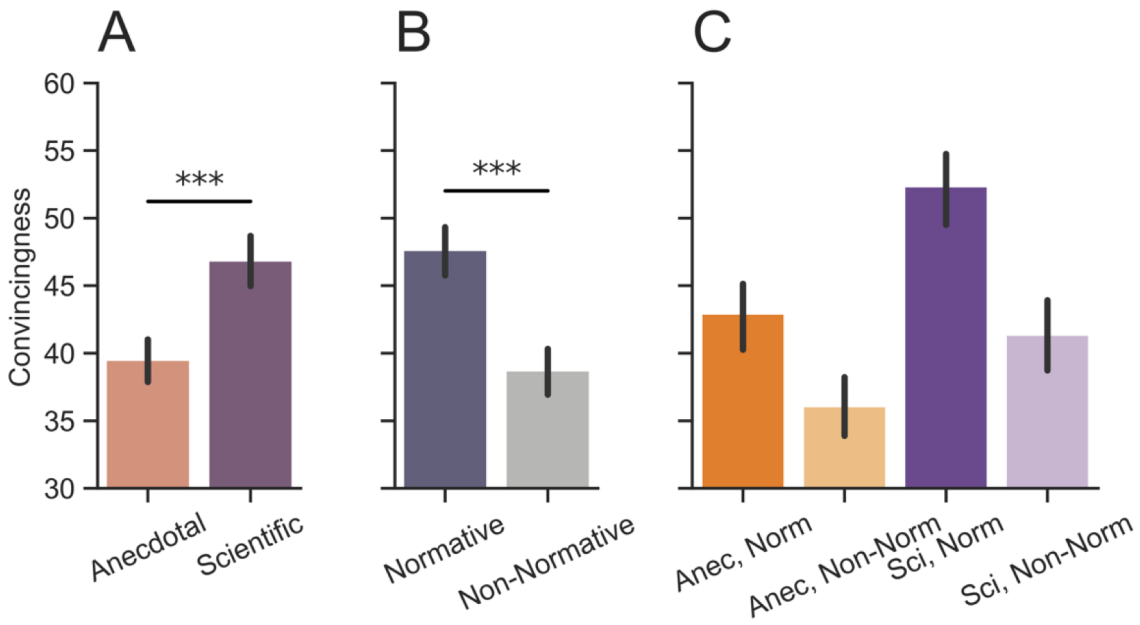


Figure 5.2: Evidence convincingness as a function of evidence type (Anecdotal versus Scientific in Panel A; Normative versus Non-Normative in Panel B; Anecdote-Normative, Anecdote-Non-Normative, Scientific-Normative, Scientific-Non-Normative in Panel C). Error bars represent ±1 standard errors of the mean.

First, the relationship between evidence type (anecdotal versus scientific) and rational belief update was mediated by evidence convincingness. As Figure 5.3 illustrates, the regression coefficient between evidence type (anecdotal versus scientific) and belief change was statistically significant, as were the regression coefficients between evidence type and evidence convincingness and between evidence convincingness and belief change when controlling for evidence type. I tested the significance of the indirect effect using bootstrapping procedures. The indirect effect was computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped indirect effect was 0.569, and the 95% confidence interval ranged from

0.22 to 1.01. Thus, the indirect effect was statistically significant, p<0.001 (Table 5.1; Table 5.2).

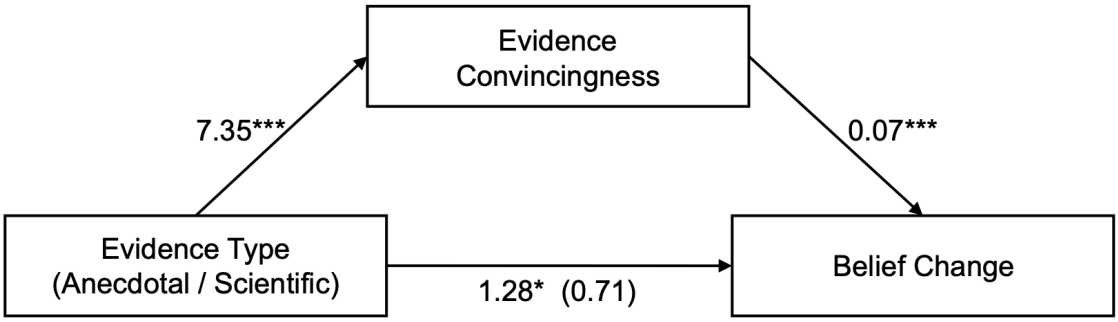


Figure 5.3: Regression coefficients for the relationship between evidence type (anecdotal versus scientific) and belief change as mediated by evidence convincingness. The standardized regression coefficient between evidence type and belief change, controlling for evidence convincingness, is in parentheses.

Table 1
Regression analyses associated with the first mediation model.

Predictors	*b* (s.e.)	*t*	*F*	*df*	R^2	*p*
Model 1						
Evidence type	1.28 (0.56)	2.27*	5.18	(1, 398)	0.01	0.023
Model 2						
Evidence type	0.71 (0.56)	1.26	12.43	(2, 397)	0.05	0.206
Evidence convincingness	0.07 (0.01)	4.40***			0.05	0.000

b = regression coefficients; s.e. = standard error
* p<0.05; ** p<0.01; *** p<0.001

Table 5.1: Regression analyses of first mediation model

Second, the relationship between evidence type (normative versus non-normative) and rational belief update was also mediated by evidence convincingness. As Figure 5.4 illustrates, the regression coefficient between evidence type (normative versus non-normative) and belief change was statistically significant, as were the regression coefficients between evidence type and evidence convincingness and between evidence convincingness and belief change when controlling for evidence type. I tested the

Table 2
Causal mediation analyses: nonparametric bootstrap CI, with 10,000 simulations

	Estimate	95%CI lower	95%CI upper	p
Indirect Effect (ACME)	0.56	0.22	1.01	0.0004***
Direct Effect (ADE)	0.71	-0.47	1.90	0.2342
Total Effect	1.28	0.14	2.39	0.0260*
Proportion Mediated	0.44	0.11	0.02	0.0264*

ACME = average causal mediation effects; ADE = average direct effect

Table 5.2: Causal mediation analyses: nonparametric bootstrap CI

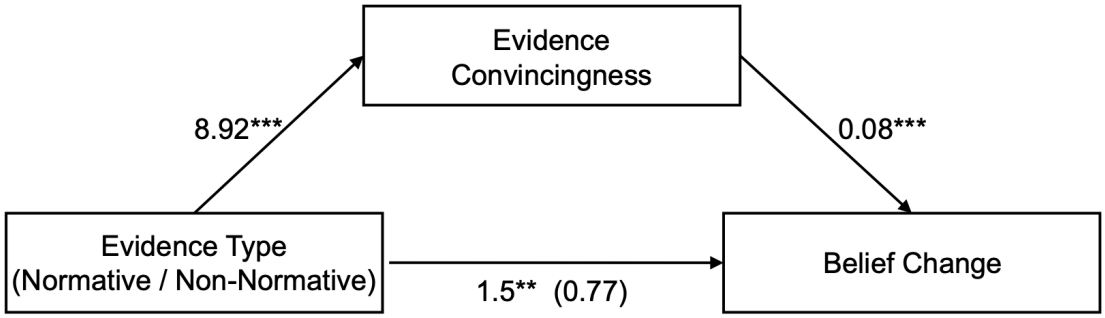


Figure 5.4: Regression coefficients for the relationship between evidence type (normative versus non-normative) and belief change as mediated by evidence convincingness. The standardized regression coefficient between evidence type and belief change, controlling for evidence convincingness, is in parentheses.

Table 3
Regression analyses associated with the second mediation model.

Predictors	*b* (s.e.)	*t*	*F*	*df*	R^2	*p*
Model 1						
Evidence type	1.50 (0.55)	2.70**	7.33	(1, 398)	0.01	0.007
Model 2						
Evidence type	0.77 (0.56)	1.37	15.03	(2, 397)	0.06	0.169
Evidence convincingness	0.08 (0.01)	4.72***			0.06	0.000

b = regression coefficients; s.e. = standard error
* p<0.05; ** p<0.01; *** p<0.001

Table 5.3: Regression analyses of second mediation model

Table 4
Causal mediation analyses: nonparametric bootstrap CI, with 10,000 simulations

	Estimate	95%CI lower	95%CI upper	p
Indirect Effect (ACME)	0.72	0.34	1.19	0.0000***
Direct Effect (ADE)	0.77	-1.95	0.35	0.1778
Total Effect	1.50	0.43	2.61	0.0064**
Proportion Mediated	0.48	0.18	1.63	0.0064**

ACME = average causal mediation effects; ADE = average direct effect

Table 5.4: Causal mediation analyses: nonparametric bootstrap CI

significance of the indirect effect using bootstrapping procedures. The indirect effect was computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped indirect effect was 0.728, and the 95% confidence interval ranged from 0.34 to 1.19. Thus, the indirect effect was statistically significant, $p<0.001$ (Table 5.3; Table 5.4).

Discussion

People are constantly bombarded with information they could use to adjust their beliefs. Here, I show that individuals change their beliefs more in line evidence when evidence is portrayed as normative compared to when evidence is portrayed as non-normative, pointing to the strong influence social norms have on beliefs. This finding aligns with the literature showing the strong influence of social norms on behavior change (Cialdini & Trost 1998; Cialdini & Goldstein, 2004; Paluck & Green, 2009), and extends it to belief change. This finding also aligns with prior work showing that normative information is perceived as more believable than non-normative information, extending the role of social norms from the passive perception of believability (Kim, 2018) to the active changing of beliefs.

Moreover, I show that individuals change their beliefs more according to scientific evidence than according to anecdotal evidence. Given the sample is composed of Princeton University students, attuned and used to digesting scientific information, this finding might not seem surprising. To ensure the generalizability of this effect in the larger population, a replication with a more representative sample of the population is necessary. This effect is consistent with prior work showing that factual information is shared more than anecdotes given perceived social norms of sharing information (Zhang et al, 2019), and is encouraging given that public health organizations typically distribute factual and not anecdotal information (Lyles et al, 2013). Furthermore, I show that both effects of evidence normativity and evidence type on belief change are mediated by self-reported evidence convincingness. These mediation analyses provide support for the mechanism by which evidence is incorporated by the cognitive system in people's beliefs, leading to belief change.

In the present work, I use a controlled, experimental approach to studying belief change. Constraining the investigation to these minimal conditions allows us to isolate the effect of social norms on belief change, and the effect of evidence type on belief change. It is important to note, however, that in real world situations, additional factors such as (1) conversational interactions following exposure to evidence, or (2) source credibility, would likely affect the degree to which the people integrate evidence into their beliefs. First, after receiving new information, people typically discuss it with each other in conversations (Liu, Jin, Austin, 2013). Thus, this line of work would benefit from being extended from the individual to the collective belief level by incorporating conversations within communities in the investigation of information normativity and evidence type on belief change. It might be that when given the opportunity to discuss, people mostly mention the anecdotal evidence they were exposed to, omitting the scientific information. Since conversations influence how individual level processes scale to give rise to collective level phenomena (Vlasceanu,

Enz, Coman, 2018; Vlasceanu & Coman, 2020), the community's emergent collective beliefs could be different as a result of the same evidence, from those of a group of non-interacting individuals. Clarifying this process is important for policy makers interested in impacting communities (Dovidio & Esses, 2007). Second, another noteworthy aspect of belief change not included here is source credibility (Chung, Fink, Kaplowitz, 2008; Slater & Rouner, 1996; Vlasceanu & Coman, 2020). This line of work would also benefit from future investigations into how the source presenting the evidence might interact with the effects of social norms and evidence type on belief change. For instance, perceived norms have been shown to be most influential when they arise from others with whom I share a common identity (Abrams, Wetherell, Cochrane, Hogg, Turner, 1990; Centola 2011). Therefore, it could be that identifying with the person sharing anecdotal evidence might increase the likelihood of incorporating that evidence in changing beliefs.

Beyond their theoretical importance, these findings might prove useful tools in the battle against misinformation, one of the top threats faced by the world today (Farkas & Schou, 2019; Lewandowsky et al, 2012). Emerging research has been using social science to understand and counter the spread of false information (Guess, Nagler, Tucker, 2019) using strategies such as debunking (Wegner, Wenzlaff, Kerker, & Beattie, 1981), prebunking (van der Linden, Leiserowitz, Rosenthal, Maibach, 2017), or nudging accuracy (Pennycook et al, 2020). The findings suggest that, when targeting a highly educated population, the focus should be placed on communicating scientific evidence in support of accurate information, as opposed to communicating anecdotal evidence. Second, these findings also suggest that when available, normativity cues favoring accurate information should be made salient (e.g., "90% of Americans believe vaccines are safe" or conversely, "Only 10% of Americans believe vaccines cause autism"), as they can increase the endorsement of

accurate information and decrease the endorsement of misinformation.

Also at the social norms level of the cognitive structure, I investigated which information source is more likely to have an impact on belief change.

5.2 Study 5: Information Sources Differentially Trigger Belief Change

Abstract

During a global health crisis people are exposed to vast amounts of information from a variety of sources. Here, we assessed which information source could increase knowledge about COVID-19 (Study 5a) and COVID-19 vaccines (Study 5b). In Study 5a, a US census matched sample of 1060 Cloud Research participants rated the accuracy of a set of statements and then were randomly assigned to one of 10 between-subjects conditions of varying sources providing belief-relevant information: a political leader (Trump/Biden), a health authority (Fauci/CDC), an anecdote (Democrat/Republican), a large group of prior participants (Democrats/Republicans/Generic), or no source (Control). Finally, they rated the accuracy of the initial set of statements again. Study 5b involved a replication with a sample of 1876 Cloud Research participants, and focused on COVID-19 vaccine information and vaccination intention. In both studies, we found that participants acquired most knowledge when the source of information was a generic group of people. Surprisingly, knowledge accumulation from the different information sources did not interact with participants' political affiliation.

Introduction

In December 2019, a new coronavirus generated a fast-spreading pandemic, which reached 160 countries by March 2020 (WHO, 2020). Given that infectious diseases have been responsible for the greatest human death tolls in history (Scott & Duncan, 2001), the spread of COVID-19 triggered panic, confusion, and uncertainty in the population (Depoux et al., 2020). This created the perfect storm for the spread of another, equally consequential epidemic: misinformation and conspiracy theories (Starbird, 2019; van Prooijen & Douglas, 2017). In this context of uncertainty, an unregulated social media environment provided fertile ground for the dissemination of such beliefs (Ellis, 2020; Frenkel, Alba, & Zhong, 2020; McCauley & Jacques, 1979; van Prooijen & Douglas, 2017). Scientific research soon confirmed that most people held at least one misperception about COVID-19 (Pennycook, McPhetres, Bago, & Rand, 2020), particularly problematic since misinformation has been associated with harmful consequences. For example, belief in conspiracy theories has been linked to decreased vaccination rates (Jolley & Douglas, 2014), increased climate change denial (Lewandowsky, Gignac, & Oberauer, 2015), and increased intergroup prejudice (Jolley, Meleady, & Douglas, 2020). Conversely, knowledge, and belief in accurate information, have been shown to have beneficial effects in times of crisis, consistent with the idea that people's behavior is influenced by knowledge (Janz & Becker, 1984). For instance, having more COVID-19 knowledge was associated with a lower likelihood of engaging in dangerous behaviors such as going to crowded places or not wearing masks (Zhong et al., 2020). Therefore, increasing people's knowledge by promoting accurate information and reducing misinformation is essential during such a global crisis. So far, strategies to increase COVID-19 knowledge through accuracy nudges, such as reminding people to think about accuracy when reading COVID-19 related information, have been found successful at increasing the perceived accuracy of accurate information (Pennycook, McPhetres, Zhang, & Rand,

2020). Expanding this work, we aim to explore ways in which we could facilitate the acquisition of COVID-19 knowledge, by increasing belief in accurate information and decreasing belief in conspiracy theories. To this end, we will assess the role of information sources and political ideologies in the COVID-19 knowledge assimilation and belief update. Prior literature has established that the source of information has an important impact on knowledge assimilation. The credibility of the source was found to influence belief change in a variety of domains (Chung, Fink, & Kaplowitz, 2008; Slater & Rouner, 1996). Overwhelmingly, statements made by credible sources are more likely to be believed and integrated in one's mental model (Begg, Anas, & Farinacci, 1992). In the present study, we aim to compare the relative effectiveness of varying sources of information increasing overall COVID-19 knowledge acquisition. We will assess the impact of different sources of information on people's beliefs in a horse-race design, with the same information being transmitted by: (a) groups of people, either ideologically committed or not, (b) ideologically committed individuals (i.e., political figures), and (c) experts. The design employed here will serve to establish which sources are most effective at facilitating knowledge assimilation and whether there are any ideological biases to knowledge incorporation. Source credibility has also been found to influence concrete behavioral intentions such as voting (Mondak, 1995) and purchasing intentions (Lafferty & Goldsmith, 1999; Till & Busler, 1998). The current pandemic context allows us to assess whether accumulating knowledge from various sources regarding the COVID-19 vaccine leads to increased vaccination intentions, and which source leads to the highest increase in such intentions. The influence of groups of people has been investigated in the vast literature on social norms, defined as the perception of what others are doing, approve, or disapprove of (Cialdini & Goldstein, 2004). People heavily rely on social norms to understand the situations they are in, especially in contexts of uncertainty, and are a strong predictor of behavior (Cialdini & Trost 1998; Cialdini & Goldstein,

2004). However, although people are influenced by norms, their perceptions are often inaccurate, and they tend to either underestimate or overestimate others' behaviors, especially health-related ones (Miller & Prentice, 1996; Berkowitz, 2005). Leveraging these effects, we are interested in whether COVID-19 related knowledge can be promoted by portraying believing accurate information as normative and believing conspiracy theories as counter normative. Given previous work showing that group norms are also significant predictors of intentions (Fielding, Terry, Masser, & Hogg, 2008), we are also interested in whether promoting knowledge about the COVID-19 vaccines can increase vaccination intentions. But information sources might not be similarly effective across a given population. A well-established literature shows that motivations to reach particular conclusions affect information processing (Chaiken, Giner-Sorolla, & Chen, 1996). This suggests that there might be meaningful differences between liberals and conservatives in how information sources might influence beliefs and therefore knowledge (Haidt, Graham, & Joseph, 2009). The first possibility is that people are more sensitive to sources that match their ideology (e.g., a Republican might be more sensitive to information from another Republican than from a Democrat, and vice-versa). This possibility is supported by prior work showing that perceived norms are most influential when they arise from others with whom we share a common identity (Abrams, Wetherell, Cochrane, Hogg, & Turner, 1990; Centola, 2011). The second possibility involves a differentiation between liberals and conservatives, such that conservatives might be more resistant to change than liberals, as has been shown before (Jost, Glaser, Kruglanski, & Sulloway, 2003; White et al., 2020). This is also consistent with a recent study, in which Republicans tended to be less concerned about COVID-19 and less likely to share accurate information about COVID-19 than Democrats (Pennycook et al., 2020). This is perhaps not surprising given that Republican leaders such as President Trump and conservative media outlets such as Fox News have expressed skepticism regarding the risk posed

by the virus (Montanaro, 2020; Ballhaus, Armour, & Leary, 2020; Grynbaum & Abrams, 2020). In line with this messaging, a Pew poll conducted in March 2020 estimated that most (59%) Democrats but only a minority (33%) of Republicans viewed COVID-19 as a major threat. The third possibility is that ideology might not interact at all with information sources when it comes to beliefs. This possibility is supported by recent research showing that accurate beliefs about COVID-19 are broadly associated with reasoning skills regardless of political ideology (Pennycook et al., 2020). To investigate, we designed an experiment in which participants first rated the accuracy of a set of statements about COVID-19 (accurate information and conspiracies; pre-test). Then, in the second phase, they were randomly assigned to one of 10 between-subjects conditions in which we varied the source that provided belief-relevant information: a political leader (President Trump, President Biden), a health authority (Doctor Fauci, the CDC), an anecdote (of a Democrat or of a Republican), a large group of prior participants portrayed as being either Democrats (Democratic Normative), Republicans (Republican Normative), or with no ideological designation (Generic Normative). In the Control Condition, participants skipped the second phase entirely. Importantly, the source always endorsed accurate information and denied conspiracies. Therefore, trusting the source and incorporating their message would always increase scientific knowledge. Finally, participants rated the accuracy of the initial set of statements again (post-test). This experimental design has numerous strengths. First, it accomplishes a horse-race comparison between different sources using the same materials. Second, it uses a US census-matched sample to increase the generalizability of the results. And third, it is conducted during a real-time health crisis, therefore creating an ecologically valid context of investigation. Finally, we replicate this experiment in Study 5b, with different stimulus materials (i.e., regarding the Covid-19 vaccine), an increased sample size, and an additional measure of intent to get vaccinated against COVID-19. Since we are interested in establishing

the most efficient source to increase knowledge, this investigation is mainly exploratory. That said, two main hypotheses were formulated based on prior literature. Our first hypothesis was that participants in the Generic Normative Condition will change their beliefs in line with the source, therefore increasing in knowledge compared to the Control Condition. Second, we hypothesized a partisan bias in belief change in the form of an interaction between participant and source ideology, such that participants will change their beliefs in line with the source more, when the source matches their ideology. In other words, Republicans will be more sensitive to Republican sources, whereas Democrats will be more sensitive to Democratic sources.

Study 5a

Method

Open science practices. The materials and data can be found on our open science framework page:
https://osf.io/zcp3m
The pre-registrations can be found here:
Study 5a: https://aspredicted.org/blind.php?x=wg3aa5
Study 5b: https://aspredicted.org/kb4bc.pdf
The data analysis (in Python) can be accessed as a jupyter notebook on Github: https://github.com/mvlasceanu/COVIDsource

Participants. We aimed for a US census matched sample of 1000 participants, half Democrats and half Republicans. This sample size was calculated based on a power analysis including an effect size of 0.4, a significance level of 0.05, and 80% power, for each of the independent sample comparisons between the Control and Experimental conditions. Using the Cloud Research platform, we recruited a US

census matched sample of 1387 Americans, expecting, based on prior studies, to exclude 25% of them based on pre-registered criteria (i.e., failed attention checks). And indeed, 327 participants failed our attention checks. We conducted statistical analyses on the final US census matched sample of 1060 participants (57% female; Mage=48.30, SDage=16.89). This sample matched the US census quotas of age, gender, race, and ethnicity. The total sample contains 544 participants self-identified as Democrats and 516 self-identified as Republicans. The study protocol was approved by the Princeton University Institutional Review Board.

Stimulus materials. We undertook preliminary studies to develop a set of 8 statements regarding COVID-19. A pilot study was conducted on a separate sample of 269 Cloud Research participants (Mage=40.63, SDage=15.49; 66% women) to select these statements from a larger initial set of 37 statements. For each of these statements we collected believability ratings (i.e., "How accurate or inaccurate do you think this statement is" on a scale from 0-Extremely Inaccurate to 100-Extremely Accurate). The 8 statements we selected were on average moderately endorsed (M=53.03, SD=21.57, on a 0 to 100-point scale), as we chose them to avoid ceiling and floor effects. Four of them are scientifically accurate (MAccurateBeliefs=71.1, SD=29.8) and 4 are conspiracies (MConspiracyBeliefs=34.9, SD=34.4), as concluded by published scientific papers and/or by the Centers for Disease Control, at the time of data collection.

Design and procedure. The data for this study was collected between May 26, 2020 and June 4, 2020. The 1060 participants went through three experimental phases. Participants were told they would participate in an experiment about people's evaluation of information and were directed to the survey on the Qualtrics platform. After completing the informed consent form, participants were directed to the first phase (pre-test), in which they rated a set of 8 statements (one on each page) by indicating the degree to which they believed each statement (i.e., "How accurate

do you think this statement is," from 1-Extremely inaccurate to 100-Extremely accurate). Then, in the second phase, participants were randomly assigned to one of 10 between-subjects conditions. For each of the 10 conditions, participants were told the source of half (i.e., target items) of the initially rated statements was one of the following: a political leader (President Trump or President Biden), a health expert (Doctor Fauci or the CDC), an anecdote (of a Democrat or a Republican), or a group of prior participants (either Democrats, Republicans, or Generic non-ideological). Importantly, the source always endorsed the accurate information they mentioned (e.g., "This statement was part of a speech by President Trump"; 2 target items, counterbalanced with baseline items) and denied the conspiracies they mentioned (e.g., "This statement was refuted in a speech by President Trump"; 2 target items, counterbalanced with baseline items). Note that for the normative conditions (i.e., involving supposed groups of prior participants) participants were instead told they would be able to see the average accuracy rating assigned to half (i.e., target items) of the initial statements by prior participants while qualifying their political ideology as either Republican (i.e., "You will now see the average accuracy assigned to some of these statements by the Republican participants who took this survey last week"; Normative Republican), Democratic (i.e., "You will now see the average accuracy assigned to some of these statements by the Democratic participants who took this survey last week"; Normative Democratic), or Generic (i.e., "You will now see the average accuracy assigned to some of these statements by the participants who took this survey last week"; Generic-Normative). Importantly, in each of these 3 conditions, the average ratings presented to participants for the target items were very high for accurate information (e.g., "95%", "98%") and very low for conspiracies (e.g., "5%", "2%"). The non-target items, which consist of half of the items participants were presented with in the pre-test phase, were considered baseline items. We note that the 8 initial items were pseudo-randomly assigned to either

a target or a baseline status across participants, such that the source supported two accurate beliefs and opposed two conspiracy beliefs. In the Control Condition participants were not presented with any information at all, they only completed the pre-test and post-test. In the third phase (post-test) participants rated again the believability of the initial 8 statements, after which they were asked to complete a series of demographic information and were debriefed.

Measures. Statement endorsement was measured at pre-test and post-test with the question "How accurate or inaccurate do you think this statement is", on a scale from 0-Extremely Inaccurate to 100-Extremely Accurate. We asked participants to indicate their age, gender, education, and political orientation.

Analysis and coding. Participants' knowledge about COVID-19 was operationalized and computed as the difference between their belief in the accurate information and the conspiracies (i.e., belief in accurate minus belief in inaccurate statements). This score was calculated separately for the target items (i.e., statements the source mentioned) and the baseline items (i.e., statements the source omitted). Knowledge change (or accumulation) was computed as participants' knowledge at post-test minus the knowledge at pre-test.

Results

First, we ran a between-subjects ANOVA with change in knowledge as the dependent variable and condition as the between-subjects variable, and found a significant main effect of Condition $F(9, 1050)=10.08$, $p<0.001$, $\eta_p^2=0.08$ (Figure 5.5).

To test our first hypothesis, that participants in the Generic-Normative Condition will change their beliefs in line with the source, therefore increasing in knowledge compared to the Control Condition, we conducted an independent sample t-test and found that, as hypothesized, participants in the Generic-Normative Condition (M=13.22, SD=22.17) increased their knowledge more than participants in the Control Con-

dition (M=1.42, SD=9.20) t(137)=5.01, p<0.001, Cohen's d=0.69, CI [7.14, 16.45]. Additionally, we conducted independent sample t-tests assessing the differences in knowledge change between the Control Condition and all other conditions. We found that the Normative Democratic, Normative Republican, Fauci, CDC, and Trump (in the opposite direction) Conditions were significantly different from the Control Condition (Figure 5.5; statistics reported in Table 5.5). We note that the significance level was adjusted for multiple comparisons (i.e., 9 comparisons, significance threshold p<0.0055) using the Bonferroni correction.

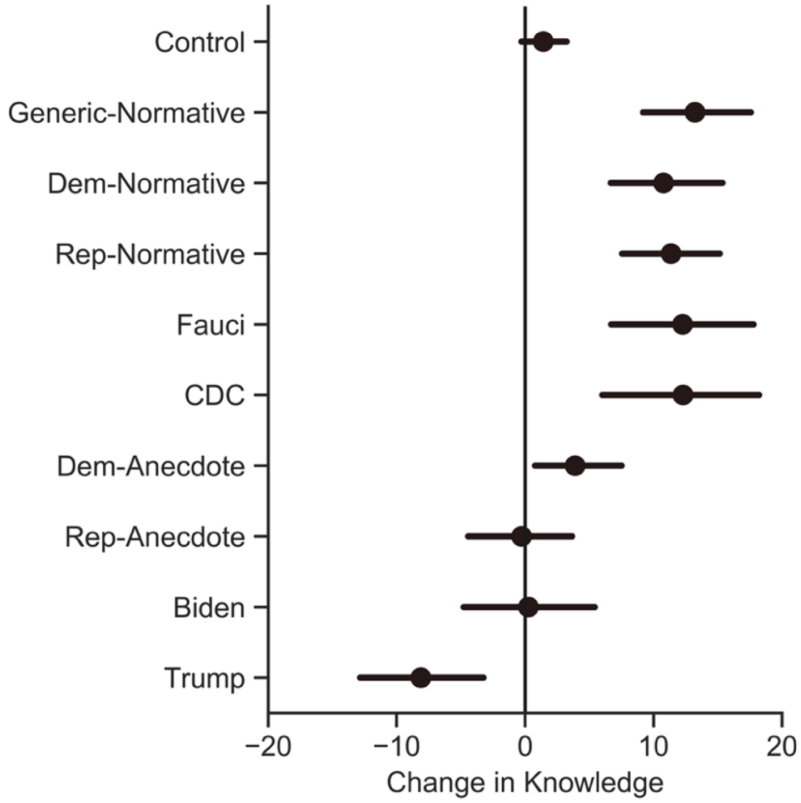


Figure 5.5: Change (post-test minus pre-test) in knowledge (belief in accurate information minus belief in conspiracy theories) for the target items, in each of the 10 between-subject conditions. Error bars represent ±1 standard errors of the mean.

To investigate our second hypothesis, of a partisan bias in knowledge change in the form of an interaction between participant and source ideology, we ran a between-subjects ANOVA with change in knowledge as the dependent variable, condition and

Condition	*M*	*SD*	*df*	*t*	*p*	*Cohen's d*	*CI*
Generic Normative	13.22	22.17	137	5.01	<0.00001*	0.69	[7.1, 16.4]
Democratic Normative	10.77	23.13	152	3.99	<0.0002*	0.52	[4.5, 14.1]
Republican Normative	11.36	20.18	150	4.64	<0.00001*	0.62	[5.6, 14.2]
Doctor Fauci	12.27	28.38	118	3.64	<0.0003*	0.51	[5.0, 16.6]
CDC	12.30	31.08	124	3.46	<0.0008*	0.47	[4.5, 17.1]
Democratic Anecdote	3.89	19.30	161	1.18	=0.236	0.16	[-1.6, 6.5]
Republican Anecdote	-0.27	21.11	139	-0.75	=0.453	0.10	[-6.1, 2.7]
Vice-President Biden	0.26	26.80	128	-0.42	=0.675	0.05	[-6.6, 4.3]
President Trump	-8.10	25.96	125	-3.49	<0.0006*	0.49	[-14.8, -4.1]

Table 5.5: Statistics of 9 independent sample t tests comparing all experimental conditions to the Control Condition. "*" marks the significant comparisons at the p=0.0055 significance level (p-value adjusted for 9 comparisons).

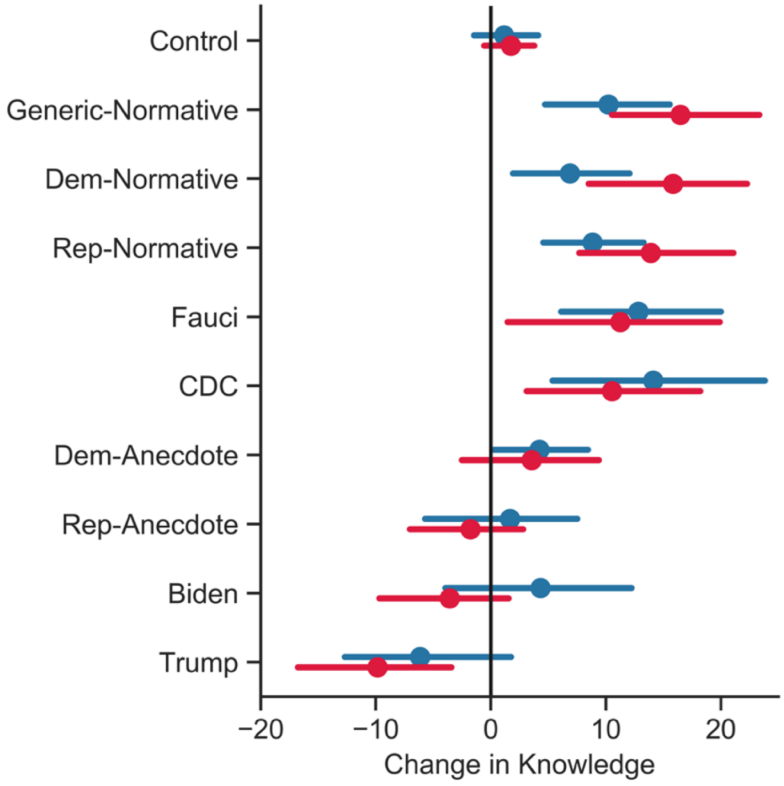


Figure 5.6: Change (post-test minus pre-test) in knowledge by participant type (Democrats in Blue vs. Republicans in Red), in each of the 10 between-subjects conditions. Error bars represent ±1 standard errors of the mean.

participant ideology (Democrats vs. Republicans) as the between-subject variable. We found a main effect of condition $F(9, 1038)=9.99$, $p<0.001$, $\eta_p^2=0.08$, but not of

participant ideology F(1, 1038)=0.004, p=0.948, η_p^2=0.00, and no interaction between the two variables F(9, 1038)=1.35, p=0.205, η_p^2=0.01 (Figure 5.6). This suggests Democrats and Republicans are similarly affected by COVID-19 information sources.

Discussion

Using a horse race experimental design and a US census matched sample, we found that individuals' COVID-19 knowledge increased compared to a Control Condition when information was provided by large groups of people (Democrats, Republicans, Generic) and health authorities (Doctor Fauci and the CDC), but not when provided by political leaders (Trump, Biden) or anecdotes. We did not find ideological differences in the knowledge integration. Intriguingly, not only did our participants not update beliefs based on information from political leaders, when the source of information was President Trump, they displayed a backfire effect, such that they changed their initial beliefs away from whatever President Trump had conveyed. Given our study design, in which all the sources supported accurate information and refuted conspiracy theories, by moving away from his message, participants decreased their level of COVID-19 knowledge. Of all ten sources tested in this study, this was the only condition in which knowledge decreased from pretest to posttest, pointing to a general skepticism towards any COVID-19 information coming from President Trump.

To ensure the generalizability and replicability of these findings, in Study 5b we investigated these effects in the context of the COVID-19 vaccine. We increased the sample size to increase the power of detecting potential interactions with participants' political affiliation. Finally, in Study 5b, we were also interested in whether vaccine-related knowledge accumulation would predict vaccination intention.

Study 5b

Participants. To increase the power of detecting potential ideological differences in the effect observed in Study 5a, we now calculated the sample size based on the power analysis of each of the independent sample comparisons between Democratic and Republican participants in each condition, at an effect size of 0.4, a significance level of 0.05, and 80% power. Thus, we aimed for a sample of 2000 participants, which is the sample size we pre-registered. Using the Cloud Research platform, we recruited a sample of 2075 Americans. Of these, 199 were excluded based on pre-registered criteria (i.e., failed attention checks). We conducted statistical analyses on the final sample of 1876 participants (61% female; Mage=49.24, SDage=18.02). The total sample contains 911 participants self-identified as Democrats and 965 self-identified as Republicans. The study protocol was approved by the Princeton University Institutional Review Board.

Stimulus Materials. We developed a set of 8 statements regarding COVID-19 vaccines. Four of them are scientifically accurate and four are inaccurate, as concluded by published scientific papers and/or by the Centers for Disease Control, at the time of data collection.

Design and Procedure. The data for this study was collected between February 1st and February 2nd, 2021. The design and procedure were the same as in Study 5a, with one exception. At the end of this study, we asked participants' intention to get vaccinated against COVID-19.

Measures. We used the same measures as in Study 5a, with the addition of a measure of intention to get vaccinated against COVID-19. First, participants were asked if they had already been vaccinated against COVID-19. If their answer was "no", then the follow-up question appeared on their screen: "If you were offered the CDC currently recommended COVID-19 vaccine (Moderna or Pfizer) today, would

you agree to get vaccinated?" which they had to answer on a scale from 0-"Absolutely not" to 100-"Absolutely yes".

Analysis and coding. Just like in Study 5a, we operationalized knowledge about COVID-19 vaccines as the difference between participants' belief in the accurate and inaccurate information (i.e., belief in accurate minus belief in inaccurate statements). Knowledge change was computed as participants' knowledge at post-test minus the knowledge at pre-test.

Results

As in Study 5a, we began our analyses by running a between-subjects ANOVA with change in knowledge as the dependent variable and condition as the between-subjects variable, and found a significant main effect of Condition $F(9, 1866)=2.088$, $p=0.027$, $\eta_p^2=0.01$ (Figure 5.7). To test our first pre-registered hypothesis, that participants in the Generic-Normative Condition will change their beliefs in line with the source, therefore increasing in knowledge compared to the Control Condition, we conducted an independent sample t-test and found that, as hypothesized, participants in the Generic-Normative Condition ($M=11.78$, $SD=24.04$) increased their knowledge more than participants in the Control Condition ($M=3.92$, $SD=13.02$) $t(330)=4.06$, $p<0.001$, Cohen's $d=0.39$, CI [3.86, 11.85], replicating the result we found in Study 5a. Additionally, we conducted independent sample t-tests assessing the differences in knowledge change between the Control Condition and all other conditions. This time, none of these conditions were significantly different from the Control Condition when adjusting the significance level for multiple comparisons (i.e., 9 comparisons, significance threshold $p<0.0055$) using the Bonferroni correction (statistics reported in Table 5.6). We note that the pattern of results is similar between the two studies, but the differences do not reach corrected statistical significance levels in Study 5b,

mainly because participants in the Control Condition now increased in knowledge from pretest to posttest.

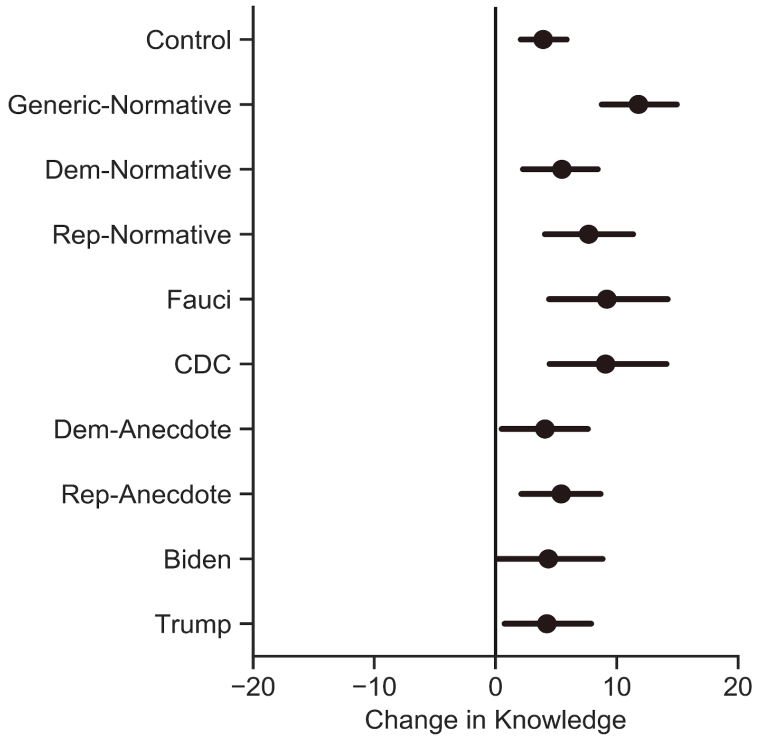


Figure 5.7: Change (post-test minus pre-test) in knowledge (belief in accurate information minus belief in conspiracy theories) for the target items, in each of the 10 between-subject conditions. Error bars represent ±1 standard errors of the mean.

Condition	*M*	*SD*	*df*	*t*	*p*	*Cohen's d*	*CI*
Generic Normative	11.78	24.04	330	4.06	=0.00006**	0.39	[3.86, 11.85]
Democratic Normative	5.47	22.74	322	0.81	=0.413	0.08	[-2.29, 5.38]
Republican Normative	7.68	23.83	271	1.83	=0.068	0.19	[-0.28, 7.79]
Doctor Fauci	9.18	33.92	224	1.91	=0.057	0.20	[-0.16, 10.66]
CDC	9.07	34.86	244	1.89	=0.059	0.19	[-0.37, 10.66]
Democratic Anecdote	4.07	23.62	295	0.07	=0.940	0.01	[-3.83, 4.12]
Republican Anecdote	5.41	23.37	274	0.73	=0.461	0.07	[-2.48, 5.46]
President Biden	4.35	32.56	277	0.17	=0.862	0.01	[-4.73, 5.59]
President Trump	4.23	24.40	278	0.14	=0.882	0.01	[-3.79, 4.40]

Table 5.6: Statistics of 9 independent sample t tests comparing all experimental conditions to the Control Condition. "*" marks the significant comparisons at the p=0.0055 significance level (p-value adjusted for 9 comparisons).

To investigate our second hypothesis, of a partisan bias in knowledge change in the form of an interaction between participant and source ideology, we ran a between-subjects ANOVA with change in knowledge as the dependent variable, condition and participant ideology (Democrats vs. Republicans) as the between-subject variable. We found a main effect of condition $F(9, 1856)=2.08$, $p=0.027$, $\eta_p^2=0.01$, but not of participant ideology $F(1, 1856)=1.62$, $p=0.20$, $\eta_p^2=0.001$, and no interaction between the two variables $F(9, 1856)=0.68$, $p=0.724$, $\eta_p^2=0.003$ (Figure 5.8). This suggests Democrats and Republicans are similarly affected by COVID-19 information sources, replicating the finding in Study 5a.

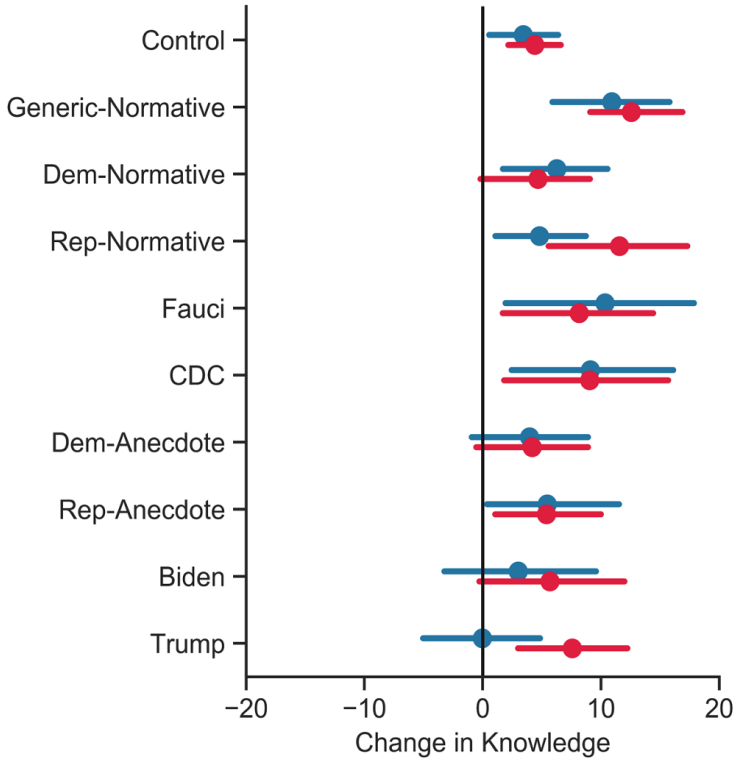


Figure 5.8: Change (post-test minus pre-test) in knowledge by participant type (Democrats in Blue vs. Republicans in Red), in each of the 10 between-subjects conditions. Error bars represent ±1 standard errors of the mean.

Next, we wanted to explore the relation between vaccine knowledge and vaccination intention. Given that 13% of participants had already been vaccinated, we excluded them from the following analyses. First, we ran a linear mixed model

with intention to get vaccinated as the dependent variable, fitting pretest vaccine knowledge as well as participant ideology as fixed effects, and included by-participant random intercepts. We found a significant interaction between Democrats and Republicans (β=0.10, SE=0.04, t(1629)=2.33, p<0.019) in how their pretest knowledge predicted vaccination intention, such that this effect was stronger for Democrats (β=0.75, SE=0.02, t(1630)=29.43, p<0.001) than for Republicans (β=0.61, SE=0.02, t(1630)=22.79, p<0.001) (Figure 5.9A). Given the result that more knowledge about the vaccine is associated with a stronger intention to get vaccinated, next, we wanted to assess whether integrating knowledge also results in an additionally stronger intention to get vaccinated. We ran a linear mixed model with intention to get vaccinated as the dependent variable, fitting vaccine knowledge change as well as participant ideology as fixed effects, and including by-participant random intercepts. We found that change in vaccine knowledge (i.e., knowledge accumulation) predicts intention to get vaccinated positively for Democrats (β=0.15, SE=0.05, t(1630)=2.88, p=0.003) and negatively for Republicans (β=-0.14, SE=0.05, t(1630)=-2.76, p=0.005), but we found no significant interaction between the two political ideologies (β=0.06, SE=0.07, t(1629)=0.872, p=0.384) (Figure 5.9B).

Moreover, we wanted to assess whether knowledge accumulation predicts intention to get vaccinated, differently, depending on the source of the information. We were interested in whether participants are more likely to get vaccinated if the source of the knowledge accumulation is a generic normative group, an ideologically consistent source (e.g., for Democrats: the group of Democrats, the anecdote of a Democrat, President Biden), an ideologically inconsistent source (e.g., for Democrats: the group of Republicans, the anecdote of a Republican, or President Trump), or a health expert. We ran a linear mixed model with intention to get vaccinated as the dependent variable, fitting change in knowledge as well as participant ideology, and the collapsed conditions as fixed effects, and included by-participant random

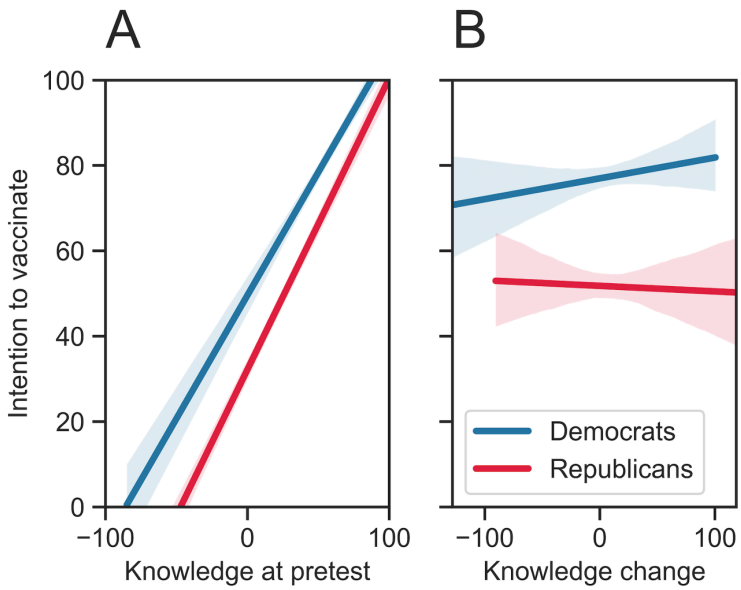


Figure 5.9: Intention to get vaccinated against COVID-19 as predicted by vaccine knowledge at pretest (Panel A) and by vaccine knowledge change (Panel B), split by participant type (Democrats in Blue vs. Republicans in Red). Note that knowledge change can be negative (in Panel B) if participants decrease in knowledge from pretest to posttest. Shaded regions represent 95% confidence intervals.

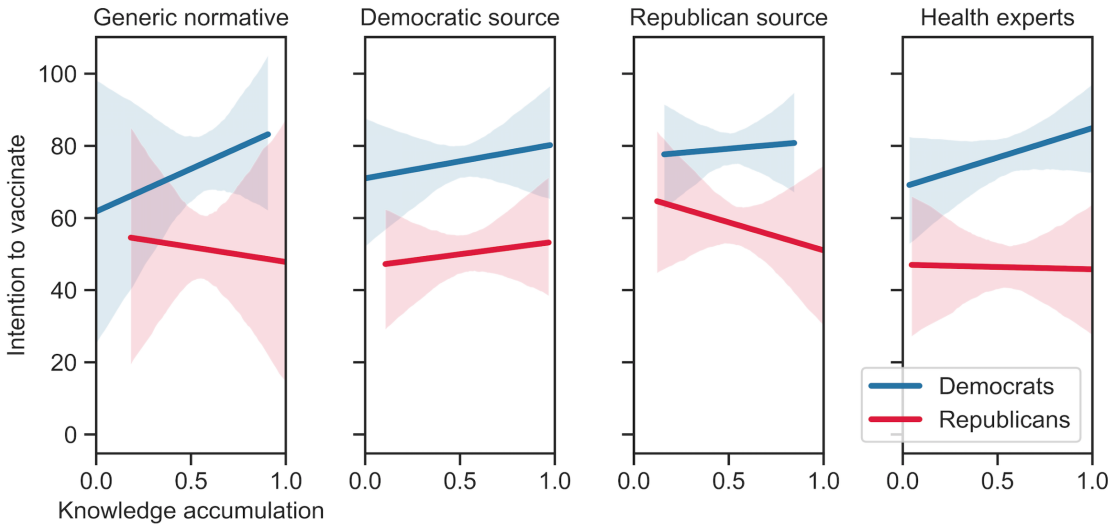


Figure 5.10: Vaccination intention predicted by change in knowledge (i.e., knowledge accumulation) for Democrats (in blue) and Republicans (in red) in the collapsed conditions: generic normative sources, Democratic sources, Republican sources, and health experts.

intercepts. We found that knowledge change predicts vaccination intention only for Democratic participants in the health experts condition (β=0.18, SE=0.08, t(1622)=2.10, p=0.035) (Figure 5.10). Even though, for simplicity, in this analysis we collapsed the conditions in this manner, in the extended mixed model we find that knowledge change predicts vaccination intention only for Democratic participants in the Doctor Fauci condition (β=0.33, SE=0.13, t(1612)=2.51, p=0.012).

Discussion

Using different informational content, we replicated the finding that normative cues that involve generic sources are most impactful at increasing scientific knowledge. We also replicated the finding that political ideology does not impact knowledge integration. In contrast to Study 5a, we now found that no other source of information led to more knowledge accumulation compared to the control condition, given that participants in the control condition increased in knowledge from pretest to posttest, perhaps as a result of increased epistemic vigilance the second time they rated the statements (Sperber et al., 2010). This difference may also have been due to the difference in participants' initial levels of knowledge between the two studies, as we found that participants in Study 5b were more knowledgeable at pretest than participants in Study 5a. Further investigations are however necessary to pinpoint the specific mechanism leading to this difference.

In addition to Study 5a, Study 5b included a measure of intention to get vaccinated against COVID-19. We found that initial vaccine knowledge is a strong predictor of vaccination intention for both Democrats and Republicans. We also found that the more information Democrats accumulated, the more likely they were to want to be vaccinated, but the more information Republicans accumulated, the less likely they were to want to be vaccinated. We speculate that Republicans' lower initial level of knowledge about COVID-19 vaccines compared to Democrats'

may have contributed to this effect. When further investigating these findings, while also taking into account the source of the information, we found that information accumulation increases vaccination intention only for Democrats in the health experts condition. More specifically, they were most likely to increase in vaccination intention when Doctor Fauci was the source of their information accumulation.

General Discussion

In two studies we used an experimental horse-race design to assess which information source is most likely to lead to COVID-19 and COVID-19 vaccine knowledge incorporation. We found that individuals' knowledge increases most when information is provided by a generic group of people. This finding aligns with seminal work showing that social norms have a meaningful impact on people's attitudes and behaviors (Cialdini & Goldstein, 2004; Paluck & Green, 2009), especially under uncertainty (Cialdini & Trost 1998). Here, we show that portraying information as normative (i.e., highly endorsed by others) increases people's belief in its accuracy, and portraying information as counter normative (i.e., highly opposed by others) decreases people's belief in its accuracy. However, several sources (e.g., the ideological groups, the health experts) emerged as impactful in Study 5a, but not Study 5b. The pattern looks similar between the two studies, when looking at the degree of each source's impact, but an investigation into the context in which these sources have a weaker versus a stronger effect is certainly warranted.

Also in both studies, we find that knowledge accumulation occurs similarly for Democrats and Republicans. This result diverges from previous studies which found that norms are most influential when they arise from others with whom people share a common identity (Abrams et al., 1990; Centola 2011). We also expected ideological differences in knowledge integration from congruent versus incongruent sources based on prior work showing that conservatives are more resistant to change than liberals

(Jost et al., 2003; White et al., 2020) and that Republicans are less concerned about COVID-19 than Democrats (Pew, 2020). However, in the present work we did not find ideological differences in knowledge integration, consistent with prior work in which Democrats and Republicans updated their beliefs similarly as a function of evidence (Vlasceanu, Morais, & Coman, 2021). Along the same lines, Pennycook and colleagues found that accurate beliefs about COVID-19 are associated with reasoning skills regardless of political ideology (Pennycook et al., 2020). One possible explanation for the lack of ideological differences could be that when stakes are high, people might moderate their ideological biases. Another possibility is that ideological differences do not surface in a short-term context and only become apparent over longer periods of time as novel information integrates with preexisting ideological schemas. This possibility is consistent with prior work in the persuasion literature, such as the sleeper effect (for a meta-analysis see Kumkale & Albarracín, 2004). However, these speculative explanations should be empirically tested in future work. The lack of an effect of political leaders on belief change is also in line with prior work showing that ideological messages are not effective when communicated by leaders (Grant & Hofmann, 2011). Interestingly, we also found no effect of anecdotes on knowledge change. This is surprising, as prior work has already established the effectiveness of anecdotal evidence in persuasion (Slater & Rouner, 1996). One explanation for our null finding may be that the conditions of stress and uncertainty caused by the COVID-19 pandemic, might have reduced the impact of single voices as persuasive sources. To confirm this assumption, one could programmatically manipulate perceived threat and observe the impact of anecdotal evidence on knowledge change, a direction we deem worthwhile pursuing.

When it comes to translating the knowledge accumulation to a concrete behavior (i.e., vaccination intention), we found that Democrats were most likely to increase their vaccination intention when Doctor Fauci was the source of their knowledge

increase. This finding is consistent with prior work showing that credible sources have more influence on people's beliefs (Begg, Anas, & Farinacci, 1992; Chung, Fink, & Kaplowitz, 2008; Slater & Rouner, 1996) and intentions such as voting (Mondak, 1995) and purchasing behavior (Lafferty & Goldsmith, 1999; Till & Busler, 1998). Here, we replicate this finding in the context of vaccination. This finding is very informative from a policy perspective, as it suggests that short-term interventions to impact behavioral intentions involving knowledge have limited efficiency, even if based on ideologically congruent sources. For Republicans, other types of interventions (e.g., non-knowledge based) would need to be created and tested.

A meaningful expansion of this work could involve the investigation of how information from different sources propagates through social networks (Vlasceanu, Enz, & Coman, 2018). Once these sources communicate information in real-world circumstances, people often communicate with one another and share this information (Liu, Jin, & Austin, 2013). Thus, it would be important to understand how these conversations amplify the impact of the source, especially in homogeneous communities, given homophily characteristics. Critically, individual level effects have been found to propagate in social networks (Coman, Momennejad, Duker, & Geana, 2016; Vlasceanu, Morais, Duker, Coman, 2020; Vlasceanu & Coman, 2020), and social networks can amplify the spread of behaviors that are both harmful and beneficial during an epidemic (Christakis & Fowler, 2013). Thus, tracking COVID-19 information propagation in fully mapped social networks would be critically important, especially given policymakers' interests in impacting community-wide knowledge and behavior (Dovidio & Esses, 2007).

Finally, these findings might prove useful in the battle against misinformation, a prominent threat facing the world today (Lewandowsky et al., 2012). Emerging research is using social science to understand and counter the spread of false information (Guess, Nagler, & Tucker, 2019). One approach is refutation, or debunking

(Berinsky, 2017; Lewandowsky et al., 2012), which has been found to backfire in some contexts (Porter, Wood, & Kirby, 2018; Swire & Ecker, 2018). Another approach is prebunking, or inoculating (van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017) by preemptively exposing people to small doses of misinformation techniques (including scenarios about COVID-19) which reduce susceptibility to fake news (Basol, Roozenbeek, & van der Linden, 2020; Roozenbeek & van der Linden, 2019). A third approach is nudging accuracy which has been found to reduce belief in false news (Pennycook & Rand, 2019). Here, we show that generic normativity cues are most successful at increasing knowledge across the ideological spectrum, and that expert communications are most successful at increasing Democrats' vaccination intentions.

The processes I identified in the first 5 Chapters as effective at changing beliefs could readily be used as tools in the design of strategies of combating misinformation. However, scaling these strategies from individuals to communities may prove a more efficient route of attacking the misinformation epidemic we are facing globally (Vlasceanu, Enz, Coman, 2018). To this end, we first need to understand how individual-level processes that impact beliefs affect not just individuals, but also how they are impacted by dyadic conversations, and how they scale up in larger social networks.

Part III

Collective Beliefs

Chapter 6

From Individuals to Dyads

6.1 Study 6: Dyadic Interactions Trigger Belief Change

Abstract

In a high-risk environment, such as during an epidemic, people are exposed to a large amount of information, both accurate and inaccurate. Following exposure, they typically discuss the information with each other in conversations. Here, we assessed the effects of such conversations on their beliefs. A sample of 126 M-Turk participants first rated the accuracy of a set of COVID-19 statements (pre-test). They were then paired and asked to discuss either any of these statements (low epistemic condition) or only the statements they thought were accurate (high epistemic condition). Finally, they rated the accuracy of the initial statements again (post-test). We did not find a difference of epistemic condition on belief change. However, we found that individuals were sensitive to their conversational partners, and changed their beliefs according to their partners' conveyed beliefs. This influence was strongest for initially moderately held beliefs. In exploratory analyses, we found that pre-test COVID-19 knowledge was predicted by trusting Fauci, not

trusting Trump, and feeling threatened by COVID-19. Conversely, pre-test COVID-19 conspiracy endorsement was predicted by trusting Trump, not trusting Fauci, news media consumption, social media usage, and political orientation. In further exploration of the political orientation predictor, we found that Democrats were more knowledgeable than Republicans, and Republicans believed more conspiracies than Democrats.

Introduction

With the rise of globalization, infectious diseases have proven more and more far-reaching (Saker, Lee, Cannito, Gilmore, Campbell-Lendrum, 2004). There is hardly a year without the emergence of a highly threatening pandemic, from H1N1 (swine flu) in 2009, to Ebola in 2014, to the Zika virus in 2017, and to COVID-19 today. Fighting an epidemic involves not only developing effective treatments and ensuring wide distribution, but also informing the public of the symptoms, protective measures, and treatments associated with the disease. It becomes critically important, thus, to understand how information is transferred from medical labs to policy makers, program planners, and the lay public (Green, Ottoson, García, Hiatt, 2009).

Although the diffusion of information in social networks has been studied in a variety of ways (Watts, 2004), very few models incorporate psychologically grounded assumptions about information search, acquisition, and propagation. One relevant area of research in this domain is the recent literature on social aspects of memory, which provides insights into how knowledge acquisition is affected by conversations (Hirst, & Echterhoff, 2012). According to this research, the communicative act of jointly recalling information shapes the memory of both the transmitter and the recipient of information (Cuc, Koppel, Hirst, 2007). When a speaker in a social interaction repeats something they already know, their pre-existing memory of that information is strengthened (Blumen & Rajaram, 2008). And stronger memory has been shown

to increase the believability of information (Hasher, Goldstein, Toppino, 1977; Fazio, Brashier, Payne, Marsh, 2015; Vlasceanu & Coman, 2018). Conversely, their pre-existing memory of related but not discussed information is weakened, and a weaker memory has been shown to decrease the believability of information (Vlasceanu & Coman, 2018). These effects are particularly prominent in the case of moderately endorsed pieces of information (Vlasceanu & Coman, 2018). Thus, one strategy to diminish the believability of misinformation and to encourage the dissemination of accurate information might be to impose a higher threshold for both communicating and accepting information. This higher threshold could be created through instructions involving high epistemic accuracy, such as encouraging people to question the veracity of information before communicating it (Lewandowsky, Ecker, Seifert, Schwarz, Cook, 2012). Here I will assess the effectiveness of deploying such a strategy during communicative interactions on knowledge acquisition and transfer.

Moreover, most information diffusion models fail to account for the fact that information search, acquisition, and propagation often happens during public health emergencies. These high risk, uncertain situations creating a state of increased anxiety might facilitate the perfect context for inaccurate information to spread, mainly because people do not have the cognitive resources to assess the veracity of the information they receive (Coman & Berry, 2015). Indeed, a large body of psychological research shows that information is differentially processed by the cognitive system depending on the emotional state of the recipients (Rozin & Royzman, 2001). Contexts high in emotionality result in high information propagation rates (Harber, & Cohen, 2005), in "viral" successes (Berger & Milkman, 2012), and in communicative advantages in dyadic interactions (Nyhof & Barrett, 2001). These findings showcase the impact of emotional states created by high risk and uncertainty contexts on information propagation. Thus, in the current study I am incorporating this factor by conducting the experiment during the COVID-19 pandemic. These prior studies

suffer however from the important limitation that they rarely involve experimental manipulations, which constrains the causal explanations that can be drawn. Here I aim to address this limitation.

Furthermore, when individuals learn about an epidemic, they engage in behaviors aimed at acquiring information, such as turning to news and social media for relevant content (Saker et al, 2004; Frenkel, Alba, Zhong, 2020). They also seek information from experts (Thiriot, 2018) and leaders (Grant & Hofmann, 2011). Recent work has shown for instance that COVID-19 information conveyed by health experts was incorporated in people's beliefs increasing their COVID-19 knowledge, whereas information conveyed by political leaders was not (Vlasceanu & Coman, 2020). Subsequently, after being exposed to such large amounts of crisis-relevant information, people typically discuss the acquired information with each other (Liu, Jin, Austin, 2013). It is therefore critical to understand how accurate and inaccurate health-relevant information is acquired, incorporated, and discussed in high risk environments, as well as to unveil the predictors of endorsing such beliefs.

To investigate the effects of conversational interactions on people's beliefs during a high-risk high-uncertainty environment caused by a global health crisis, I designed an experiment in which participants first rated the accuracy of a set of statements about COVID-19 (accurate, inaccurate, and conspiracies) and filled a series of emotion scales. Then, they were assigned to pairs, and were asked to discuss the statements with each other, in 5-minute dyadic conversations. To manipulate the likelihood of individuals sharing accurate information in their conversational interactions, the instructions encouraged a random subset of the pairs to discuss any piece of information from the study (low epistemic accuracy condition), and the other subset of the pairs to discuss only the pieces of information they were confident were correct (high epistemic accuracy condition). Lastly, participants rated again the believability of the initial statements and their emotions.

The first hypothesis was that participants in the high epistemic accuracy condition would become more knowledgeable than those in the low epistemic accuracy condition, given that the focus of their conversations would be on the accurate rather than the inaccurate information or conspiracies. The second hypothesis was that participants would be sensitive to their conversational partners' beliefs expressed during their conversations and adjust their own beliefs accordingly. Based on previous research (Vlasceanu & Coman, 2018) I also hypothesized this adjustment would be strongest in the case of initially moderately held beliefs. In exploratory analyses for which I did not have a priori hypotheses, I tested the predictors of COVID-19 knowledge and conspiracy beliefs in models that included trust in politicians/experts, media consumption, threat and anxiety, and positive and negative emotions.

Method

Open science practices. The data and stimulus materials can be found on my open science framework page: https://osf.io/sk4dt/ The data analysis (in Python) can be accessed as a jupyter notebook on Github: https://github.com/mvlasceanu/coviddyad

Participants. To detect an effect size of 0.5 with 80% power at a significance level of 0.05 in an independent sample comparison, I recruited a total of 140 participants. Participants were recruited on Amazon Mechanical Turk and were compensated at the platform's standard rate. After discarding participants who failed the pre-established attention checks, the data from the final sample of 126 participants (61% women; Mage=37.84, SDage=11.35) was included in the statistical analyses. The study was approved by the Institutional Review Board at Princeton University.

Stimulus materials. I undertook preliminary studies to develop a set of 22 statements regarding COVID-19. A pilot study was conducted on separate sample of 269 Cloud Research workers (Mage=40.63, SDage=15.49; 66% women) to select these statements from a larger initial set of 37 statements. For each of these

statements I collected believability ratings (i.e., "How accurate or inaccurate do you think this statement is?" on a scale from 0-Extremely Inaccurate to 100-Extremely Accurate). The 22 statements I selected were on average moderately endorsed (M=51.95, SD=20.08, on a 0 to 100-point scale). In reality, 9 of them were actually accurate (e.g., "The sudden loss of smell or taste is a symptom of being infected with COVID-19"), 9 were inaccurate (e.g., "Antibiotics can kill COVID-19"), and 4 were conspiracies (e.g., "COVID-19 was built as an intended bioweapon"), as concluded by published scientific papers and/or by the CDC at the time of data collection.

Design and procedure. The data was collected in May 2020. The 126 participants went through five experimental phases. Participants were told they would participate in an experiment about people's evaluation of information and were directed to the survey on SoPHIE (i.e., Software Platform for Human Interaction Experiments) a platform that allows fluent computer-mediated interactions among participants. After completing the informed consent form, participants were directed to the first phase (pre-test), in which they rated a set of 22 statements (one on each page) by indicating the degree to which they believed each statement (i.e., "How accurate do you think this statement is," from 1-Extremely inaccurate to 10-Extremely accurate). Then, in the second phase (pre-emotions phase), participants were asked to fill a series of emotion scales (see Measures). In the third phase (conversational phase), participants were randomly paired in groups of two, and were instructed to discuss the information from the pre-test phase with another participant, in a 5-minute dyadic conversation. The instructions encouraged a random subset of the pairs to discuss any piece of information from the study (N=58; Low Epistemic Condition; "In what follows you will have a chat conversation with another participant who answered the same questions about COVID-19 like yourself. In this conversation, please discuss the information about COVID-19 I asked about at the beginning of this study. As you mention a piece of information please be as specific as possible so

that your conversational partner can identify what information you are referring to"), and the other subset of the pairs to discuss only the pieces of information they were confident were correct (N=68; High Epistemic Condition; "In what follows you will have a chat conversation with another participant who answered the same questions about COVID-19 like yourself. In this conversation, please discuss the information about COVID-19 I asked about at the beginning of this study. Importantly, only discuss information you believe is true, and correct the other participants if they bring up information you believe is false. As you mention a piece of information please be as specific as possible so that your conversational partner can identify what information you are referring to"). Conversations took the form of interactive exchanges in a chat-like computer-mediated environment in which participants typed their responses. In the fourth phase (post-test) participants rated again the believability of the initial 22 statements. Finally, in the fifth phase (post-emotions) participants rated their emotions again, after which they were asked to complete a demographic questionnaire and were debriefed.

Measures. Statement endorsement was measured at pre-test and post-test with the question "How accurate or inaccurate do you think this statement is?", on a scale from 0-Extremely Inaccurate to 10-Extremely Accurate. The emotion scales included COVID-19 anxiety, measured in the pre-emotion and post-emotion phase with the question "How anxious are you about the COVID-19 pandemic?" on a scale from 0-Not at all to 100-Extremely. Another emotion scale was a shorter version of the Positive and Negative Affect Schedule (PANAS; Watson, Clark, Tellegen, 1988), in which I included 8 emotions: Calm, Tense, Relaxed, Worried, Content, Fearful, Hopeful, Anxious, Lonely. The instructions were: "Read each statement and select the appropriate response to indicate how you feel right now, that is, at this moment. There are no right or wrong answers. Do not spend too much time on any one statement and give the answer which seems to describe your present

feelings best" and participants rated each emotion from 0-Not at all to 5-Extremely. I included the PANAS both in the pre-emotion phase and in the post-emotion phase. In the analyses, I aggregated the 4 positive emotions and the 5 negative emotions to create a measure of PANAS positive emotions and one of PANAS negative emotions. Moreover, dynamic anxiety was measured by the question "Would you say that during the past 6 weeks you have become more or less anxious about the COVID-19 pandemic?" on a scale from 1-Much less anxious to 7-Much more anxious. I only measured dynamic anxiety in the pre-emotions phase. The final emotion I measured was COVID-19 threat, with the question "How threatening is the COVID-19 pandemic?" from 0-Not at all to 10-Extremely. Participants' news media and social media usage was measured as part of the demographic section at the end of the experiment. Engagement with news media was measured with the question "During a regular day the last 2 weeks, how many hours a day have you been watching the following media outlets (approximate to whole number)" on a scale from 0-0 hours to 5-5 or more hours. I included the following media outlets: MSNBC, CNN, FOX, ABC, NBC, CBS, and PBS/NPR. In the analyses, I used both the individual media outlet data as well as the aggregated score of all 7 media outlets. The aggregated score consisted of the sum of all the answers on each of the 7 scales, creating a single news media measure. Similarly, for social media, I asked participants "During a regular day the last 2 weeks, how many minutes a day have you been on the following social media platforms (approximate to whole number)" on a scale from 0-0 minutes to 10-100 minutes or more. The social media platforms I included were Facebook, Instagram, Twitter, and Snapchat. I aggregated these 4 social media platforms in the analyses by summing all the answers on each of these scales to create a single social media measure. Also in the demographic section I measured participants' trust in President Trump with the question "How much do you trust the COVID-19 information provided by President Trump?" and trust in Dr. Fauci with the question

"How much do you trust the COVID-19 information provided by Doctor Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases?" on a scale from 0-Not at all to 10-Extremely. Finally, I asked participants to indicate their age, gender, education, and political orientation.

Analysis and coding. Participants' knowledge about COVID-19 was computed as the difference between their endorsement of the accurate and the inaccurate information. Participants' COVID-19 conspiracy beliefs were analyzed separately. Together, knowledge (belief in accurate minus inaccurate information) and conspiracies are referred to as beliefs. Belief change was computed as participants' statement endorsement at post-test minus endorsement at pre-test. The conversations' content was coded for belief endorsement. This entailed marking the statements that were endorsed or refuted in conversation, or simply not brought up at all. I used a coding rubric by which a mentioned statement was labeled as either strongly endorsed (+3), endorsed (+2), slightly endorsed (+1), not mentioned (0), slightly opposed (-1), opposed (-2), or strongly opposed (-3). For each participant I accounted for both their own input (i.e., self-endorsement, coded from -3 to +3) as well as their conversational partner's input (i.e., partner-endorsement, coded from -3 to +3) for each statement. Ten percent of the data were double coded for reliability (Cohen k ¿0.88), and all disagreements were resolved through discussion between coders.

Results

Low versus High Epistemic Conditions. To test the first hypothesis, that participants in the high epistemic condition would increase more in knowledge compared to those in the low epistemic condition, I ran a repeated measures ANOVA nested by conversational dyads to account for participants' interactions in the conversational phase, with belief change as the dependent variable, condition (low and high epistemic) as the between-subject variable, and belief type (accurate, inaccurate, conspiracy)

as the within-subject variable. I found a non-significant main effect of belief type $F(2, 114)=0.761$, $p=0.47$, $\eta_p^2=0.013$, a non-significant main effect of condition, $F(1, 57)=0.151$, $p=0.699$, $\eta_p^2=0.003$, and a non-significant interaction $F(2, 114)=0.015$, $p=0.955$, $\eta_p^2=0.001$. Thus, I found that participants did not change their beliefs differently in the two epistemic conditions, so participants in the high epistemic condition did not increase more in knowledge than participants in the low epistemic condition as hypothesized.

To investigate differences in the conversational content between the two epistemic conditions, I ran a repeated measures ANOVA nested by conversational dyads with conversational belief endorsement as the dependent variable, condition (low and high epistemic) as the between-subject variable, and belief type (accurate, inaccurate, conspiracy) as the within-subject variable. I found a significant main effect of belief type $F(2, 114)=20.58$, $p<0.001$, $\eta_p^2=0.265$, not of condition, $F(1, 57)=1.00$, $p=0.321$, $\eta_p^2=0.017$, and a significant interaction $F(2, 114)=7.87$, $p<0.001$, $\eta_p^2=0.121$. Post-hoc analyses revealed that accurate statements were endorsed more in conversations in the high ($M=4.05$, $SD=4.27$) than in the low epistemic condition ($M=1.23$, $SD=1.84$), $t(74)=4.66$, $p<0.001$, Cohen's $d=0.88$, CI [1.67, 3.95]. However, inaccurate statements were similarly endorsed in the high ($M=-0.84$, $SD=3.18$) and low ($M=-0.11$, $SD=0.82$) epistemic conditions ($p=0.09$), and conspiracies were also similarly endorsed in the high ($M=-0.48$, $SD=2.17$) and low ($M=0.22$, $SD=3.62$) epistemic conditions ($p=0.18$). Therefore, I found that, as intended, participants endorsed more accurate statements in their conversations in the high epistemic condition (Figure 6.1).

Belief change as a result of conversational interactions. To investigate the second hypothesis, that participants would change their beliefs to align with their conversational partner, I ran a linear mixed model with belief change as the dependent variable, partner conversational endorsement as the fixed effect, and by-participant, by-dyad,

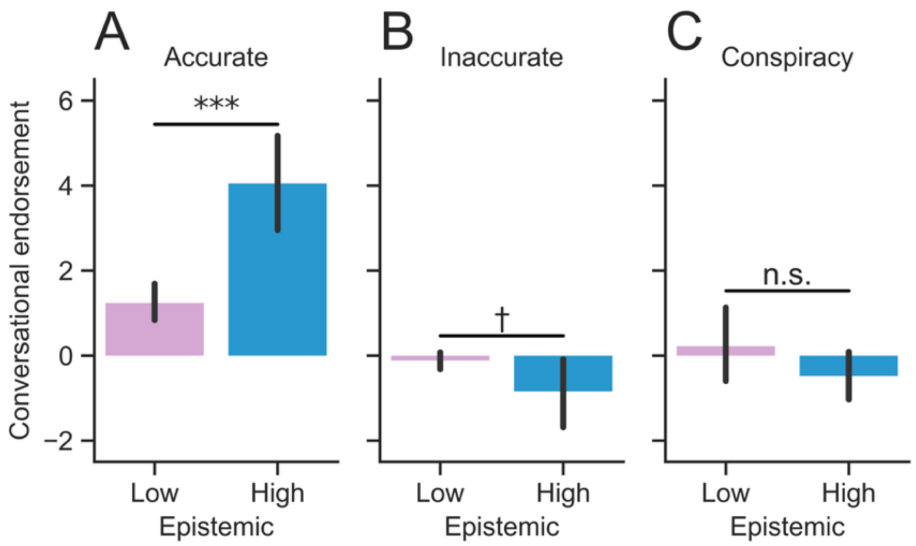


Figure 6.1: Conversational statement endorsement (joint self and partner) in the Low (Pink) and High (Blue) Epistemic Conditions, split by statement type: Accurate (Panel A), Inaccurate (Panel B) and Conspiracy (Panel C). Error bars represent ±1 standard errors of the mean.

and by-item random intercepts. I found that indeed, partners' conversational belief endorsement triggered participants' belief change (β=0.25, SE=0.06, t(2402)=4.00, p<0.001). This relationship remained significant (β=0.22, SE=0.06, t(2545)=3.33, p<0.001) even when controlling for participants' own conversational belief endorsement, by including self-conversational belief endorsement as another fixed effect in the model. Therefore, participants were sensitive to their conversational partners' endorsement of the statements, and changed their beliefs accordingly, such that the more their partner expressed disagreement with a statement in conversation the more the participant decreased their belief in that statement, and the more their partner expressed agreement with a statement in conversation, the more the participant increased their belief in that statement (Figure 6.2A).

Furthermore, to uncover which beliefs were most susceptible to change, I split the 22 statements into 3 categories for each participant, according to their pre-test ratings, as: low endorsement (lowest rated 7 statements), moderate endorsement (middle 8 statements), and high endorsement (highest 7 statements). I then ran a

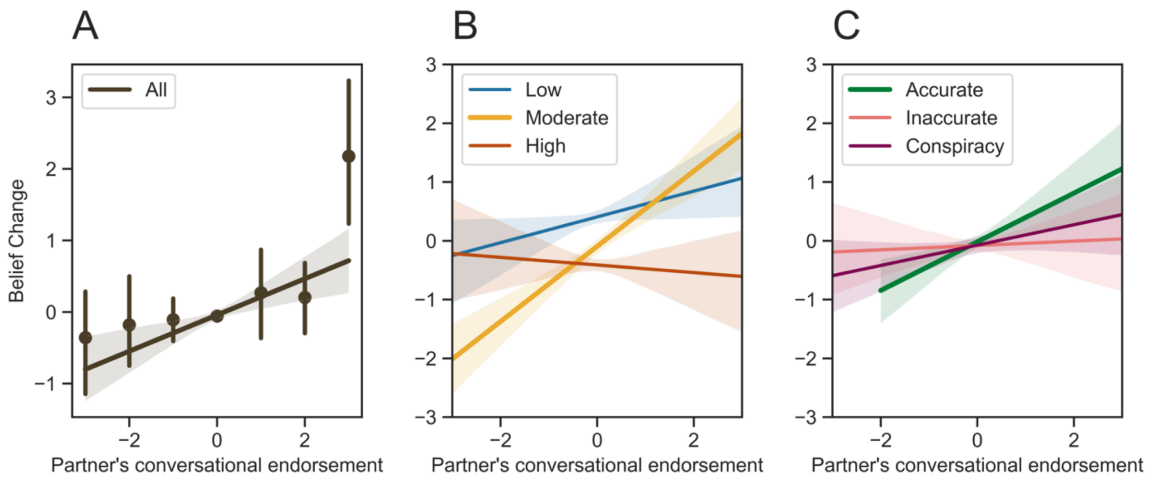


Figure 6.2: Belief update (self) as a function of the conversational partner's belief endorsement as conveyed in the conversational interaction, for all statements (Panel A), for the statements split into Low (blue), Moderate (yellow), and High (orange) endorsed statements at pre-test (Panel B), and for statements split into Accurate (Green), Inaccurate (Red), and Conspiracy (Purple) statements. Error bars represent 95% bootstrapped confidence intervals on the means.

linear mixed model with belief change as the dependent variable, partner conversational endorsement and type (low, mod, high) as fixed effects, with by-participant, by-dyad, and by-item random intercepts. I found that partner conversational endorsement significantly triggered belief change for the initially moderately endorsed statements (β=0.60, SE=0.09, t(2578)=6.35, p<0.001), but not for the initially low (β=0.19, SE=0.12, t(2582)=1.51, p=0.133) or high (β=-0.19, SE=0.11, t(2415)=-1.69, p=0.090) endorsed statements. Therefore, as hypothesized, participants' sensitivity to their conversational partners' statement endorsement was driven by the beliefs they initially moderately endorsed (Figure 6.2B).

Lastly, to further explore which statements were most susceptible to change, I split the 22 statements by their actual accuracy, into accurate (9 statements), inaccurate (9 statements), and conspiracy (4 statements). To investigate whether item type as a function of partner conversational endorsement rendered any differences in belief change, I ran a linear mixed model, with belief change as the dependent variable, partner conversational endorsement and type (accurate, inaccurate, conspiracy) as

fixed effects, and by-participant and by-item random intercepts. I found that partner conversational endorsement significantly triggered belief change for the accurate statements (β=0.42, SE=0.09, t(1862)=4.39, p<0.001), but not for the inaccurate (β=0.04, SE=0.12, t(2581)=0.31, p=0.752) or the conspiracy statements (β=0.19, SE=0.11, t(2588)=1.70, p=0.088). Therefore, participants' sensitivity to their conversational partners' statement endorsement was driven by the accurate statements (Figure 6.2C).

Exploratory analyses. In exploratory analyses, I first tested which variables predicted knowledge, and which predicted believing conspiracies. For knowledge (i.e., belief in accurate information minus belief in inaccurate information), I ran a linear mixed model with knowledge at pre-test as the dependent variable; the fixed effect variables I included were: education level, age, gender, political orientation, trust in Trump, trust in Fauci, news media, social media, COVID-19 threat, COVID-19 anxiety, dynamic anxiety, PANAS positive emotions, PANAS negative emotions. Of these, the significant predictors of knowledge were not trusting Trump, trusting Fauci, and COVID-19 threat (Table 6.1).

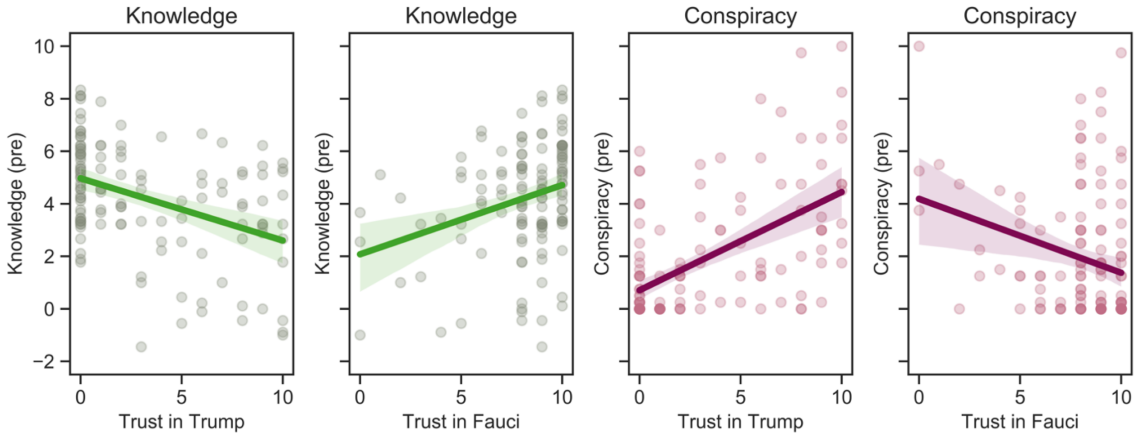


Figure 6.3: Knowledge (Green) and Conspiracy (Purple) at pre-test, as a function of trust in Trump and trust in Fauci.

For conspiracy beliefs, I ran the same linear mixed model, except the dependent variable was conspiracy endorsement at pre-test. The significant predictors of believ-

	β	SE	df	t	p	
(Intercept)	3.06	1.31	112	2.32	0.021	*
Trust in Trump	-0.13	0.06	112	-2.02	0.045	*
Trust in Fauci	0.21	0.08	112	2.64	0.009	**
COVID-19 threat	0.18	0.07	112	2.35	0.020	*
COVID-19 anxiety	-0.02	0.08	112	-0.27	0.784	
Dynamic anxiety	-0.11	0.13	112	-0.85	0.394	
PANAS positive	-0.31	0.20	112	-1.54	0.125	
PANAS negative	-0.35	0.24	112	-1.45	0.149	
News media	-0.04	0.03	112	-1.34	0.182	
Social media	-0.01	0.02	112	-0.43	0.662	
Age	0.02	0.01	112	1.77	0.077	
Education	0.03	0.17	112	-0.20	0.837	
Gender (F)	0.05	0.33	112	0.17	0.858	
Political orientation (D)	-0.01	0.41	112	-0.02	0.981	
Political orientation (R)	-0.16	0.55	112	-0.30	0.760	

Table 6.1: Linear mixed model predicting knowledge at pre-test.

ing conspiracies were trusting Trump, not trusting Fauci, news media, social media, and political orientation (Table 6.2).

To more intuitively display the significant predictors of knowledge and conspiracy beliefs in the two mixed models above, I plotted the regressions of trust in Trump and trust in Fauci (Figure 6.3) as well as news media and social media (Figure 6.4), against knowledge and conspiracy belief.

Given the surprising result that news media consumption predicted conspiracy beliefs, I wanted to investigate whether this effect was driven by a particular news source, or whether it was a general effect of all networks. I ran a linear mixed model with conspiracy endorsement at pre-test as the dependent variable, news media consumption and network (FOX, CNN, ABC, MSNBC, NBC, CBS, PBS) as fixed effects, and by-participant random intercepts. I found no interaction of news consumption

	β	SE	df	t	p	
(Intercept)	1.79	1.19	112	1.50	0.135	
Trust in Trump	0.21	0.05	112	3.68	<0.001	***
Trust in Fauci	-0.28	0.07	112	-3.81	<0.001	***
COVID-19 threat	0.09	0.06	112	1.34	0.180	
COVID-19 anxiety	0.02	0.07	112	0.33	0.740	
Dynamic anxiety	-0.07	0.12	112	-0.65	0.517	
PANAS positive	0.22	0.18	112	1.20	0.229	
PANAS negative	0.14	0.21	112	0.67	0.498	
News media	0.11	0.03	112	3.68	<0.001	***
Social media	0.04	0.02	112	2.01	0.046	*
Age	0.004	0.01	112	0.31	0.755	
Education	-0.11	0.16	112	-0.70	0.482	
Gender (F)	-0.43	0.29	112	-1.45	0.149	
Political orientation (D)	-0.77	0.37	112	-2.08	0.039	*
Political orientation (R)	0.01	0.49	112	0.02	0.982	

Table 6.2: Linear mixed model predicting conspiracy at pre-test.

with network (p=0.9), suggesting that news consumption of any of the 7 networks predicted conspiracy beliefs (Figure 6.5). For the sake of completion, even though not significant, I also plotted the news media consumption of each network as it predicted knowledge (Figure 6.5).

In further exploring ideological differences in people's beliefs, I ran a repeated measures ANOVA with belief at pre-test as the dependent variable, participant ideology (Democrats and Republicans) as the between-subject variable, and item type (knowledge and conspiracy) as the within-subject variable, and found a significant main effect of ideology $F(1, 93)=5.91$, $p=0.017$, $\eta_p^2=0.06$, a significant main effect of type, $F(1, 93)=29.72$, $p<0.001$, $\eta_p^2=0.24$, and a significant interaction $F(1, 93)=20.73$, $p<0.001$, $\eta_p^2=0.182$. Post-hoc analyses suggested that Democrats (M=4.58, SD=2.08) were more knowledgeable than Republicans (M=3.15, SD=2.15), $t(55)=3.03$, $p=0.0037$,

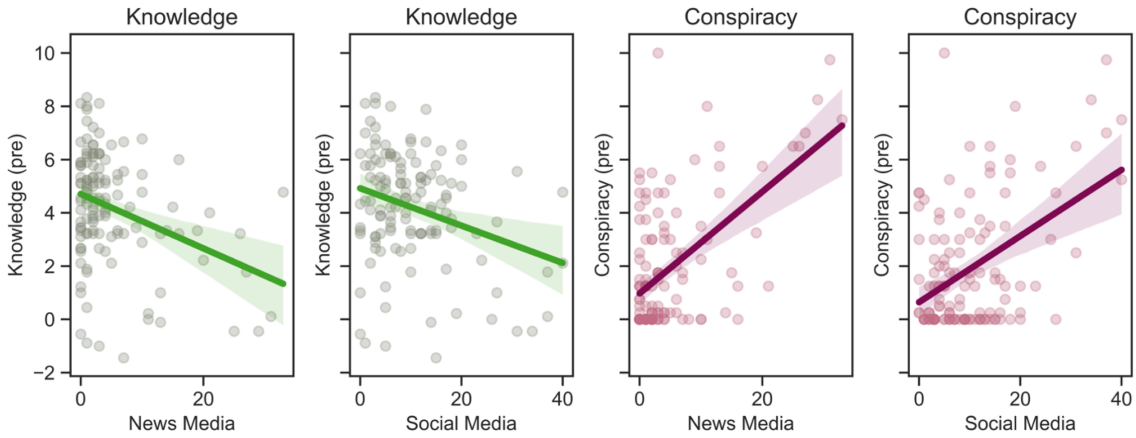


Figure 6.4: Knowledge (Green) and Conspiracy (Purple) at pre-test, as a function of News Media consumption and Social Media usage.

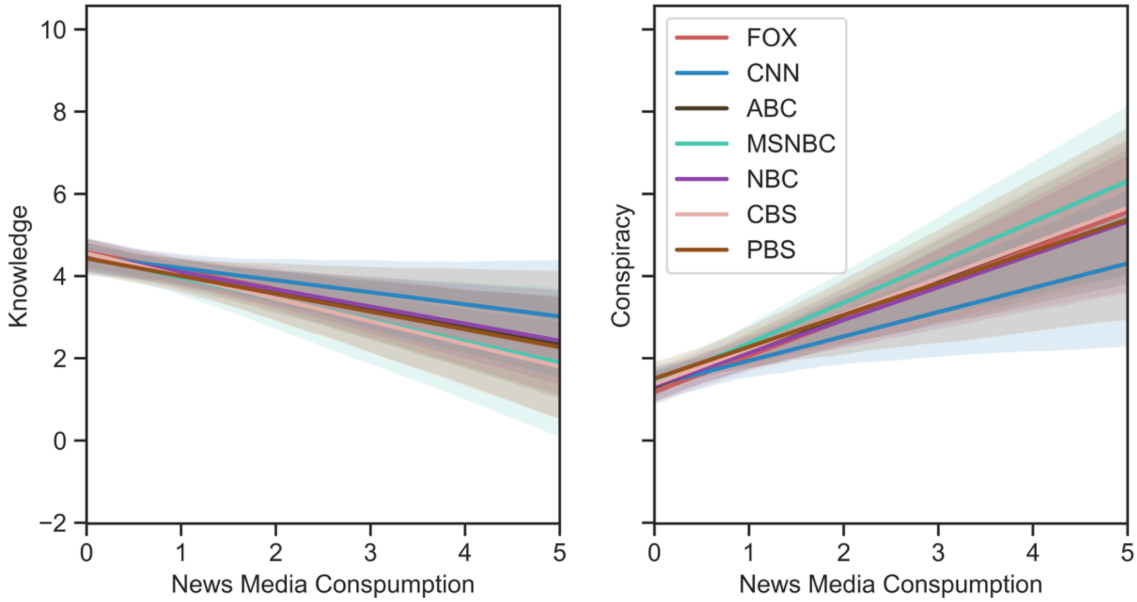


Figure 6.5: Knowledge (left) and Conspiracy (right) at pre-test, as a function of News Media consumption, by news networks.

Cohen's d=0.67, CI [0.49, 2.36], and Republicans (M=3.65, SD=2.41) believed more conspiracies than Democrats (M=1.15, SD=2.23), t(53)=4.81, p<0.001, Cohen's d=1, CI [1.49, 3.52].

Lastly, in a repeated measures ANOVA with trust as the dependent variable, participant ideology (Democrats and Republicans) as the between-subject variable,

and source (Trump and Fauci) as the within-subject variable, and found a significant main effect of ideology $F(1, 93)=49.69$, $p<0.001$, $\eta_p^2=0.34$, a significant main effect of source, $F(1, 93)=178.2$, $p<0.001$, $\eta_p^2=0.65$, and a significant interaction $F(1, 93)=78.7$, $p<0.001$, $\eta_p^2=0.45$. Post-hoc analyses suggested that Republicans (M=7.56, SD=2.34) trusted Trump more than Democrats (M=1.73, SD=2.76) $t(65)=10.6$, $p<0.001$, Cohen's d=2.2, CI [4.66, 6.99], and Democrats (M=8.44, SD=1.76) trusted Fauci marginally more than Republicans (M=7.66, SD=2.42), $t(43)=1.76$, $p=0.08$, Cohen's d=0.39, CI [0.1, 1.66].

Discussion

In a high-risk environment, such as during an epidemic, people are exposed to a large amount of information, both accurate and inaccurate, which they typically discuss with each other in conversations. Here, I show that during the coronavirus pandemic, individuals talking to each other are sensitive to their conversational partners, by changing their beliefs regarding COVID-19 information according to their partners' conveyed beliefs. This influence is strongest for initially moderately held beliefs (compared to initially endorsed or opposed beliefs), and for accurate information (compared to inaccurate or conspiracy information). This finding extends prior research showing this effect at the level of memory (Cuc et al, 2007), and aligns with prior research showing individuals synchronize their beliefs after engaging in conversations with other participants in social networks (Vlasceanu, Morais, Duker, Coman, 2020).

Moreover, I found that participants in the high versus low epistemic condition discussed more accurate information, consistent with prior work in which epistemic accuracy manipulations have been found successful at diminishing misinformation spread (Lewandowsky et al, 2012). However, this difference in conversational content did not lead to differences in how knowledgeable participants in the two conditions became as a result of conversations, pointing to a higher resistance to this manipulation

when it comes to changing one's beliefs compared to simply choosing what to discuss and propagate. This finding is consistent with work on audience tuning showing that while people tune to their audiences, they only change their memories consistent with this tuning when motivated to create a shared reality with the audience (Echterhoff, Higgins, Groll, 2005). Thus, it could be that the context created by the design, in which strangers interacted with each other, did not trigger the necessary conditions for an epistemic accuracy manipulation to meaningfully impact beliefs.

Lastly, in exploratory analyses, I found that having COVID-19 knowledge is predicted by trusting Fauci, not trusting Trump, and feeling threatened by COVID-19. Conversely, endorsing conspiracies is predicted by trusting Trump, not trusting Fauci, news media consumption, social media usage, and political orientation. In further exploration, I found that Democrats are more knowledgeable than Republicans, and Republicans believe more conspiracies than Democrats. These findings, although needing to be confirmed through replications by future research, also align with prior work and have important implications in the current socio-political context. The interaction between ideology and conspiracy endorsement is consistent with prior instances in which Republicans endorsed a conspiracy theory more than Democrats (Pasek, Stark, Krosnick, Tompson, 2015), and with the general trend in the wider political literature of Republicans being more likely to believe conspiracies about Democrats and vice versa (Hollander, 2018; Smallpage, Enders, Uscinski, 2017; Miller, Saunders, Farhart, 2016; Oliver & Wood, 2014; Radnitz & Underwood, 2015). These trends are applicable in the case of COVID-19, which was labeled a "hoax" by President Trump, and a "Democratic hoax" by Eric Trump. A surprising finding was that news media consumption positively predicted believing conspiracies regarding COVID-19, even when controlling for demographic variables such as age, gender, education, and political orientation. This effect was not driven by a particular news network, instead it was a general effect of news media consumption. This finding

counters prior work suggesting that people consuming news media are less likely to believe conspiracies (Hollander, 2018; Stempel, Hargrove, Stempel III, 2007) and that people who are more knowledgeable about news media are also less likely to endorse conspiracy theories (Craft, Ashley, Maksl, 2017). Thus, clarifying the mechanism of this discrepancy would be a worthwhile future trajectory.

Several other research trajectories emerge from this work. For instance, an important aspect of belief change that was omitted here is source credibility (Chung, Fink, Kaplowitz, 2008; Slater & Rouner, 1996). This line of work would benefit from future investigations into how the source presenting information might influence the conversational content and belief change, and how this influence might be amplified or attenuated by conversations. Also, future work could investigate whether revealing features of the conversational partner, such as their ideological orientation, might moderate individuals' willingness to change their beliefs as a function of their conversations. The present work could also be extended from the dyadic level to the collective belief level (Vlasceanu, Enz, Coman, 2018; Vlasceanu & Coman, 2020) by investigating the effect of multiple conversations within communities on belief change. Critically, these dyadic level influences (i.e., from speaker to listener) have been found to propagate in social networks (Coman, Momennejad, Duker, Geana, 2016; Vlasceanu et al, 2020). In line with existing research, it is likely that the high perceived risk of infection might influence the propagation of information through social networks. Tracking information propagation in fully mapped social networks would be critically important, especially given policymakers' interests in impacting communities (Dovidio & Esses, 2007).

In Chapter 6 I showed the effects of dyadic conversational interactions on belief change. In the next two Chapters I will zoom out one step further, and investigate the effects of conversations in social networks on belief change.

Chapter 7

From Dyads to Networks

7.1 Study 7: Network Structure Shapes Collective Beliefs

Abstract

People's beliefs are influenced by interactions within their communities. The propagation of this influence through conversational social networks should impact the degree to which community members synchronize their beliefs. To investigate, we recruited a sample of 140 participants and constructed fourteen 10-member communities. Participants first rated the accuracy of a set of statements (pre-test) and were then provided with relevant evidence about them. Then, participants discussed the statements in a series of conversational interactions, following pre-determined network structures (clustered/non-clustered). Finally, they rated the accuracy of the statements again (post- test). The results show that belief synchronization, measuring the increase in belief similarity among individuals within a community from pre-test to post-test, is influenced by the community's conversational network structure. This synchronization is circumscribed by a degree of separation effect and is equivalent in the clustered and non- clustered networks. We also find that

conversational content predicts belief change from pre-test to post-test.

Introduction

Human societies are characterized by extensive communicative exchanges. This dynamic information flow has been shown to exert a strong influence on people, impacting their individual memories (Cuc, Koppel, & Hirst, 2007), their beliefs (Vlasceanu & Coman, 2020), and their behaviors (Frankel & Swanson, 2002). It has also been found to affect collective-level phenomena, such as the formation of collective memory (Coman, Momennejad, Drach, Geana, 2016), collective beliefs (Vlasceanu, Morais, Duker, Coman, 2020), and collective decision-making (Bahrami et al, 2012). A growing body of work has been focusing on the cognitive and social processes involved in these collective phenomena (Vlasceanu, Enz, Coman, 2018; Borge, Ong, Rose, 2018), revealing the importance of network structure in their emergence. However, little is known about the impact of network structure on the formation of collective beliefs, a construct of vital social importance given its potential to impact behavior (Shariff & Rhemtulla, 2012; Mangels, Butterfield, Lamb, Good, Dweck, 2006; Ajzen, 1991; Hochbaum, 1958). While a belief is defined as a statement held to be true (Schwitzgebel, 2010), a collective belief is characterized by a group of individuals' joint commitment towards a particular belief (Gilbert, 2000; Bouvier, 2004). A central feature of beliefs is their dynamic nature, making them susceptible to change (Bendixen, 2002). Indeed, prior work on the synchronization of beliefs revealed that conversations within 3 member groups can change beliefs, leading to their coordination (Wilkes-Gibbs & Clark, 1992). Additionally, Vlasceanu, Morais, Duker, and Coman (2020) showed that community members, after being exposed to a public speaker's beliefs and having conversations about them within their social networks, align with the public speaker's beliefs and therefore become more synchronized with each other. This effect, they showed, was driven by a memory mechanism by which

beliefs become stronger when their mnemonic accessibility is increased and weaker when their accessibility is decreased (Vlasceanu & Coman, 2018).

Here, I am interested in expanding this work and programmatically exploring how the synchronization of collective beliefs is influenced by a community's network structure. I am studying this influence considering both a time-independent topological mapping typically used in network analysis (Watts & Strogatz, 1998), as well as the temporal sequencing of conversational interactions in networks (Tang,Musolesi, Mascolo, Latora, 2009). Both modalities of experimentally manipulating networks have been found to impact collective-level phenomena. Coman and colleagues (2016) manipulated the clustering coefficient of conversational networks and found that non-clustered networks reached higher mnemonic convergence than clustered networks. To showcase the impact of temporal sequencing of conversations in driving the formation of collective memories, Momennejad, Duker, & Coman (2019) manipulated when conversations occurred between people who bridged network clusters (either early or late during the community's conversations). They found that early conversations between bridge individuals lead to increased community-wide convergence compared to late conversations between bridge individuals. No such investigation has been conducted to explore the dynamics involved in the formation of collective beliefs.

To investigate how network structure affects collective belief synchronization in social networks, I designed an experiment in which participants were invited in the lab in groups of 10, each group forming a lab-created community (Figure 7.1). One hundred and forty participants enrolled in the study through Princeton University's recruitment system. First, participants rated the accuracy of a set of statements, pretested and selected for their moderate believability (belief pre-test). In reality, half of the statements were scientifically accurate, and the other half were scientifically inaccurate. They were then provided with relevant evidence supporting or refuting half of the initial statements (target items); the other half of the initial statements

for which no evidence was provided were considered baseline items. The pieces of evidence were constructed to vary in a randomized fashion on various features (e.g., anecdotal/scientific) to increase the credibility of the stimuli and the external validity of the study. After reading the evidence, participants were asked to discuss the statements with each other, in a series of dyadic conversational interactions within their community. These interactions were computer-mediated and followed one of two pre-determined network structures, clustered and non-clustered (Figure 7.1). Participants were assigned to their position in clusters randomly and did not have knowledge of the structure of the network. Finally, participants rated the accuracy of the initial set of statements again (belief post-test).

The first hypothesis was that the conversational network structure would impact the synchronization of collective beliefs, measured as the increase in belief similarity of all pairs of individuals within each network from pre-test to post-test. I also hypothesized that this effect would be circumscribed by a degree of separation effect, such that individuals closer to each other in the network would become more similar than individuals further away from each other in the network. Lastly, I hypothesized that the conversational content would influence the direction and degree of belief change, such that statements endorsed in conversations would increase in believability whereas statements opposed in conversations would decrease in believability.

To compute the dependent variables, I operationalized rational belief update as the belief change from pre-test to post-test in the direction corresponding to incorporating the available evidence (Table 7.1, eq. 1). For statements with supporting evidence – the rational update is to increase in believability from pre-test to post-test. For statements followed by refuting evidence – the rational update is to decrease in believability from pre-test to post-test. Through counterbalancing, I ensured that participants could not trivially infer that "correct" updates must be in one direction. I also operationalized belief similarity as the correlation of the

beliefs held by a pair of individuals at a given timepoint (Table 7.1, eq. 2), and belief synchronization as the increase from pre-test to post-test in the average of all belief similarity scores within a network (i.e., all correlations of beliefs of all the possible pairs of individuals; table 7.1, eq. 3). The degree of separation denoted the minimum number of links a given pair of nodes within a network are away from each other (i.e., how many conversations would need to occur for information from one node to reach the other). Moreover, I coded the conversations' content for recall and for belief endorsement. The recall coding involved a binary system in which a statement was labeled as either mentioned or not mentioned in each conversation by either the participant (listener recall), their partner (speaker recall), or either (joint recall). Coding for belief endorsement in conversations entailed the additional factor of valence, which denoted whether a mentioned statement was endorsed (strongly, moderately, slightly) or refuted (strongly, moderately, slightly) in conversation, by either the participant (listener belief), their partner (speaker belief), or either (joint belief).

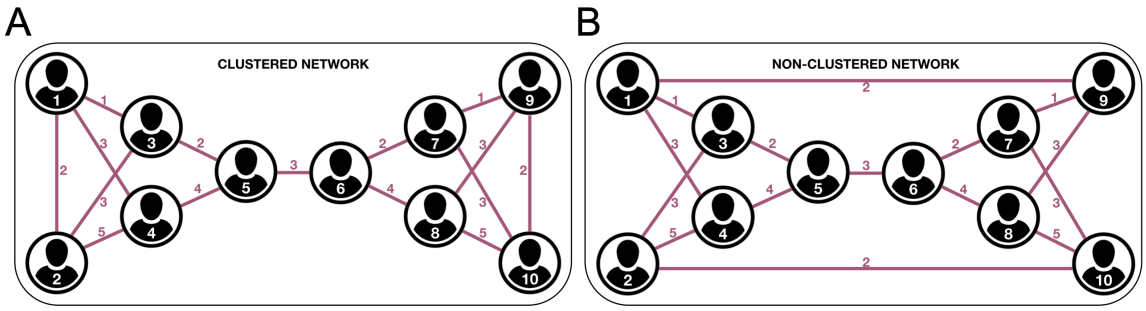


Figure 7.1: Network structures: clustered (Panel A), non-clustered (Panel B). Circles represent participants, and links represent conversations. Numbers in red indicate the sequence of conversations.

Methods

Open science practices. The data and stimulus materials can be found on my open science framework page: https://osf.io/72ayt/ The data analysis (in python) can be

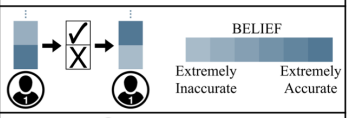
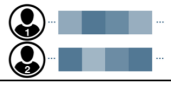
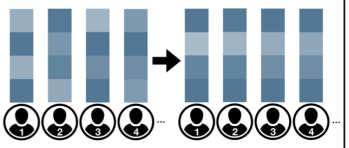
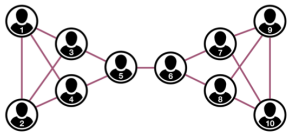
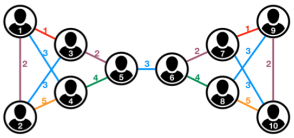
MEASURE/DEFINITION	FORMULA	FIGURE
RATIONAL BELIEF UPDATE The belief change from pre-test to post-test in the direction corresponding to incorporating the available evidence.	$\Delta B_r = E \times \Delta B$ (1) Evidence = {Refuting, Supporting} $\rightarrow E = \{-1, 1\}$ ΔB = change in belief from pre to post	
BELIEF SIMILARITY The correlation of the beliefs held by a pair of individuals at a given timepoint	$BSim_{i,j} = r_S(B_{P_i}, B_{P_j})$ (2) B_{P_i} = belief vector of participant i	
BELIEF SYNCHRONIZATION The increase from pre-test to post-test in the average of all belief similarity scores within a network	$BSynch = \overline{r_S^{post}} - \overline{r_S^{pre}}$ (3) $\overline{r_S} = \frac{1}{\binom{N}{2}} \sum_{i<j} r_S(B_{P_i}, B_{P_j})$ (4) $\overline{r_S}$ = average belief similarity in a network N = number of nodes (participants in a network) $\binom{N}{2}$ = number of pairs of nodes	
CLUSTERING COEFFICIENT (STATIC) Measure of the degree to which nodes in a graph tend to cluster together.	$\overline{C} = \frac{1}{N} \sum_i \frac{\#\text{ of edges in } G^{\mathcal{N}_i}}{\binom{k_i}{2}}$ (5) $\mathcal{N}_i$ = neighbors of node i, G = network, $G^{\mathcal{N}_i}$ = neighbor subnetwork of node i k_i = # of neighbors of node i	
CLUSTERING COEFFICIENT (TEMPORAL) Measure of the degree to which nodes in a graph tend to cluster together at each moment in time.	$\overline{C_{0,T}} = \frac{1}{N} \sum_i \sum_{t=0}^{T} \frac{\#\text{ of edges in } G_t^{\mathcal{N}_i}(0,T)}{T \times \binom{k_i(0,T)}{2}}$ (6) T = conversational round	

Table 7.1: Definitions, equations, and figures for the dependent variables

viewed as a jupyter notebook here: https://github.com/mvlasceanu/CollectiveBeliefs

Participants. A total of 140 Princeton undergraduate students (63% women; Mage=19.47, SDage=1.53) were recruited for the study. They participated in the study for either monetary compensation or research credit. Participants were grouped into fourteen 10-member networks. The statistical power afforded by this sample size was deemed adequate given effect sizes obtained in previous studies using a similar sample size and experimental paradigm (Coman et al, 2016; Vlasceanu et al, 2020). I note that the aggregation procedure used to compute network-wide belief synchronization reduces the standard errors around the mean and results in a highly accurate estimate of the true value. The study was approved by the Institutional Review Board at Princeton University, and was conducted in accordance with IRB guidelines and regulations. Informed consent was obtained from all participants.

Stimulus materials. A set of 16 statements of moderate believability were selected from a larger set of 32 statements used in a study by Vlasceanu and Coman (e.g., "Eating carrots will make eyesight sharper"). This study, conducted on a sample of Princeton undergraduate students (N=200; Mage=19.49, SDage=1.39; 64% women), collected believability ratings for each statement (i.e., "How accurate or inaccurate do you think this statement is" on a scale from 0-Extremely Inaccurate to 100-Extremely Accurate). The final set of 16 statements I used in the current study were selected such that for each of them, their level of believability in this population of interest was in the moderate range (M=51.41, SD=4.18, on a 0 to 100-point scale). Even though each of these 16 statements was equally believable by design, half of them were scientifically accurate, while the other half were scientifically inaccurate, as determined by published scientific papers.

I constructed a set of 8 pieces of direct evidence, either in favor or against half of the initial 16 statements. The statements for which no evidence was presented were considered baseline items. The pieces of evidence were constructed such that they argued in favor of the initial statement if the statement was accurate (e.g., "Children who spend less time outdoors are at greater risk to develop nearsightedness, study shows") and against the initial statement if the statement was inaccurate (e.g., "Eating carrots does not makes eyesight sharper, study shows"). To increase external validity, these pieces of evidence were constructed and displayed to participants as if they were tweets collected from the Twitter platform. To increase the variability of these pieces of evidence, mirroring content frequently found on Twitter (Zhang et al, 2019), I counterbalanced the phrasing suggesting the evidence posted is the result of a study, with phrasing suggesting the evidence posted is anecdotal, such that each statement was in either counterbalancing condition with equal probability randomly assigned across participants. Also consistent with content on Twitter (Kim, 2018), I counterbalanced the number of retweets each of these tweets had,

such that each statement either had a large or a small number of retweets also with equal probability randomly assigned across participants. Moreover, for each piece of evidence, the sources were constructed to be as similar as possible while allowing some variability, to maintain the credibility of the stimuli. In each case, the person tweeting was depicted as a white middle-aged male, with a common name and appearance. The dates of the tweets were randomly chosen from dates in the month on July 2019. Even though these varying features of the evidence varied randomly across participants and items, they were held constant within a network.

Design and procedure. The 140 participants were split in 14 lab-created communities of 10 participants. Each community was assembled separately and was comprised of individuals who arrived in the lab at the same time. These communities were randomly assigned to either the Clustered network structure condition (7 networks) or the Non-Clustered network structure condition (7 networks), following procedures by Coman and colleagues. These network structures resemble real-world small networks characterized by varying degrees of clustering (Watts & Strogatz, 1998; Tang et al, 2009). Once assigned to condition, participants went through four experimental phases. Participants were told they would participate in an experiment about people's evaluation of information and were directed to the survey on the Qualtrics platform. After completing the informed consent form, participants were directed to the first phase (pre-test), in which they rated a set of 16 statements (one on each page) by indicating the degree to which they believed each statement (i.e., "How accurate do you think this statement is," from 1-Extremely inaccurate to 100-Extremely accurate). Then, in the second phase (evidence phase), participants were exposed to a sub-set of 8 pieces of evidence (in the form of tweets, one on each page), half of which argued in favor and the other half argued against the initial statements (i.e. target statements). Eight statements presented initially constituted baseline items. Each of the 16 statements were randomly assigned to either a target

or a baseline status. To ensure participants processed the evidence information, they were instructed to rate each tweet on how convincing, rigorous, widespread, and personal it appears to them, as well as how likely they would be to share it. In the third phase (conversational phase), participants were directed to another software platform (i.e., Software Platform for Human Interaction Experiments; SoPHIE) that allows fluent computer-mediated interactions among participants. At this stage, participants were instructed to discuss the information from the study with each other, in a series of dyadic conversations (each with a different partner). Conversations took the form of interactive exchanges in a chat-like computer-mediated environment in which participants typed their responses. Each participant had three conversations, each lasting 150 seconds. Participants in the Clustered condition (n=70 participants; seven 10-member networks) communicated according to a network structure characterized by two subclusters (Figure 7.1A), whereas participants in the Non-clustered condition (n=70 participants; seven 10-member networks), communicated according to a network structure characterized by a single large cluster (Figure 7.1B). The number of participants per network (n=10), the sequencing of conversational interactions, and the number of conversations each participant had (i.e., three) were kept constant between the two conditions. Finally, in the fourth phase (post-test) participants rated again the believability of the initial 16 statements, after which they were asked to complete a series of demographic information and were debriefed.

Conversational Coding. I coded the conversations' content for recall and for belief endorsement. The recall coding involved a binary system in which a statement was labeled as either mentioned or not mentioned by each participant in each of their conversations. Thus, a participant could have mentioned a given statement anywhere from 0 times up to 3 times (once in each of their 3 conversations). For each participant, their conversational partners could have also mentioned a particular statement up to 3 times. Thus, for each participant, the joint recall measure I

used denotes the number of times either the participant or their interaction partners mentioned each statement (i.e., from 0 to 6). Coding for belief endorsement in conversations entailed the additional factor of valence, which denotes whether a mentioned statement is being endorsed or refuted in conversation. Thus, the belief coding involved a larger interval (i.e., from -3 to +3), in which a mentioned statement was labeled as either strongly endorsed (+3), moderately endorsed (+2), slightly endorsed (+1), just mentioned (0), slightly opposed (-1), moderately opposed (-2), or strongly opposed (-3). Again, for each participant I accounted for both their own input and their 3 conversational partners' inputs for each statement, to form a measure of joint belief (i.e., from -18 to +18). To make this interval comparable and consistent to the recall interval, I normalized it by dividing it by 3. In each case, ten percent of the data were double coded for reliability (Cohen k ¿0.89).

Results

Since there was no difference in the initial level of believability of scientifically accurate statements (M=51.36, SD=12.68) and scientifically inaccurate statements (M=51.93, SD=12.61), p=0.66599, which were pre-tested to be as similar as possible, I combined them for the rest of the analyses conducted. Similarly, I did not find any differences in rational updating between the varying features of the pieces of evidence (e.g., anecdotal/scientific) in a repeated-measures ANOVA with rational belief update as the dependent variable, and evidence type as a within-subject variable $F(3, 417)=1.09$, p=0.353. Therefore, I combined them for the rest of the analyses. To investigate whether individuals' beliefs became more similar post conversations relative to pre-conversations, I ran a linear mixed model with belief similarity (for the target items) as the dependent variable, time-point (pre-test vs. post-test) as the fixed effect, and by-network random intercepts which nests the data by networks, and found that belief similarity scores (i.e., pairwise belief correlations) at post-test (M=0.0649, SD=0.380)

were significantly higher (β=0.06, SE=0.01, t(139)=3.5, p<0.001) than at pre-test (M=0.0079, SD=0.374) (Figure 7.3A).

Collective belief synchronization is dependent on network structure. To investigate the first hypothesis, that network structure would impact belief synchronization, I conducted two sets of analyses. First, I compared the levels of synchronization achieved by the two network structures employed. I found that synchronization was equivalent in the Clustered (M=0.0257, SD=0.090) compared to the Non-Clustered (M=0.0254, SD=0.067) condition (Figure 7.3B), in a linear mixed model with belief similarity difference (for the target items) as the dependent variable, condition (clustered vs. non-clustered) as the fixed effect, and by-network random intercepts (β=0.02, SE=0.01, t(13)=1.52, p=0.15).

A more complex analysis that takes into account the degree of separation among community members involved the construction of a hypothesis matrix for each of the 2 network structures used (clustered and non-clustered; Figure 7.2A), following methods by Coman and colleagues. I reasoned that participants who had a conversation with one another would synchronize their beliefs more than participants who were two degrees away from one another in the network, and so on. In this context, a hypothesis matrix represents the hypothesized distance in belief ratings between any two participants in a community, depending on their distance in the network. In constructing these hypotheses matrices, I used the range of the entire empirical belief synchronization scores (i.e., post minus pre scores of the target items belief correlations of each pair of participants within a network) collected in each condition. I split the intervals of these distributions into quintiles (clustered) or terciles (non-clustered), and mapped the boundary values of these intervals to the corresponding degree of separation in the participant by participant matrix (e.g., first quantile boundary value was mapped on the first degree of separation, second quantile boundary value was mapped on the second degree of separation, etc.). This density matching procedure was employed

such that the hypotheses matrices can be comparable to the corresponding empirical matrices. The empirical data matrices were constructed by allocating the observed synchronization scores of each pair of participants to the corresponding cells in the participant by participant matrices.

Once I constructed both the hypothesis matrices and the empirical matrices (Figure 7.2A), I ran a non-parametric statistical test to assess the first hypothesis – that network structure would circumscribe belief synchronization. Specifically, I tested whether the empirical matrices are better explained by their corresponding hypothesis matrix than by any other hypothesis matrices that could exist based on the same constraints of building the conversational network. To do so, I simulated 10,000 bootstrapped random hypothesis matrices matched to the true hypothesis matrices' features (10 nodes, 3 ties per node) using the same data mapping procedure used for the true hypothesis matrices. I then compared the empirical matrices to both the true hypothesis matrices and to each of these 10,000 random hypothesis matrices by running a series of quantile correlations. I found that the empirical data was better explained by the true hypothesis matrices (the true network structures) than by most of the simulated hypothesis matrices for both the Clustered ($p=0.0462$) and the Non-Clustered ($p=0.0075$) conditions (Figure 7.2B). The p-values in this non-parametric test represent the probability that any of these simulated hypothesis matrices better explain the data than the true ones. Thus, I found support for the first hypothesis, that belief synchronization is determined by the structure of the network individuals are embedded in.

Further support for the hypothesis that the level of belief synchronization exhibits a degree of separation effect, comes from a repeated measures ANOVA nested by network, with belief similarity difference as the dependent variable and degree of separation (1-5) as the independent within-subject variable, showing a main effect of degree of separation $F(4, 52)=4.21$, $p<0.004$, $\eta_p^2=0.24$. Moreover, in a linear mixed model

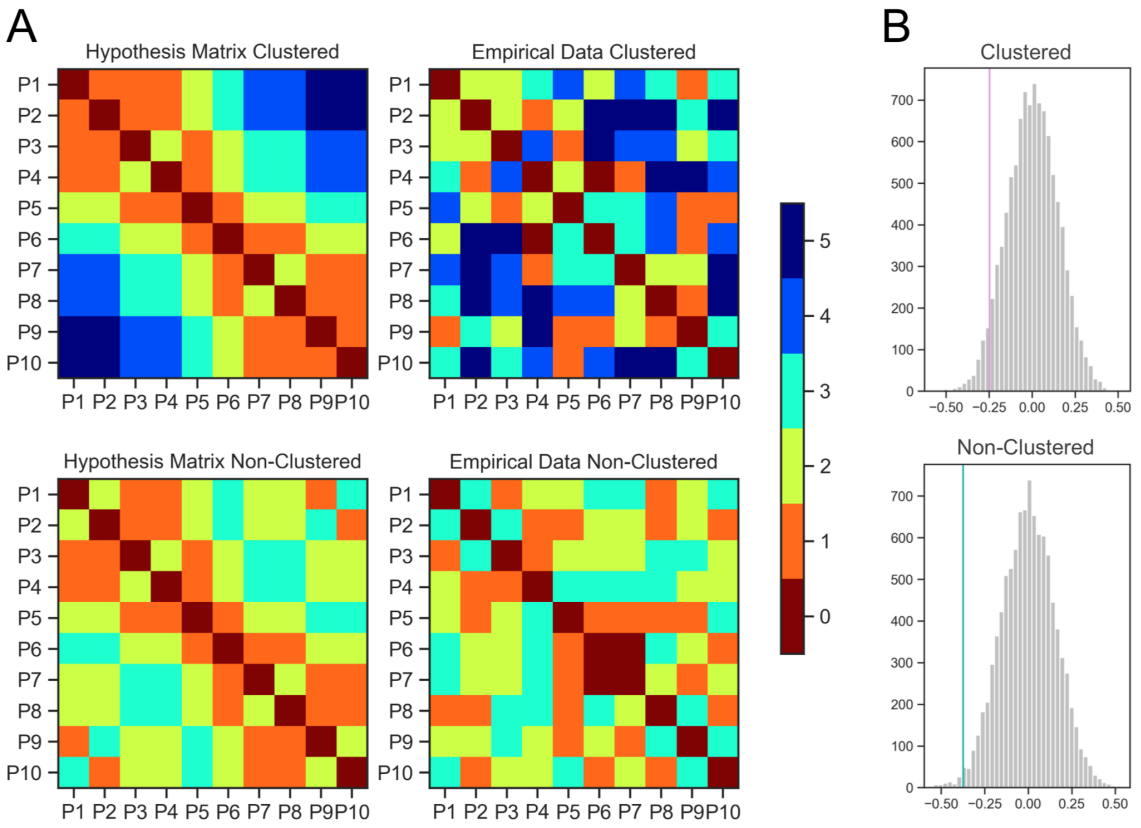


Figure 7.2: Panel A. Hypothesis Matrix of Clustered (top left) and Non-Clustered (bottom left) Network Structures. Empirical Data of Clustered (top right) and Non-Clustered (bottom right) Network Structures. Panel B. Histogram of belief similarity correspondence between empirical data (Clustered on the top, Non-Clustered on the bottom) and 10,000 bootstrapped random hypothesis matrices matched to the experiments' network features (10 nodes and 3 ties per node). The vertical lines represent the correspondence of the hypothesis matrices with the true network structures.

with belief similarity difference (for the target items) as the dependent variable, degree of separation (with 5 levels) as the fixed effect, and by-network random intercepts, I find that, as hypothesized, individuals situated 1-degree of separation away (β=0.057, SE=0.017, t(48)=3.39, p<0.0014) became significantly more similar to each other, whereas individuals 2-degrees away (β=0.026, SE=0.017, t(47)=1.54, p=0.129) or further did not (3-degrees away: β=0.009, SE=0.02, t(101)=0.46, p=0.641, 4-degrees away: β=-0.034, SE=0.03, t(230)=-1.13, p=0.258, and 5-degrees away: β=0.032, SE=0.04, t(447)=0.79, p=0.427) (Figure 7.3C). This suggests that the belief influence

within a community only travels one degree of separation away from the originating source.

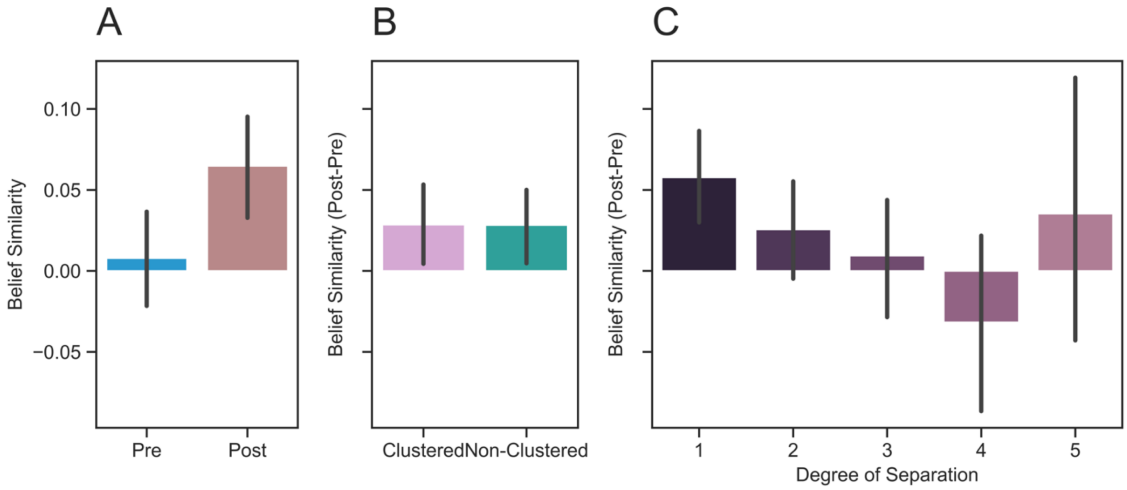


Figure 7.3: Panel A. Belief similarity at pre-test and post-test. Panel B. Belief similarity change (pre-test minus post-test) in the Clustered versus Non-Clustered Networks. Panel C. Belief similarity change (pre-test minus post-test) by degree of separation. Error bars represent ±1 standard errors of the mean.

Conversational content predicts belief change. So far, I have showed that people's beliefs become more similar after conversations, and this synchronization occurs according to their position in conversational networks. However, I have not yet considered the content of their conversations, and how these conversations lead to belief change. Here, I am investigating the last hypothesis, that endorsing or opposing beliefs in conversations influences rational belief update. First, focusing on simple recall, I conducted a linear mixed model with rational belief update (for the target items) as the dependent variable, joint conversational recall as the fixed effect, as well as by-participant random intercepts and by-network random intercepts (Figure 7.4B). I found that joint recall significantly predicts rational belief update (β=1.62, SE=0.90, t(941)=3.21, p<0.0014). This was true regardless of whether the speaker (β=2.91, SE=0.47, t(892)=3.42, p<0.001) or the listener (β=2.54, SE=0.83, t(850)=3.04, p<0.0024) was the one mentioning the beliefs. Second, to assess the impact of endorsement/opposition on rational belief update, I conducted a linear mixed

model with rational belief update (for the target items) as the dependent variable, joint conversational belief endorsement as the fixed effect, as well as by-participant random intercepts and by-network random intercepts (Figure 7.4A). I found that joint belief endorsement also significantly predicts rational belief update (β=1.92, SE=0.33, t(1079)=5.71, p<0.001). Again, just like for the memory effect, the belief effect was true regardless of whether the speaker (β=2.99, SE=0.62, t(1097)=4.78, p<0.001) or the listener (β=2.58, SE=0.54, t(1110)=4.93, p<0.001) was the one supporting/opposing the belief.

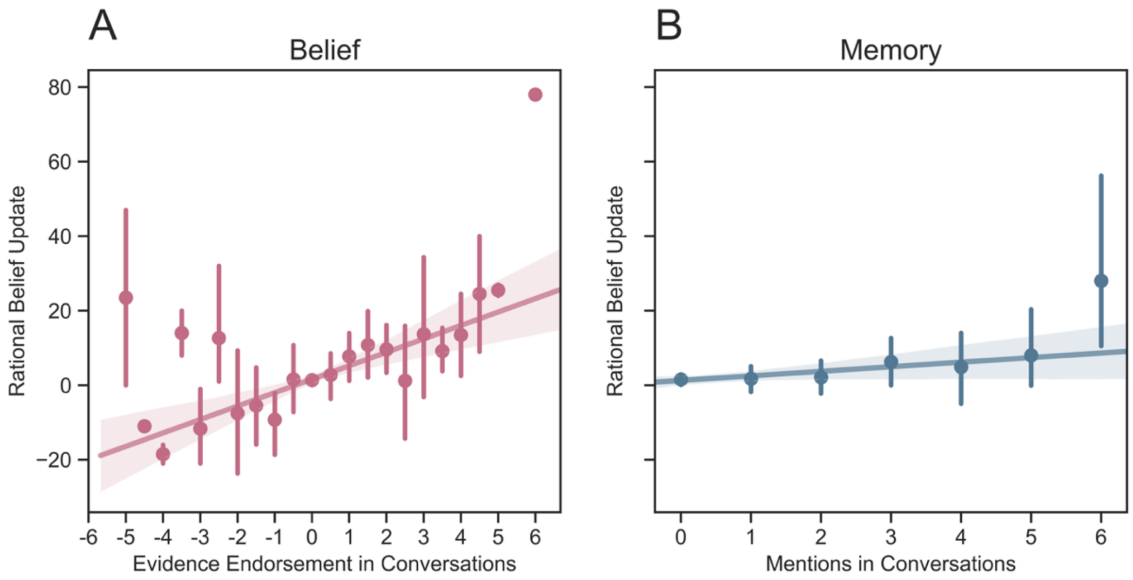


Figure 7.4: Rational belief update as a function of conversational belief endorsement (Panel A) and as a function of conversational recall (Panel B). Error bars represent 95% bootstrapped confidence intervals on the means.

When running these two predictors together in the same linear mixed model with rational belief update (for the target items) as the dependent variable, and both joint conversational recall and joint conversational belief endorsement as fixed effects, as well as by-participant random intercepts and by-network random intercepts, I found that only belief endorsement (β=1.73, SE=0.35, t(1099)=4.92, p<0.001) and not recall (β=0.92, SE=0.49, t(950)=1.87, p=0.06) significantly predicts rational belief update. This suggests that both mentioning a statement and qualifying

the level of endorsement towards it seem to explain rational belief update; when taken together most of the variance in belief update is explained by the latter variable.

Discussion

Human societies are organized in social networks of interconnected individuals that exchange information. Here, I show that when individuals have conversations within their communities, they synchronize their beliefs according to the conversational structure of the network they are embedded in, such that more closely connected individuals become more similar in their beliefs than more distantly connected individuals. Moreover, the content of their conversations explains their belief change.

These findings complement and extend prior research in meaningful ways. First, the findings reported here are consistent with prior work on the formation of collective memory, extending it to the formation of collective beliefs. While Coman and colleagues (2016) showed that network structure influences a community's collective memory, here, I show that network structure influences a community's collective beliefs. Moreover, in both studies this influence was explained by a degree of separation effect, by which the smaller the degree of separation between any two participants, the more similar their beliefs became. Second, these results align with past work on the formation of collective beliefs showing that following conversational interactions, people's beliefs become more coordinated (Wilkes-Gibbs & Clark, 1992), and synchronized (Vlasceanu et al, 2020). These findings extend these effects and in addition to showing an increased overall synchronization following conversations, I show how the network structure of the conversations determines the degree of synchronization. Moreover, in addition to prior work, I show that not only do conversations increase belief similarity among communities, the content of the conversations is also important in determining the direction and degree of belief change. The more a statement is endorsed in conversation, the more its endorsement increases post conversation for both

conversational partners. Similarly, the more a statement is opposed in conversation, the more its endorsement decreases post-conversation for the conversational partners. And third, these findings are also consistent with work suggesting that even though there are 6 degrees of separation in real-world networks (Milgram, 1967), influence only spreads 3 degrees away from the originating source (Fowler & Christakis, 2010). In the domain of beliefs, it seems that the degree of influence is even shorter, with belief influence only propagating one degree away from the source. This suggests that the propagation of beliefs might involve different processes from those prompting the propagation of other psychologically grounded phenomena such as memory (Coman et al., 2016), altruism (Fowler & Christakis, 2010), and loneliness (Cacioppo, Fowler, Christakis, 2009).

Moreover, in addition to a topological mapping typically used in classic studies involving static networks (Watts & Strogatz, 1998; table 7.1, eq. 5), here, I over imposed a temporal network framework, to gain additional ecological validity (Table 7.1, eq. 6). This distinction between the topological and temporal mapping of networks is important from a theoretical perspective – while static networks aggregate information flowing through a network over time, dynamic networks account for the temporal dimension of real-world networks (Tang et al, 2009). This difference between static and dynamic networks can also explain the equivalent level of belief synchronization produced by the two network structures used in this experiment: clustered and non-clustered. These two network structures were chosen to be as distinct as possible in terms of clustering from a topological perspective (conditioned on maintaining the constraint of regular networks with 10 nodes and 3 links per node). When only taking into account the topological mapping of the network structure, the clustered network's global clustering coefficient is 0.4, and the non-clustered network's global clustering coefficient is 0 (Freeman, 1978; table 7.1, eq. 5). However, the two network structures tested here were too similar in terms of clustering coefficient when taking

into account the temporal dimension. When considering the temporal aspect of the networks, the clustered network's temporal clustering coefficient is 0.066, and the non-clustered network's temporal clustering coefficient is 0 (Tang et al, 2009; table 7.1, eq. 6). Therefore, these two temporal structures are too similar to each other to render observable differences in belief synchronization, at a small sample size. I posit that increasing the sample size or increasing the difference in temporal clustering coefficient between the network structures, would indeed render meaningful differences in belief synchronization, as suggested by the main finding that network structure plays an important role in the way people align their beliefs.

The current investigation opens several avenues for future research. First, an important extension of this work involves programmatically investigating the impact of other network features on belief synchronization. What are the temporal and topological features of the network that results in the highest degree of synchronization? Second, another noteworthy trajectory could involve assessing the impact of trust on belief synchronization. Would manipulating trust among conversational partners (e.g., high vs. low) before individuals engage in conversational interactions differentially impact belief synchronization? Lastly, investigating the effects of belief synchronization on behavior would be a research trajectory of critical importance especially in the current socio-political climate. What is the impact of collective belief change on communities' behaviors such as voting or living sustainably? Understanding the mechanisms by which collective beliefs take shape and change over time is essential from a theoretical perspective (Vlasceanu, Enz, Coman, 2018), but perhaps even more urgent from an applied point of view. This urgency is fueled by recent findings showing that false news diffuse farther, faster, deeper, and more broadly than true ones in social networks (Vosoughi, Roy, Aral, 2018), and that news can determine what people discuss and even change their beliefs (King, Schneer, White, 2017). And given that beliefs influence people's behaviors (Shariff

& Rhemtulla, 2012; Mangels, Butterfield, Lamb, Good, Dweck, 2006; Ajzen, 1991; Hochbaum, 1958), understanding the dynamics of collective belief formation is of vital social importance as they have the potential to affect some of the most impending threats our society is facing from pandemics (Pennycook, McPhetres, Zhang, Rand, 2020) to climate change (Benegal & Scruggs, 2018). Thus, policy makers could use such findings in designing misinformation reduction campaigns targeting communities (Dovidio & Esses, 2007; Lewandowsky et al, 2012). For instance, these findings suggest such campaigns be sensitive of the conversational network structures of their targeted communities. Knowing how members of these communities are connected, and leveraging the finding that people synchronize their beliefs mainly with individuals they are directly connected to, could inform intervention designers how communities with different connectivity structures might respond to their efforts.

At the collective level, I showed how dyadic interactions between people and how the network structure they are embedded in impacts their beliefs. In the final study of this dissertation, I will show how an individual-level cognitive process propagates through sequential dyadic interactions within a social network, giving rise to an emergent phenomenon: collective beliefs.

Chapter 8

Collective Belief Formation

8.1 Study 8: Memory Accessibility and Conversations Shape Collective Beliefs

This Chapter is based on the paper "The synchronization of collective beliefs: From dyadic interactions to network convergence" published in the *Journal of Experimental Psychology: Applied* in 2020. The co-authors of this publication are Michael J. Morais, Ajua Duker, and Alin Coman. Results in this Chapter have also been presented at the 30th APS Annual Convention.

Abstract.

Systems of beliefs organized around religion, politics, and health constitute the building blocks of human communities. One central feature of these collectively held beliefs is their dynamic nature. Here, we study the dynamics of belief endorsement in lab-created 12-member networks using a 2-phase communication model. Individuals first evaluate the believability of a set of beliefs, after which, in Phase 1, some networks listen to a public speaker mentioning a subset of the previously evaluated beliefs while other networks complete a distracter task. In Phase 2, all participants

engage in conversations within their network to discuss the initially evaluated beliefs. Believability is then measured both post conversation and after one week. We find that the public speaker impacts the community's beliefs by altering their mnemonic accessibility. This influence is long-lasting and amplified by subsequent conversations, resulting in community-wide belief synchronization. These findings point to optimal socio-cognitive strategies for combating misinformation in social networks.

Introduction.

People's beliefs meaningfully impact their behavior. Religious beliefs about a punishing deity are associated with reduced crime rates (Shariff & Rhemtulla, 2012), beliefs about the flexibility of human abilities cause improvements in academic performance (Mangels et al, 2006), and beliefs about immigration guide voting behavior (Schildkraut, 2010). While an important body of psychological research has explored the relation between beliefs and behavior (Ajzen, 1991), there is scarce research on how communities of individuals synchronize their beliefs. Understanding these synchronization dynamics will reveal not only how to better disseminate accurate beliefs in the population (Osterholm et al, 2015), but will also elucidate how to diminish the spread of misinformation in vulnerable communities (Hough-Telford et al, 2017; Lewandowsky et al, 2012).

Information, and beliefs based on that information, propagate through communities both because of broad exposure to public sources (e.g., politicians, pundits, celebrities), and because individuals within these communities interact with and influence one another. Accordingly, I employ a two-step flow communication model to experimentally investigate the community-wide synchronization of such beliefs (Katz & Lazarsfeld, 1955). Individuals have a set of initial beliefs, then listen to a public speaker that reiterates some of these beliefs, and, finally, communicate with one another within their social networks about these beliefs, (Vlasceanu, Enz, & Coman,

2018). This framework simulates situations in which, for example, beliefs espoused by influential individuals on Twitter or Facebook are then discussed by their followers in subsequent interactions either online or face-to-face (Bhattacharya, Srinivasan, & Polgreen, 2014; Hilbert et al, 2016).

Exposure to social sources of information has been shown to meaningfully impact people's beliefs in surprising ways. Previous research suggested that superficial features of the belief evaluation experience can impact belief endorsement (Gilbert, Krull, & Malone, 1990; Hasher, Goldstein, & Toppino, 1977). For example, information encountered in the past becomes more believable, a phenomenon known as the illusory truth effect (Hasher, Goldstein, & Toppino, 1977; Begg, Anas, & Farinacci, 1992; Ozubko & Fugelsang, 2011; Fazio, Brashier, Payne, & Marsh, 2015). In a typical illusory truth effect paradigm, participants are first asked to assess the truth-status of a series of statements; then, participants are presented with the initial statements again, interspersed with new statements and are asked again to rate the degree to which they think each statement is true. The finding is that repeated statements are judged as more true than novel statements (for a meta-analytic review, see Dechêne, Stahl, Hansen, & Wänke, 2010). The illusory truth effect has been shown to occur due to increased familiarity, through the up-regulation of memory or, in other words, the increase of mnemonic accessibility of information (Ozubko & Fugelsang, 2011). Vlasceanu and Coman (2018) used the same principle to investigate the effect of down-regulating memory, or decreasing mnemonic accessibility, on belief endorsement (Chapter 2). They found that mnemonic accessibility influences believability in both directions: a belief that is easier to recall (due to increased mnemonic accessibility) becomes more believable, and a belief that is harder to recall (due to decreased mnemonic accessibility) becomes less believable compared to a baseline belief whose mnemonic accessibility is not manipulated. This research is grounded in a well-established method to alter mnemonic accessibility: retrieval-induced forget-

ting (RIF; Anderson, Bjork, & Bjork, 1994; Murayama, Mityatsu, Buchli, & Storm, 2014). According to RIF, selectively remembering previously encoded information results in increased mnemonic accessibility for the remembered information (i.e., rehearsal effect) while at the same time leads to forgetting the information that was unmentioned, but related to the mentioned information. In a typical selective practice paradigm, participants first study category-exemplar pairs (e.g., the "Fruit" category contains the "Apple" and "Pear" exemplars; the "Tree" category contains the "Oak" and "Pine" exemplars) and then selectively practice half of the exemplars from half of the categories in a stem completion task (e.g., "Fruit-Ap.."). Analyses of a final cued-recall test show that practiced items (RP+ items: Fruit-Apple) are remembered better than unpracticed unrelated items (NRP items: exemplars in the "Tree" category) - a rehearsal effect. Unpracticed items related to those practiced (RP- items: Fruit-Pear) are remembered worse than NRP items - a retrieval-induced forgetting effect (RIF). RIF is thought to occur due to inhibitory processes triggered by the response competition during the practice phase (Kuhl, Dudukovic, Kahn, & Wagner, 2007; but see Mensink & Raaijmakers, 1988). Moreover, the selective practice of information occurring in conversational settings has been found to trigger similar effects (Coman, Manier, & Hirst, 2009). Specifically, when a listener monitors a speaker selectively practicing previously encoded information, the listener experiences socially-shared retrieval-induced forgetting (SS-RIF). That is, she forgets information related to what the speaker mentioned in the conversation. This phenomenon occurs because listeners concurrently retrieve the information along with the speaker, which triggers response competition from related memories, just like in the case of RIF (Cuc, Koppel, & Hirst, 2007).

Building on this literature, Vlasceanu and Coman (2018) showed how altering mnemonic accessibility can result in individual-level belief change. They first asked participants to rate the believability of a set of 24 statements (pretest) organized in 4

categories (e.g., in the "Allergy" category: "Children can outgrow peanut allergies" and "Some babies are allergic to their mother's milk"; in the "Health" category: "The majority of people infected with malaria are children" and "Crying helps babies' lungs develop"). Then, participants listened to a person mentioning 2 statements in each of 2 categories (e.g., "Allergy-Children can outgrow peanut allergies"). Following this selective practice phase, participants were asked to recall as many of the 24 initial statements as they could in a cued recall test (i.e., the category was provided). Finally, participants were asked to rate again (posttest) the believability of the 24 statements. The memory results showed increased recall rates of selectively practiced statements (i.e., increased mnemonic accessibility), and decreased recall rates of unpracticed but related statements (i.e., decreased mnemonic accessibility), both compared to unpracticed and unrelated statements. Moreover, the belief-level results indicated that the believability of the practiced statements (e.g., "Allergy-Children can outgrow peanut allergies") increased from pretest to posttest, and the believability of the unpracticed but related statements (e.g., "Allergy: Some babies are allergic to their mother's milk") decreased from pretest to posttest compared to the believability of unpracticed and unrelated statements (e.g., statements in the "Health" category).

Here, I am interested in whether the findings reported by Vlasceanu and Coman (2018) occur in a context in which a public speaker addresses large communities (Phase 1) that will subsequently engage in networked conversations (Phase 2). Previous research has documented numerous instances in which public speakers significantly impacted societal level outcomes, from shaping public perceptions of climate change (Hmielowski et al, 2014), to increasing consumerism (Kumar et al, 2016) or even influencing voting behaviors (e.g., Oprah Winfrey's public endorsement of Barack Obama was estimated to have led to 1 million additional votes for Obama; Garthwaite & Moore, 2008). My question is: do public speakers impact communities' beliefs by influencing the mnemonic accessibility of those beliefs? If

so, how do conversations that take place after the public speaker's intervention impact people's beliefs? Do people's beliefs synchronize according to the influence of the public speaker and the conversations? Finally, are these effects long lasting? I aim to contribute to the two-step flow communication model by integrating recent psychological advances on the impact of communicative exchanges on memory and belief (Vlasceanu & Coman, 2018; Sperber, 1996).

Methods.

Open science practices. The data can be found on the Open Science Foundation website: https://osf.io/8vjym/

Participants. A total of 168 participants (65% women; mean age 21.36 years old) associated with Princeton University were recruited for the study. They participated in the study for either monetary compensation or research credit. Participants were grouped into fourteen 12-member networks (Figure 8.1). The sample size was determined based on the effect sizes reported in previous research: the average belief suppression effect size was d=0.26, while the average belief rehearsal effect size was d=0.36 (Vlasceanu & Coman, 2018). While a conventional effect size analysis would weigh the false alarm probability against the detection probability and choose a sample size accordingly, I argue that there is an additional consideration involving the network level of analysis. In contrast to the previous research, the current study involves repeated conversational interactions. Because these conversations have been shown to have a cumulative impact on the dependent variable (Coman & Hirst, 2012), and considering the prior reported effect size (Vlasceanu & Coman, 2018), I would expect an effect size between 0.4 and 0.6. The lower bound, an effect size of 0.4 would require a sample size of 200 participants to be detected. The upper bound, an effect size of 0.6 would only require 42 participants to be detected. Thus, I decided on a stopping rule of 14 networks (168 participants), also taking into consideration

previous studies investigating collective-level phenomena that used 14 (Coman et al., 2016) and 12 networks (Momennejad, Duker, Coman, 2018). This sample size would give us .90 power to detect an effect size of .5 in a between-condition comparison. Of the 168 participants, 115 (59 in the Experimental and 56 in the Control condition) completed a one-week follow-up survey (69% women; mean age 21.63 years old). The study was approved by the Institutional Review Board at Princeton University.

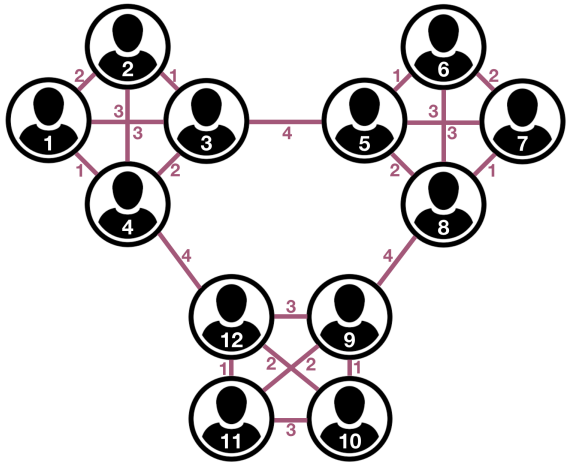


Figure 8.1: Network structure with 3 clusters, characterized by frequent within cluster interactions. Circles represent participants and links represent conversations. Numbers along the links indicate the conversational sequence.

Stimulus materials. The materials were borrowed from Vlasceanu & Coman (2018). They consisted of 24 belief statements grouped in four categories (2 myths and 4 correct pieces of information in each category). The myths and the facts were selected based on a pretested dataset collected on MTurk (112 participants), such that they were not statistically significantly different on believability, perceived scientific support, and personal relevance. In addition, the beliefs were correctly categorized as being part of a category by more than 75% of the sample (Vlasceanu & Coman, 2018). The myths were comprised of statements commonly endorsed by individuals as true, but in fact are demonstrably false, whereas the facts were scientifically ac-

curate statements. For example, a myth was that "reading in dim light can damage children's eyes," while a fact was that "children who spend less time outdoors are at greater risk to develop myopia."

Design and procedure. The 168 participants were split in 14 lab-created communities of 12 participants. Each community was assembled separately, and was comprised by individuals who arrived in the lab at the same time. Sessions were overbooked (by 2-3 participants) to ensure that I had the required community size even if some participants did not show up at the scheduled time. These communities were randomly assigned to either the Control (7 networks) or the Experimental (7 networks) condition. Once assigned to the condition, participants went through 5 experimental phases. In the study/evaluation phase, participants were presented with 24 statements that, they were told, are "frequently encountered on the internet." The presentation occurred in a randomized category-blocked fashion and participants were instructed to carefully read these statements. They were also asked to rate the degree to which they believe each statement is accurate, on a scale from 1-(Not at all) to 7-(Very much so) and has scientific support, on a scale from 1-(Definitely not) to 7-(Definitely yes). Next, participants in the Experimental condition listened to an audio of a participant who, supposedly, recalled the information to which s/he was exposed during the experiment in a previous session. In reality, the speaker was a confederate mentioning the statements with minor hesitations to indicate a naturalistic recall. Each participant listened to an audio containing half of the correct statements (i.e., 2 statements) from each of the 4 initially studied categories. Thus, each participant listened to 8 pieces of correct information in the audio, which constituted the RP+ (Retrieval Practice +) items. RP+ items were always correct pieces of information. The RP- (Retrieval Practice -) items were the 16 initially studied items not mentioned in the audio. Participants in the Control condition did not go through a selective practice phase, instead they completed an unrelated distracter

task. Next, participants engaged in sequential dyadic anonymous chat conversations (computer-mediated) as part of the conversational recall phase. Each participant took part in a sequence of three or four 3-minute conversations, during which they were asked to collaboratively remember as many of the statements they initially studied as possible: "In this phase you will have a series of chat conversations with other participants about the materials you studied. During these chat conversations you will be asked to jointly remember the information that you studied initially about child rearing. Please be patient and engaged in the task throughout." The conversations were characterized by turn-taking, with virtually all conversational recalls involving collaboration between the interacting partners. A qualitative analysis of the conversations revealed that all of the participants stayed on task throughout the duration of the study. The conversational sequence created a communicative network characterized by 3 clusters, with frequent within-cluster interactions, a network structure that mimics the types of networked interactions one might have in one's community (Watts & Strogatz, 1998). After the conversational recall phase, participants were randomly presented with the initially read beliefs and were asked to rate them on the same two 7-point scales as before (i.e., degree of belief and scientific support). This rating occurred at two time points, once immediately after the recall phase in the lab (post-evaluation phase) and one week later through a Qualtrics link (follow-up-evaluation phase). All stimuli and procedures were approved by Princeton University's Institutional Review Board (IRB).

I offer three points of clarification about the public speaker. First, I differentiate the usage of the term from the more colloquial usage implying a mass media context. I simply mean that the speaker's message is broadcasted to all participants in the community. Second, the participants were told that the public speaker only selectively remembered some of the beliefs due to time constraints, to diminish the possibility that participants inferred the importance of the beliefs based on whether

they were mentioned or not by the public speaker. And third, my main aim in implying that the speaker in the audio was another participant, was to reduce the impact of source credibility and expertise on the dependent variables.

Analysis and coding. Each belief was coded as successfully remembered if the recall captured the gist of the original statement. For instance, if for the statement "Eating carrots will make babies' eyesight sharper," participants remembered "Carrots improve vision," their recall was coded as accurate, since it captures the gist of the original statement. Ten percent of the data were double-coded for reliability (Cohen's kappa=0.93) and all disagreements were resolved through discussion between coders. Items that were negated in conversations were excluded from the analyses of both the participant who mentioned the statement as false, and their conversation partner. On average, for each participant, only a small number of beliefs (1.66 out of the 24) were discarded due to negation during the conversational phase, with no difference between the accurate and inaccurate beliefs. I decided to exclude these items because I am interested in observing the impact of mnemonic accessibility on statement believability, which would be contaminated by a discussion of their truth-value.

Results.

I hypothesized that a belief's mnemonic accessibility would impact its believability. First, I predicted that the public speaker would influence the mnemonic accessibility of the initially studied statements and cause believability changes across the community. Second, I expected the influence of the public speaker to be amplified in subsequent conversations. Statements mentioned by the public speaker should be more likely to be discussed in ensuing conversations (relative to the Control condition) and would experience an increase in believability from pre to post-conversation (belief rehearsal effect). Statements not mentioned by the public speaker, but related to those men-

tioned should be less likely to be remembered in subsequent conversations (relative to the Control condition), which would result in a decrease in believability from pre to post-conversation (belief suppression effect). Finally, these belief rehearsal and suppression effects should circumscribe the degree of belief convergence across the community.

I first wanted to establish whether there are differences in believability and memorability between the accurate (i.e., facts) and inaccurate beliefs (i.e., myths). Believability was measured by averaging the two highly-correlated evaluations (i.e., perceived accuracy and scientific support, r=0.80) for each belief. I found no differences between the pre-conversational believability scores for myths (M-Myths=4.22, SD=0.79) and for facts (M-Facts=4.19; SD=0.61), p=0.74. For memorability, I first computed the recall proportion of each belief by coding the conversational recalls of participants. I then averaged these conversational recall proportions across all rounds of conversation for each participant. This comparison only involved the Control condition, since the recall proportion of the beliefs in the Experimental condition was influenced by the status of the belief (i.e., RP+/RP-). There was no difference between the myths (M-Myths=0.31, SD=0.14) and the facts (M-Facts=0.29, SD=0.13), p=0.19. The two types of items were, therefore, indistinguishable from one another, as was found in preliminary studies involving these stimulus materials (Vlasceanu & Coman, 2018). Since there was no main effect for a variable coding for truth value of belief nor an interaction with other variables in the analyses, I collapsed across myths and facts here forth.

Next, I investigated whether listening to the public speaker leads to rehearsal and retrieval-induced forgetting effects on memory. I computed the recall proportion of beliefs mentioned in each participant's conversational recalls. Regardless of who brought up an item in the conversation, it was counted as remembered for both conversational participants, a method typically employed to address the interdependency

of individual recalls in conversational interactions (Kashy & Kenny, 1999); this decision was also based on previous studies documenting a similar effect size for speakers as for listeners (Coman, Manier, Hirst, 2009). For the Experimental condition, I averaged these scores across the rounds that each participant was part of, separating between items mentioned in the audio (RP+ items) and items related to those mentioned in the audio (RP- items). For the Control condition, I separated the beliefs according to the Experimental condition's RP+ beliefs and RP- beliefs, but note that in this condition none of the items were actually practiced before the conversational recalls started. Using these recall proportions as the dependent variable, I conducted a Mixed Factorial ANOVA, with Retrieval Type (RP+ vs. RP-) as a within-subject variable, and Condition (Experimental vs. Control) as a between-subject variable. I found a significant main effect for Retrieval Type, $F(1, 160)=105.99$, $p<0.001$, for Condition, $F(1, 160)=40.26$, $p<0.001$, and for their interaction, $F(1, 160)=286.37$, $p<0.001$. When exploring the interaction, I found that the recall proportion of RP+ items was significantly larger in the Experimental condition ($M=0.55$, $SD=0.18$) than in the Control condition ($M=0.25$, $SD=0.14$), ($p<0.001$, Cohen's $d=1.86$), suggesting that listening to the public speaker selectively practicing the beliefs leads to increased recall of those beliefs. Similarly, the recall proportion of RP- items was significantly lower in the Experimental condition ($M=0.27$, $SD=0.10$) than in the Control condition ($M=0.32$, $SD=0.12$), ($p<0.002$, Cohen's $d=0.45$) indicating that unmentioned beliefs related to those mentioned were forgotten, relative to the control condition (Figure 8.2).

Does this recall pattern result in believability changes? In other words, is believability dependent on the mnemonic accessibility of the belief? Given the recall data, I expected a belief rehearsal effect, such that practiced beliefs should increase in believability, and a belief suppression effect, such that beliefs related to the practiced ones should decrease in believability. To explore these predictions, I first standardized

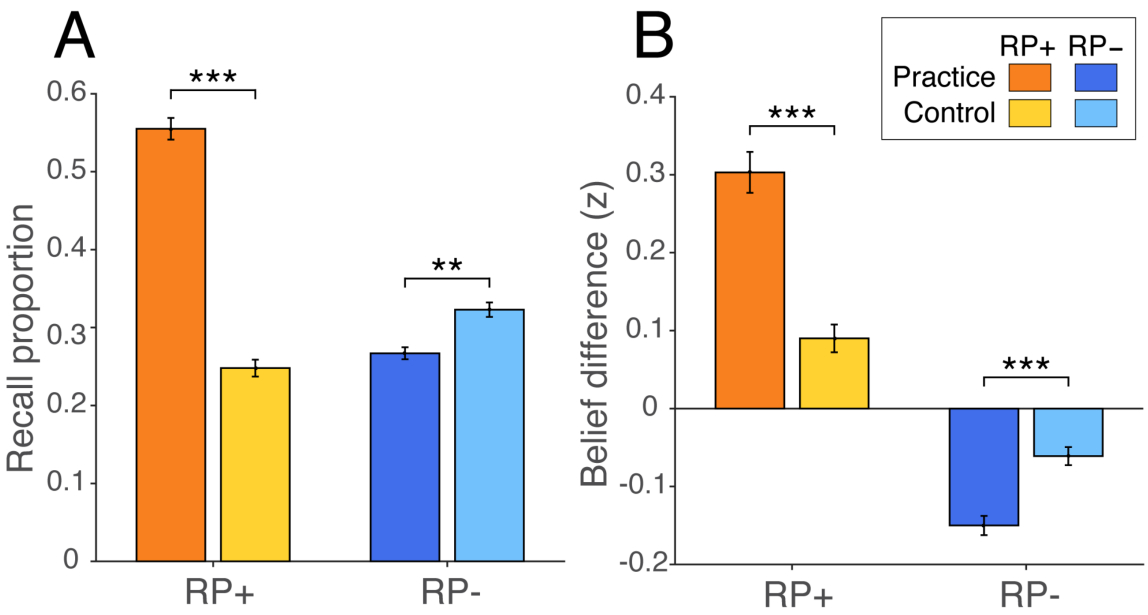


Figure 8.2: Panel A. The recall proportion of beliefs averaged over each participant's conversations and separate between the RP+ and RP- beliefs (Experimental) and their corresponding beliefs in the Control condition. Panel B. The post-pre belief difference score, separate between RP+ and RP- beliefs (Experimental) and their corresponding beliefs in the Control condition.

the belief ratings, using z-scores, within-participant. This standardization procedure allowed controlling for participant-specific particularities of rating scale use, ensuring any observed effects are not driven by outliers (Fischer & Milfont, 2010). I did so separately for the pre-conversational and post-conversational belief ratings. For each participant, I subtracted the pre-conversational z-score from the post-conversational z-score for each belief, separating between RP+ and RP- beliefs in the Experimental condition and the corresponding beliefs in the Control condition. Using this belief difference score as a dependent variable I conducted a Mixed Factorial ANOVA, with Retrieval Type (RP+ vs. RP-) as a within-subject variable and Condition (Experimental vs. Control) as a between-subject variable. I found a significant main effect for Retrieval Type, $F(1, 164)=78.22$, $p<0.001$, for Condition, $F(1, 160)=31.17$, $p<0.001$, and for their interaction, $F(1, 160)=19.31$, $p<0.001$. Post-hoc analyses showed that the RP+ beliefs became more believable (from pre- to post-conversation) in the Ex-

perimental condition (M=0.30, SD=0.34) than in the Control condition (M=0.09, SD=0.23), $p<0.001$, Cohen's d=0.72. Listening to a public speaker selectively practicing beliefs leads, thus, to increased believability on the part of the listener. Similarly, the RP- items became less believable (from pre- to post-conversation) in the Experimental condition (M=-0.15, SD=0.16) than in the Control condition (M=-0.06, SD=0.15), ($p<0.001$, Cohen's d=0.58), indicating a belief suppression effect for beliefs related to those mentioned (Figure 8.2B). Listening to a public speaker results, thus, in diminished believability for beliefs related to those mentioned. These results are consistent with the pattern I obtained with an individual-level paradigm (Vlasceanu & Coman, 2018).

I showed that mnemonic accessibility changes triggered by a public speaker's selective practice affects believability. I next assessed the conversational recall's independent impact on believability as well as the cumulative effect of the public speaker and the conversational recall on believability. To test these effects, I analyzed the content of participants' conversations, following previous work (Coman, Momennejad, Drach, Geana, 2016). I computed cumulative reinforcement/suppression (R/S) scores for each of the 24 initially studied beliefs for each participant as follows. If a belief was mentioned during a conversation, it received a (+1) score on the R/S scale. Similarly, if a belief was not mentioned during a conversation, but was related to a belief that was mentioned it received a (-1) score on the R/S scale. Unmentioned and unrelated to the mentioned beliefs received a score of 0 on the R/S scale. The final R/S score for each participant was cumulated across the three/four conversations s/he had in the network and was computed separately for each belief. For instance, if a belief was mentioned in all three conversations that a participant had in the network, then it's cumulative R/S score was (+3), while if the belief was part of the category mentioned during all the conversations that the participant was engaged in, but was itself never mentioned in any of the three conversations, then it's R/S cumu-

lative score was (-3). I did not account for the source of the information during the conversation (i.e., who was the speaker and who was the listener) since previous research showed that during conversational recall the speakers and listeners experience similar degrees of rehearsal and retrieval-induced forgetting effects (Coman & Hirst, 2012). I predicted that items with positive cumulative R/S scores will experience a belief rehearsal effect such that they will become more believable post-conversation (relative to pre-conversation), while items with negative R/S scores will experience a belief suppression effect, such that they will become less believable post-conversation (relative to pre-conversation). To capture these effects, for each belief I computed a belief difference score by subtracting its pre-conversational z-score from its post-conversational z-score. Since there was no participant who had belief z-score values for all nine R/S levels (-4 to 4), I collapsed all the beliefs that had positive R/S scores by averaging the belief z-scores across the R/S scores that ranged from (+1) to (+4). Similarly, I collapsed all negative R/S scores by averaging the belief z-scores across the (-4) to (-1) R/S scores. A positive value for this belief difference score indicates a belief rehearsal effect, whereas a negative value indicates a belief suppression effect.

I wanted to investigate whether the conversations had an independent effect on believability. Using the belief difference score as a dependent variable, I ran a Mixed Factorial ANOVA with R/S Item Type as a within-subject variable (Negative R/S; Zero; Positive R/S) and Condition (Experimental vs. Control) as a between-subject variable. I found a main effect for R/S Item Type, $F(2, 114)=6.10$, $p<0.003$, but not for Condition ($p=0.43$). As predicted, the interaction between R/S Item Type and Condition was significant, $F(2, 114)=3.42$, $p<0.036$. Posthoc analyses revealed evidence for the belief suppression effect, with Negative R/S scores being remembered worse in the Experimental condition ($M=-0.12$, $SD=0.22$) than in the Control condition ($M=-0.01$, $SD=0.23$), $t(164)=3.14$, $p<0.02$, Cohen's $d=0.49$. No belief rehearsal effect was found, even though the direction of the difference was consistent with the

hypothesis that Positive R/S scores would be remembered better in the Experimental condition (M=0.15, SD=0.39) than in the Control condition (M=0.06, SD=0.46) (p=0.19, Cohen's d=0.21) (Figure 8.3).

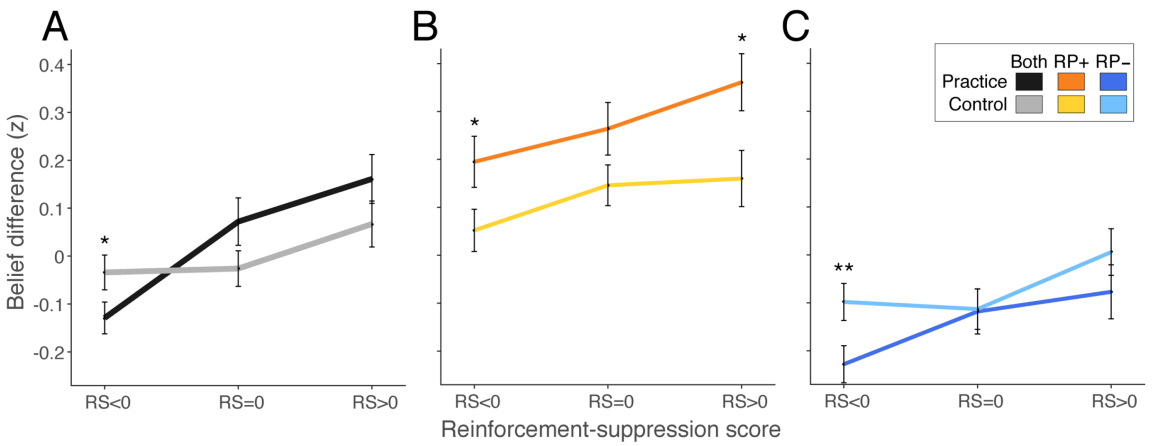


Figure 8.3: Panel A. The effect of conversational recall on belief difference. R/S indicates the Cumulative Reinforcement/Suppression score; beliefs with RS¡0 were unmentioned in the participant's conversations, but were related to those mentioned, while beliefs with RS¿0 were mentioned during the participant's conversations. Panel B. The effect is separated for the beliefs mentioned by the public speaker and their corresponding beliefs in the Control condition. Panel C. The effect is separated for the beliefs that were related to those mentioned by the public speaker and their corresponding beliefs in the Control condition.

This analysis did not differentiate, however, between items mentioned (vs. those related to those mentioned) by the public speaker. If the conversational recall amplified the effects triggered by the public speaker, I should observe higher belief difference scores between the Experimental and Control conditions for RP+ items that had positive R/S scores, and significantly lower belief difference scores between the Experimental and Control conditions for RP- items that had negative R/S scores. And indeed, both differences were statistically significant: $t(143)=2.31$, $p<0.022$, Cohen's $d=0.40$, and $t(164)=3.02$, $p<0.003$, Cohen's $d=0.46$, respectively (Figure 8.3B,C). This indicates that the impact of a public speaker on people's beliefs is stronger if its influence is further propagated in their subsequent conversations. It is the con-

junction of a public speaker's interventions and people's ensuing conversations that facilitate both the belief rehearsal and belief suppression effects.

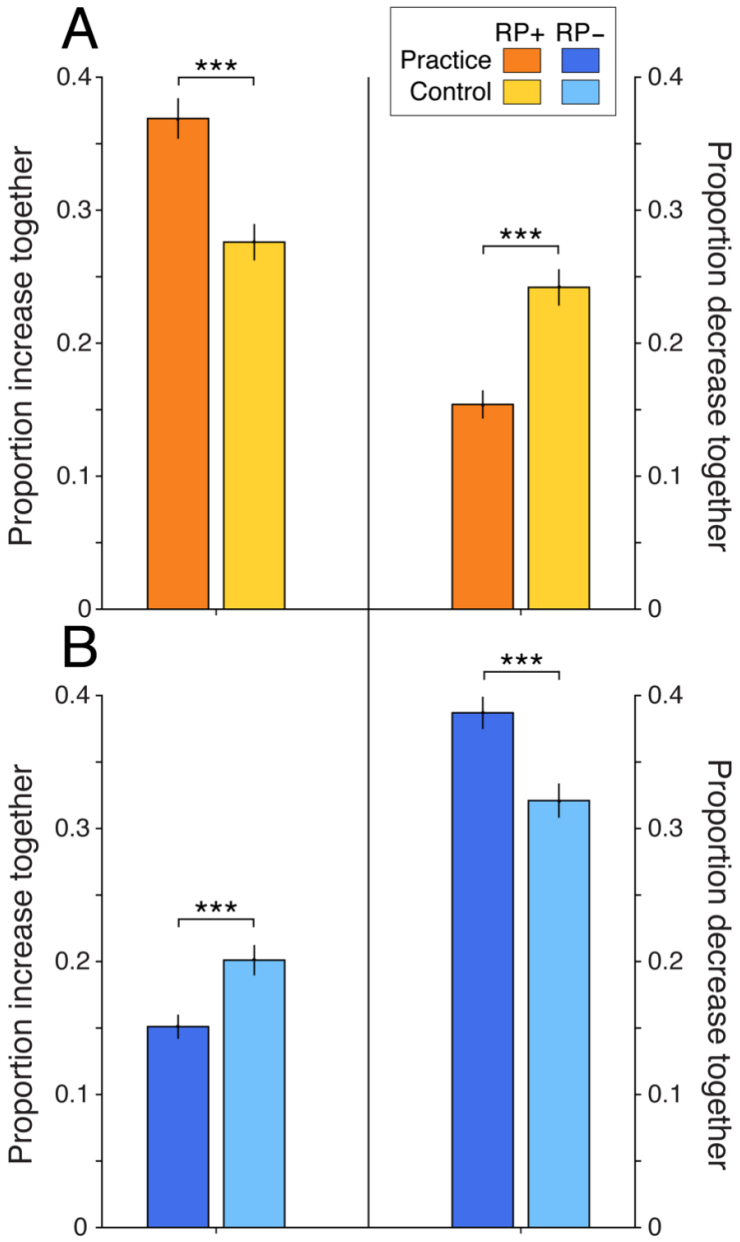


Figure 8.4: The proportion of beliefs that increase together or decrease together, from pre to post-conversation, averaged across all pairs of participants in the community, and separate for (A) RP+ beliefs (Experimental) and their corresponding beliefs in the Control condition and (B) RP- beliefs (Experimental) and their corresponding beliefs in the Control condition.

Does the public speaker lead to belief synchronization at a community level? To test for belief synchronization, I reasoned that the beliefs mentioned by the speaker (RP+) would be more likely to increase in believability in unison among community members, while beliefs related to those mentioned (RP-) would decrease in believability, again, in unison across the community, relative to the Control condition. I used the post-pre belief difference scores to compute the proportion of beliefs that increased together (in believability) and the proportion of beliefs that decreased together between every pair of participants in a network. I then separated between RP+ and RP- beliefs. For the Control condition, I computed the same proportions, for the items corresponding to the RP+/RP- beliefs. Using these pair-wise proportion scores as the dependent variable, I conducted a Mixed Factorial ANOVA with Retrieval Type (RP+ vs. RP-) and Pre-Post Dynamic (Increase vs. Decrease) as within-subject factors and Condition (Experimental vs. Control) as a between-subject variable. I found a significant three-way interaction, $F(1, 900)=151.39$, $p<0.001$. As predicted, the proportion of RP+ beliefs that increased together was higher in the Experimental ($M=0.37$, $SD=0.19$) than in the Control condition ($M=.28$, $SD=.17$), $p<0.001$, Cohen's $d=0.50$. Conversely, the proportion of RP+ beliefs that decreased together was lower in the Experimental ($M=0.15$, $SD=0.13$) than in the Control condition ($M=0.24$, $SD=0.17$), $p<0.001$, Cohen's $d=0.59$. The RP- beliefs exhibited the opposite pattern, such that the proportion of RP- items that decreased together was larger in the Experimental ($M=0.39$, $SD=0.15$) than in the Control condition ($M=0.32$, $SD=0.16$), $p<0.001$, Cohen's $d=0.45$, and the proportion of RP- items that increased together was lower in the Experimental ($M=0.15$, $SD=0.11$) than in the Control condition ($M=0.20$, $SD=0.14$), $p<0.001$, Cohen's $d=0.40$ (Figure 8.4). This pattern suggests that exposure to a public speaker triggers community-wide belief synchronization.

In summary, I found that a public speaker affects both what people remember and what they believe. Moreover, this impact leads to the synchronization of people's beliefs across the community. But how long-lasting are these effects? Previous research has found that the rehearsal and retrieval-induced forgetting effects can last for at least a week (Storm, Bjork, Bjork, 2012). No research to date has investigated how long-lasting the belief rehearsal and belief suppression effects are. In order to answer this question, I conducted a follow-up, 7.31 days (SD=1.29) after the initial participation in the study. After contacting all 168 participants from the first phase of the study, I collected belief evaluation data from 115 participants (59 participants in the Experimental and 56 participants in the Control condition).

The effect of the public speaker on belief endorsement at follow-up. I first subtracted the pre-conversational belief z-score from its follow-up z-score, separating between RP+ and RP- beliefs in the Experimental condition and the corresponding beliefs in the Control condition. Using this belief difference score as a dependent variable, I conducted a Mixed Factorial ANOVA, with Retrieval Type (RP+ vs. RP-) as a within-subject variable, and Condition (Experimental vs. Control) as a between-subject variable. I found a significant main effect for Retrieval Type, $F(1, 113)=62.33$, $p<0.001$, for Condition, $F(1, 113)=13.52$, $p<0.001$, and for their interaction, $F(1, 113)=6.35$, $p<0.013$. Post-hoc analyses that the RP+ beliefs became significantly more believable (from pre-conversation to follow-up) in the Experimental condition (M=0.29, SD=0.31) than in the Control condition (M=0.14, SD=0.27), $p<0.005$, Cohen's d=0.52, suggesting that listening to the public speaker selectively practicing beliefs leads to increased believability on the part of the listener one week after exposure. Similarly, the RP- items became marginally less believable (from pre-conversation to follow-up) in the Experimental condition (M=-0.15, SD=0.16) than in the Control condition (M=-0.08, SD=0.18), p=0.066, Cohen's d=0.41, indicating

a marginal belief suppression effect one week after initial exposure for beliefs related to those mentioned.

The independent and cumulative impact of conversational recall on belief endorsement at follow-up. As in previous analyses, in order to explore whether the conversations had an impact on believability at follow-up, I employed a Mixed Factorial ANOVA with R/S Item Type as a within-subject variable (Negative R/S; 0; Positive R/S) and Condition (Experimental vs. Control) as a between-subject variable. The dependent variable was the difference between a belief's z-score at follow-up and its z-score in the pre-conversational phase. I found a main effect for R/S Item Type, $F(2, 77)= 5.46$, $p<0.006$, but not for Condition ($p=0.67$). As predicted, I found an interaction between R/S Item Type and Condition, $F(2, 114)=3.72$, $p<0.029$. Posthoc tests revealed that the belief suppression effect was significant, with Negative R/S scores becoming less believable in the Experimental condition ($M=-0.11$, $SD=0.24$) than in the Control condition ($M=0.00$, $SD=0.21$), $t(113)=2.85$, $p<0.005$, Cohen's $d=0.49$. I also found a belief rehearsal effect, with Positive R/S scores becoming more believable in the Experimental condition ($M=0.14$, $SD=0.33$) than in the Control condition ($M=-0.03$, $SD=0.47$) $t(113)=2.20$, $p<0.03$, Cohen's $d=0.42$. This indicates that people's conversations meaningfully impact the believability scores of the studied beliefs one week after they occur.

I showed that the effect of the public speaker on people's beliefs is amplified in their subsequent conversations as measured immediately after the public speaker's intervention. Is this amplification effect long-lasting? If the conversational recall facilitated the effects triggered by the public speaker, I should observe significantly higher belief difference scores between the Experimental and Control conditions for RP+ items with positive R/S scores, and significantly lower belief difference scores between the Experimental and Control conditions for RP- items with negative R/S scores. Neither the belief rehearsal ($p=0.10$), nor the belief suppression ($p=0.20$)

effects reached statistical significance at follow-up. These results suggest that the cumulative impact of the public speaker and ensuing conversations on beliefs loses strength with time.

Is belief synchronization long-lasting? As in previous analyses, I used the increase together/decrease together pair-wise proportion scores as a dependent variable. This time, the increase/decrease was computed using the change between the belief's pre-conversational z-scores and the follow-up z-scores. I conducted a Mixed Factorial ANOVA with Retrieval Type (RP+ vs. RP-) and Pre-Post Dynamic (Increase vs. Decrease) as within-subject variables and Condition (Experimental vs. Control) as a between-subject variable. The three-way interaction was significant, $F(1, 309)=29.11$, $p<0.001$. Post-hoc analyses revealed that the synchronization pattern observed in the post-conversational analysis was also present at follow-up. The proportion of RP+ beliefs that increased together was higher in the Experimental condition ($M=0.38$, $SD=0.18$) than in the Control condition ($M=0.29$, $SD=0.17$), $p<0.001$, Cohen's $d=0.51$, while the proportion of RP+ beliefs that decreased together was lower in the Experimental ($M=0.16$, $SD=0.14$) than in the Control ($M=0.19$, $SD=0.15$) condition, $p<0.04$, Cohen's $d=0.21$. The proportion of RP- items that increased together was lower in the Experimental ($M=0.17$, $SD=0.10$) than in the Control condition ($M=0.21$, $SD=0.13$), $p<0.001$, Cohen's $d=0.34$, while the proportion of RP- items that decreased together was marginally higher in the Experimental condition ($M=0.34$, $SD=0.13$) than in the Control condition ($M=0.32$, $SD=0.16$), $p=0.083$, Cohen's $d=0.14$.

Discussion

I have shown that a public speaker influences a community's beliefs by impacting the mnemonic accessibility of those beliefs. Beliefs mentioned by the public speaker become more believable, while beliefs related to those mentioned become less believ-

able compared to a situation in which no public speaker existed. Importantly, these effects are amplified by conversations within networks, revealing a cumulative impact of the public speaker and the conversations. These effects also regulate the degree of belief convergence across communities, with networks exposed to a public speaker synchronizing their beliefs more than control networks. Finally, I observe the effects, with sizes typical of similar paradigms (Murayama et al., 2014; Vlasceanu & Coman, 2018; Coman et al., 2016), lasting for at least one week.

The strength of this approach to investigating collective belief endorsement is its controlled, experimental nature, which adds meaningfully to research investigating social network data from platforms such as Twitter, which is mostly correlational (Hilbert et al, 2016). It is important to note, however, that in real-world situations the social context in which a public speaker communicates information to an audience is far more complex than tested in the current investigation. Factors such as expertise, credibility, and similarity of the public speaker would likely affect the degree to which the audience integrates or resists the information that is conveyed (Fiske & Taylor, 2015). Moreover, real-world conversations can also be more complex than the ones elicited in my experiment, especially if they contain strong opinions and evaluations. However, my goal was to provide empirical evidence of the impact of mnemonic accessibility on collective beliefs under minimal conditions, which ensure a highly controlled experimental design. The investigation of these minimal conditions allows us to claim that it is mnemonic accessibility, rather than the characteristics of the social source of information, that triggered the observed belief change. Now that I established this effect, future studies can investigate the impact of variables such as the public speaker's expertise, credibility, and similarity on collective belief endorsement.

Remarkably, a 20-minute session in which participants were exposed to a public speaker and then engaged in networked conversations was sufficient to trigger belief

change that lasted for at least one week. The two sources of social influence (i.e., public speaker and conversational interactions) had an independent impact on people's beliefs, both immediately post-conversation and at follow-up. The influence of these sources was also cumulative, such that beliefs mentioned by the public speaker that were also rehearsed in the conversations became most believable among all beliefs, while beliefs related to those mentioned by the public speaker, which were also related to beliefs discussed in people's conversations became least believable. This cumulative effect was found to be temporally limited, though diminishing at follow-up, which points to the boundary conditions of the impact of mnemonic accessibility on believability. For the cumulative effect of the two sources of influence to be long-lasting, additional factors might need to be implemented. Repeated interactions over time following the public speaker's intervention (Centola, 2010) and the ideological consistency between the participants' beliefs and those espoused by the public speaker (Coman & Hirst, 2012) constitute two such factors that I plan to investigate in future research. In this experiment, I created a conversational network structure that mimicked the main characteristics of real-world social networks: clustered communities characterized by frequent within-cluster interactions (Watts & Strogatz, 1998). This methodology allowed us to situate individual-level cognitive processes in a framework (Vlasceanu, Enz, Coman, 2018) aimed at programmatically investigating how micro-level cognitive processes (i.e., rehearsal, retrieval-induced forgetting) could give rise to collective-level large-scale phenomena (i.e., the synchronization of beliefs across communities). I have not undertaken, however, a programmatic investigation of how the network structure, or temporal sequencing of conversations impacts collective-level outcomes. Manipulating the topological (Coman et al, 2016) and temporal (Momennejad, Duker, Coman, 2018; Li et al, 2017) features of a community's conversational network could lead to significant advances in understanding how collective beliefs are formed and maintained in networked communities.

Moreover, in the current study participants were exposed to information organized in meaningful categories. While organizing information into structured categories is a naturally occurring process (Chapman, 1967), there may be variation in how individuals spontaneously group information into categories. Thus, a noteworthy expansion of this investigation could be exploring how these idiosyncrasies interact with processes I explored herein. Finally, these findings provide promising possibilities for interventions aimed at countering the spread of inaccurate beliefs in vulnerable communities. A typical intervention targeted at reducing the spread of misinformation involves refuting misinformation by overtly discussing its false nature (Wegner, Wenzlaff, Kerker, Beattie, 1981). But refutations have been found to reinforce misconceptions (Lewandowsky et al, 2012), especially when individuals are confronted with a complex informational environment (Lamb, King, Kling, 2003; Contractor & DeChurch, 2014) and when they are ideologically committed (Nyhan & Reifler, 2010; Flaxman, Goel, Rao, 2016). I have shown here that in order to diminish the believability of inaccurate beliefs, one does not necessarily need to discuss them. It is sufficient to trigger response competition from the part of inaccurate beliefs by repeatedly broadcasting conceptually related accurate beliefs in the population (Vlasceanu & Coman, 2018). This will result in the suppression of misinformation, which, as I have shown here, will reduce their believability across the entire community.

Part IV

Conclusion

Chapter 9

Conclusion

9.1 Summary

Beliefs are a mental construct fundamental to human societies (Shariff & Rhemtulla, 2012; Mangels et al., 2006; Ajzen, 1991; Lund, 1925). Here, I introduce a theoretical generative framework for investigating factors influencing beliefs, which makes predictions about how beliefs could be changed (Chapter 1). Then, guided by this framework, in a series of online and lab experimental studies (N=5192) I show how psychological processes such as memory accessibility (Chapter 2), emotional arousal (Chapter 3), predictions errors (Chapter 4), and social norms (Chapter 5) can be leveraged to change people's beliefs. For instance, in a series of laboratory experiments, I find that strengthening the memory of a statement increases its believability, while weakening its memory in a targeted fashion decreases its believability (Chapter 2). Building on these findings, in another series of online experiments, I also show that pairing emotionally arousing images with statements subsequently increases the believability of these statements compared to statements that had been associated with neutral or no images (Chapter 3). Furthermore, in another series of online experiments including a US census matched sample, I explore the effect of prediction

errors on belief update, by testing the relationship between the size of the errors people make when engaging in predictions regarding belief related evidence, and the subsequent updating of the corresponding beliefs. I find that people update their beliefs as a function of the size of their errors, and that making large errors leads to more belief update than not engaging in prediction, while controlling for the evidence available. Importantly, I also find that these effects hold across ideological boundaries (Democrats and Republicans, evaluating Neutral, Democratic, and Republican beliefs; Chapter 4). Lastly in this part, in a series of online and lab studies, I show that people change their beliefs more in line with evidence portrayed as normative (e.g., shared by many on social media platforms) compared to evidence portrayed as non-normative (Chapter 5.1). Upon further investigation involving a US census matched sample of over a thousand US citizens, I also find that normativity cues signaled by large groups of people are the most effective at changing beliefs. This is true regardless of participants' political affiliation, suggesting that Democrats and Republicans are similarly affected by information sources (Chapter 5.2). All of these effects occur at the individual level of investigation, and recent advances in cognitive science suggest that cognition does not occur in a vacuum - instead, such cognitive processes are highly sensitive to the social context in which they manifest. Therefore, in another series of experiments this time at the collective level of investigation, I uncover how macro-level societal outcomes can emerge from micro-level psychological processes. I show how conversational interactions trigger belief change, as individuals talking to each other change their beliefs to match those of their conversational partners (Chapter 6). Then, I find that communities' network structures determine their members' collective beliefs synchronization (Chapter 7), and that individual level effects are amplified when people interact in social networks (Chapter 8). Beyond contributing to the psychological literature on belief formation, this exploration of individual and collective beliefs has important applications for misinformation prevention.

9.2 Policy Recommendations to Fight Misinformation

False beliefs (i.e.,misinformation) are among the top threats faced by the world today (Farkas & Schou, 2019; Lewandowsky et al, 2012). To effectively address this global epidemic, policy makers must act in ways that are guided by recommendations supported by empirical research (Oxman et al., 2010; Snilstveit, Vojtkova, Bhavsar, Gaarder, 2013; Reimers & McGinn, 1997). Understanding which cognitive processes are successful in changing false beliefs is a crucial first step in informing policies directed at preventing misinformation spread.

Typical misinformation reduction interventions involving refutation or debunking (Berinsky, 2017; Wegner, Wenzlaff, Kerker, & Beattie, 1981), have been found to backfire by reinforcing misconceptions in some contexts (Lewandowsky et al, 2012; Porter, Wood, Kirby, 2018; Swire & Ecker, 2018), such as when individuals are confronted with a complex informational environment (Contractor & DeChurch, 2014) or when ideologically committed (Flaxman, Goel, Rao, 2016). Emerging research has pointed to additional strategies of countering the spread of false information (Guess, Nagler, Tucker, 2019) such as prebunking or inoculating (van der Linden, Leiserowitz, Rosenthal, Maibach, 2017), by pre-emptively exposing people to small doses of misinformation techniques which has been shown to reduce susceptibility to fake news (Basol, Roozen-beek, van der Linden, 2020; Roozenbeek & van der Linden, 2019). Nudging accuracy is another stragegy that has been found useful at reducing belief in false information (Pennycook et al, 2020).

Leveraging the findings revealed in this dissertation, I propose additional strategies of reducing misinformation in vulnerable communities.

First, when targeting moderately endorsed inaccurate beliefs, I propose that in order to diminish their believability one does not necessarily need to discuss them.

Instead, it is sufficient to trigger response competition by repeatedly broadcasting conceptually related accurate beliefs in the population. This will result in the suppression of misinformation, which will reduce their believability at the individual level (Chapter 2) and across the entire community (Chapter 8).

Likewise, when targeting accurate beliefs, I suggest strengthening their believability by facilitating associations between true statements and negative images. These associations will increase the memorability of the accurate information, increasing its believability (Chapter 3).

Moreover, when attempting to shortcut ideological biases, I recommend incorporating a prediction-then-information format in public outreach actions. This procedure involves the need to, first, map the community's estimates on relevant statistics that can be used as surprising evidence. These statistics need to be carefully compiled given that people's predictions about everyday events are fairly accurate (Griffiths & Tenenbaum, 2006). After selecting the statistics eliciting the largest misestimates, these pieces of evidence need to be disseminated back to the community in a predictions-then-feedback format. This procedure is intensive but might have a stronger impact in diminishing misinformation than existing approaches (Chapter 4).

My findings also suggest the focus should be placed on communicating scientific evidence in support of accurate information, as opposed to communicating anecdotal evidence. Moreover, whenever available, normativity cues favoring accurate information should be made salient (e.g., "90% of Americans believe vaccines are safe" or conversely, "Only 10% of Americans believe vaccines cause autism"), as they can increase the endorsement of accurate information and decrease the endorsement of misinformation. This strategy is likely to be especially effective when the normativity cues are given by information sources such as large groups of people (Chapter 5).

Furthermore, when targeting communities instead of individuals, (Dovidio & Esses, 2007), the effects of conversational interactions between community members

should be taken into consideration as they strongly impact belief change (Chapter 6). For instance, my findings suggest such campaigns be sensitive of the conversational network structures of their targeted communities (Chapter 7). Knowing how members of these communities are connected, and leveraging the finding that conversations in social networks amplify individual level effects (Chapter 8), could inform intervention designers how communities with different connectivity structures might respond to their efforts.

In future studies I strive to identify additional such effects and recommendations, as well as to unveil their effects on individual and collective behavior. For instance, a promising future direction is investigating whether raising awareness about the mechanisms driving the newly emerging political divisions in the US, would increase people's incorporation of evidence across the ideological spectrum (Macy, Deri, Ruch, & Tong, 2019).

9.3 Future Directions: From Beliefs to Behavior

Behavior is consequential in every aspect of life. Understanding, predicting, and encouraging optimal collective behaviors in the population at large is therefore of critical concern in our society. Accordingly, entire research fields have been dedicating their resources to documenting predictors of behavior, to unveil ways in which individuals can be nudged towards behaviors beneficial to themselves and to society (Thaler & Sunstein, 2009). For instance, the theory of planned behavior (Ajzen, 1985, 1991), an extension of the Theory of Reasoned Action (Fishbein & Ajzen, 1975) and one of the most influential models for predicting human social behavior, proposes that human behavior is a function of behavioral intentions and perceived behavioral control. The intentions, in turn are hypothesized to be a function of the attitudes, beliefs, and social norms surrounding the behaviors in question. And such beliefs are posited to have

emerged given factors such as personality, demographic characteristics, and exposure to media (Fishbein & Ajzen, 1975). While the theory has gathered considerable empirical support especially in the domain of health behaviors, it suffers from critiques that it's primarily based on correlational designs (Noar & Zimmerman, 2005). Another theoretical account of behavior is the health belief model, one of the first social cognition models of behavior, using beliefs and attitudes to predict health-related decision making (Hochbaum, 1958; Rosenstock, 1960, 1974). The model's hypothesized predictors of behavior are susceptibility and severity of the health problem, benefits and barriers to the behavior, cues to action (e.g., media publicity), and self-efficacy. Early empirical support for the model was given by a tuberculosis study in which individuals who believed were susceptible to the disease and believed in the benefits of early detection were more likely to have a voluntary chest X-ray (82%) than individuals who didn't hold these beliefs (21%). Subsequent empirical support for the model summarized in reviews of the literature showed that perceived barriers was the strongest predictor across behaviors, followed by perceived susceptibility and perceived benefits; severity was the weakest predictor (Janz & Becker, 1984). Although widely influential, especially in cancer and HIV research, the health belief model has also been subject to criticism. The most commonly raised limitations of the model are the lack of empirical testing of the relationships and interactions between the model's predictors and the lack of empirical testing of the cues to action hypothesized predictor (Champion & Skinner, 2008). Moreover, this model is too specific, focusing only on health-related beliefs and corresponding behaviors, therefore lacking in generalizability to broader areas. Despite these models' limitations, a recurrent theme across them is the idea that people's beliefs is likely a meaningful predictor of behavior (Ajzen, 1991; Hochbaum, 1958).

Even though the literature documenting the relationship between beliefs and behavior is extensive, it suffers from four important limitations. First, the empirical

investigations into the relationship between beliefs and behaviors provided conflicting findings. While some studies found that beliefs influence behavior, for example religious beliefs predict crime rates (Shariff & Rhemtulla, 2012), and beliefs about intelligence predict learning success (Mangels, Butterfield, Lamb, Good, Dweck, 2006), others reported that beliefs are an unreliable predictor of behavior, for example, beliefs about outgroup members do not predict behaviors towards them (Paluck, 2009). This apparent inconsistency in results could suggest that one or more additional variables are moderating the relationship between the two constructs of interest. For example, it could be that the perceived social normativity of the behavior measured moderates whether beliefs will actually consistently translate into observable behavior. It could also be that this relationship depends on the nature of beliefs, such that neutral beliefs might impact behavior whereas ideologically charged beliefs might not. Beyond speculation, these interactions can be empirically tested in controlled experiments. Second, this literature rarely involves experimental manipulations, which constrains the causal links that could be inferred from such data. For instance, the theory of planned behavior, in which beliefs one of the predictors of behavior, is mostly supported by correlational research (Noar & Zimmerman, 2005). Shariff and Rhemtulla's (2012) cross-national study in which they reported that religious beliefs predict criminal behaviors, such that the more people believe in hell the lower the crime rates, is also correlational in nature. Similarly, the literature arguing that believing intelligence is a fixed entity versus a malleable one has different consequences for learning performance (Elliot & Dweck, 2013), is also primarily based on correlational data, despite efforts to establish a causal link through neural mechanisms supported by EEG data (Mangels et al, 2006). Third, prior studies explored this relationship at the individual level, missing meaningful emergent phenomena at the collective level (Vlasceanu, Enz, Coman, 2018), such as a potential amplification of the individual level effects (Vlasceanu, Morais, Duker, Coman, 2020), or the impact

of the network structure on the overall formation of collective behavior (Vlasceanu & Coman, 2020). Looking beyond the individual, and investigating these effects in social networks is particularly important for making policy recommendations, as policy makers are primarily interested in impacting communities (Dovidio & Esses, 2007). And fourth, prior work has been mainly conducted in W.E.I.R.D. cultures (Henrich, Heine, Norenzayan, 2010) restricted to the Western Educational setting of American or European university students and non-representative online participants from Industrialized, Rich, and Developed countries, thus lacking the generalizability to the wider human population and the ability to compare findings across groups and cultures. Exploring the effects at this level has the potential to reveal additional meaningful variables that may play a role in the relationship between beliefs and behavior, such as cultural tightness/looseness (Harrington & Gelfand, 2014) or national identification (Huddy & Khatib, 2007).

Building on both canonical and recent psychological findings and methodological advances (Coman, Momennejad, Drach, Geana, 2016; Van Bavel et al, 2020), in future work I plan to address all of these limitations. As a first step, I will conduct studies to establish the relationship between beliefs and behavior by experimentally triggering belief change and observing the effects on behavior (pilot data suggest a strong causal effect; Vlasceanu & Van Bavel, in prep). I then plan to explore several variables that might moderate this relationship, such as ideology or perception of normativity. Subsequently, I plan to investigate this relationship in social networks of interconnected individuals, unveiling mechanisms underlying collective behavior change. Finally, recognizing that most psychological research to date has been based on a "small corner of the human population", an impediment to identifying universal principles of human psychology (Arnett, 2008), I will run the experimental designs in cross-cultural samples using a multinational collaborative network (Van Bavel,..,Vlasceanu,..et al, 2020). Culture is the information (e.g., beliefs, habits, ideas) learned from others that

is capable of influencing behavior in a group of people who share context and experience (Heine, 2008). The common features across all definitions of culture include the notion of a group with shared behaviors, values, and beliefs that are passed from generation to generation (Keith, 2011). At this level of investigation, new variables become available for exploration. One such variable is the tightness/looseness dimension of human societies. Recent research has noted how tight and loose cultures vary in modern societies, dimensions critical for promoting cross-cultural coordination in a world of increasing global interdependence. Nations can be "tight", meaning they have strong norms and a low tolerance of deviant behavior, or "loose", meaning they have weak norms and a high tolerance of deviant behavior (Gelfand et al, 2011). Another construct of great cross-cultural interest is national identification. National identification is the personal significance that being part of a nation holds for an individual (Cameron, 2004; Leach et al., 2008). Prior work has found that national identity plays an important role in motivating people to engage in costly behavior that benefits other members of their national community (Kalin & Sambanis, 2018) and greater civic involvement (Huddy & Khatib, 2007). Accordingly, a strong sense of shared national identity might help promote collective efforts within one's country (e.g., Dovidio, Ikizer, Kunst, Levy, 2020).

Understanding the impact of beliefs on behaviors at all of these different levels of complexity will be essential in promoting optimal collective behaviors, a critical concern to a wide range of individuals from policy makers interested in pro-environmental behaviors to public health officials interested in preventative health behaviors.

Bibliography

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British journal of social psychology*, 29(2), 97-119.

Ajzen, I. *The theory of planned behaviour: Reactions and reflections*. (2011).

Alexander, W. H., and Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* 14, 1338–1344.

Allan, S., & Zelizer, B. (Eds.). (2004). *Reporting war: Journalism in wartime*. London: Routledge.

Allport, F. H., & Lepkin, M. (1945). Wartime rumors of waste and special privilege: Why some people believe them. *Journal of Abnormal and Social Psychology*, 40, 3–36.

Amir, Y., & Sharon, I. (1990). Replication research: A̋usẗfor the scientific advancement of psychology. *Journal of Social Behavior and Personality*, 5(4), 51.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering Can Cause Forgetting: Retrieval Dynamics in Long-Term Memory. *J of Exp Psychol: LMC*, 20(5), 1063–1087.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1), 3.

Ballhaus, R., Armour, S., & Leary, A. (2020). Trump Hopes to Have US Reopened by Easter, Despite Health Experts' Warnings. Retrieved from https://www.wsj.com/articles/trump-hopes-to-have-u-s-reopened-by-easter-despite-healthexperts-guidance-11585073462

Bar, M. (2009). Predictions: a universal principle in the operation of the human brain. *Introduction. Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1181–1182.

Bartels, L. M. (2018). *Unequal democracy: The political economy of the new gilded age*. Princeton University Press.

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

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121(4), 446.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3), 800.

Benegal, S. D., & Scruggs, L. A. (2018). Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic change*, 148(1-2), 61-80.

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

Berger, J. (2014). Word of mouth and interpersonal communication: a review and directions for future research. *J. Consumer Psychol.* 24, 586–607.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British journal of political science*, 47(2), 241-262.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, 20(3), 351-368.

Berkowitz, A. D. (2005). An overview of the social norms approach. Changing the culture of college drinking: *A socially situated health communication campaign*, 1, 193-214.

Bestmann, S., Harrison, L. M., Blankenburg, F., Mars, R. B., Haggard, P., Friston, K. J., et al. (2008). Influence of uncertainty and surprise on human corticospinal excitability during preparation for action. *Curr. Biol.* 18, 775–780.

Bendixen, L. D. (2002). A process model of epistemic belief change. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (p. 191–208). Lawrence Erlbaum Associates Publishers.

Bhattacharya, S., Srinivasan, P., & Polgreen, P. (2014). Engagement with health agencies on twitter. *PLoS One*, 9(11), e112235.

Blumen, H. M., & Rajaram, S. (2008). Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory*, 16(3), 231-244.

Borge, M., Ong, Y. S., & Rosé, C. P. (2018). Learning to monitor and regulate collective thinking processes. *International Journal of Computer-Supported Collaborative Learning*, 13(1), 61-92.

Bouvier, A. (2004). Individual beliefs and collective beliefs in sciences and philosophy: The plural subject and the polyphonic subject accounts: Case studies. *Philosophy of the social sciences*, 34(3), 382-407.

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.

Brandt, M.J. & Sleegers, W.W.A. (2021). Evaluating Belief System Networks as a Theory of Political Belief System Dynamics. *Personality and Social Psychology Review*. doi:10.1177/1088868321993751

Briñol, P., Petty, R. E., & Barden, J. (2007). Happiness versus sadness as a determinant of thought confidence in persuasion: A self-validation analysis. *Journal of Personality and Social psychology*, 93(5), 711.

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6), 977.

Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). β-Adrenergic activation and memory for emotional events. *Nature*, 371(6499), 702.

Cameron, J. E. (2004). A three-factor model of social identity. *Self and Identity*, 3, 239-262.

Carpenter, C.J. (2010). A Meta-Analysis of the Effectiveness of Health Belief Model Variables in Predicting Behavior. *Health Comm.*, 25:8, 661-669.

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060), 1269-1272.

Centola D (2010) The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197.

Cermak, L. S. & Craik, F. I. M. (1979) *Levels of Processing in Human Memory* (Erlbaum, Hillsdale, NJ).

Chaiken, S., Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior*. 553–578. The Guilford Press.

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4), 556-577.

Chung, S., Fink, E. L., & Kaplowitz, S. A. (2008). The comparative statics and dynamics of beliefs: The effect of message discrepancy and source credibility. *Communication Monographs*, 75(2), 158-189.

Cialdini, R. B., & Trost, M. R. (1998). *Social influence: Social norms, conformity and compliance.*

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55, 591-621.

Colgrove, J., & Bayer, R. (2005). Could it happen here? Vaccine risk controversies and the specter of derailment. *Health Affairs*, 24, 729–739.

Coman, A., & Berry, J. N. (2015). Infectious cognition: Risk perception affects socially shared retrieval-induced forgetting of medical information. *Psychological science*, 26(12), 1965-1971.

Coman, A., Manier, D., & Hirst, W. (2009). Forgetting the unforgettable through conversation: Socially shared retrieval-induced forgetting of September 11 memories. *Psychological Science*, 20(5), 627-633.

Coman A, Hirst W (2012) Cognition through a social network: The propagation of induced forgetting and practice effects. *J Exp Psychol: Gen* 141(2):321–336.

Coman, A., & Hirst, W. (2015). Social identity and socially shared retrieval-induced forgetting: The effects of group membership. *Journal of Experimental Psychology: General*, 144(4), 717.

Coman, A., Momennejad, I., Drach, R. D., & Geana, A. (2016). Mnemonic convergence in social networks: The emergent properties of cognition at a collective level. *Proceedings of the National Academy of Sciences*, 113(29), 8171-8176.

Connors, M. H., & Halligan, P. W. (2015). A cognitive account of belief: A tentative road map. *Frontiers in psychology*, 5, 1588.

Contractor, N. S., & DeChurch, L. A. (2014). Integrating social networks and human social motives to achieve social influence at scale. *PNAS*, 111, 13650–13657.

Cook, J., Ecker, U. K. H., & Lewandowsky, S. (2015). Misinformation and its correction. *Emerging Trends in the Social and Behavioral Sciences*, 1–17.

Craft, S., Ashley, S., & Maksl, A. (2017). News media literacy and conspiracy theory endorsement. *Communication and the Public*, 2(4), 388-401.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, 104(3), 268.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3), e57410.

Cuc, A., Koppel, J., & Hirst, W. Silence is not golden: A case for socially shared retrieval-induced forgetting. *Psychological Science*, 18(8), 727-733. (2007).

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, 123, 2–22.

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.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception?. *Trends in cognitive sciences*, 22(9), 764-779.

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, 3, 548.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 14(2), 273-291.

Dovidio, J. F., & Esses, V. M. (2007). Psychological research and public policy: Bridging the gap. *Social Issues and Policy Review*, 1, 5–14.

Dovidio, J. F., Ikizler, E. G., Kunst, J. R., Levy, A. (2020). Common identity and humanity. *Together apart the psychology of COVID-19*, 142-146.

Dunn, E. W., & Spellman, B. A. (2003). Forgetting by remembering: Stereotype inhibition through rehearsal of alternative aspects of identity. *Journal of Experimental Social Psychology*, 39, 420–433.

Echterhoff, G., Higgins, E. T., & Groll, S. (2005). Audience-tuning effects on memory: the role of shared reality. *Journal of personality and social psychology*, 89(3), 257.

Ecker, U. K. H., Lewandowsky, S., Apai, J. (2011). Terrorists brought down the plane! —No, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology*, 64, 283–310.

Ecker, U. K. H., Lewandowsky, S., Swire, B., Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18, 570–578.

Ecker, U. K. H., Lewandowsky, S., Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38, 1087–1100.

Ellis, E.G. (2020). The coronavirus outbreak is a petri dish for conspiracy theories. Wired https://www.wired.com/story/coronavirus-conspiracy-theories/

Erickson, C. A., and Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J. Neurosci.* 19, 10404–10416.

Escalas, J. E. (2007). Self-referencing and persuasion: Narrative transportation versus analytical elaboration. *Journal of Consumer Research*, 33(4), 421-429.

Farkas, J., & Schou, J. (2019). *Post-truth, Fake News and Democracy: Mapping the Politics of Falsehood*. Routledge.

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.

Fein, S., McCloskey, A. L., Tomlinson, T. M. (1997). Can the jury disregard that information? The use of suspicion to reduce the prejudicial effects of pretrial publicity and inadmissible testimony. *Personality and Social Psychology Bulletin*, 23, 1215-1226.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, 58(2), 203.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*. Thousand Oaks.

Flaxman, S.and Goel, S. and Rao, J.M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80, 298–320.

Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12), 5334-5338.

Frankel, R., & Swanson, S. R. (2002). The impact of faculty-student interactions on teaching behavior: An investigation of perceived student encounter orientation, interactive confidence, and interactive practice. *Journal of Education for Business*, 78(2), 85-91.

Frenkel, S., Alba, D., & Zhong, R. (2020). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times*.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3), 215-239.

Garrett, R.K. (2011). Troubling consequences of online political rumoring. *Human Communication Research* 37(2): 255–274.

Garthwaite, C., & Moore, T. (2008). The role of celebrity endorsements in politics: Oprah, Obama, and the 2008 democratic primary. Department of Economics, University of Maryland, 1-59.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ... Aycan, Z. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033), 1100-1104.

Gilbert, M. (1987), 'Modelling Collective Belief', *Synthese* 73(1), 185-204.

Gilbert, M. (1994), 'Remarks on Collective Belief', in F. Schmitt (ed.), *Socializing Epistemology: The Social Dimensions of Knowledge*, Rowman and Littlefield, Lanham, MD, 111-134.

Gilbert, M. P. (2000). Collective belief and scientific change. *Sociality and responsibility: New essays in plural subject theory.*

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *JPSP*, 59, 601–613.

Gilbert, D. T., Tafarodi, R. W., Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221–233.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3), 213-224.

Grady, C. L., McIntosh, A. R., Rajah, M. N., & Craik, F. I. (1998). Neural correlates of the episodic encoding of pictures and words. *Proceedings of the National Academy of Sciences*, 95(5), 2703-2708.

Grant, A. M., & Hofmann, D. A. (2011). Outsourcing inspiration: The performance effects of ideological messages from leaders and beneficiaries. *Organizational Behavior and Human Decision Processes*, 116(2), 173-187.

Green, L. W., Ottoson, J. M., Garcia, C., & Hiatt, R. A. (2009). Diffusion theory and knowledge dissemination, utilization, and integration in public health. *Annual review of public health*, 30.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power. *Statistics in medicine*, 2(2), 243-251.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of memory and language*, 94, 149-165.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, 17(9), 767-773.

Grynbaum, M., & Abrams, R. (2020). Right-Wing Media Says Virus Fears Were Whipped Up to Hurt Trump.

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.

Haas, I. J., & Cunningham, W. A. (2014). The uncertainty paradox: Perceived threat moderates the effect of uncertainty on political tolerance. *Political Psychology*, 35(2), 291-302.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left-right: Ideological narratives and moral foundations. *Psychological Inquiry*, 20, 110–119.

Harber, K. D., & Cohen, D. J. (2005). The emotional broadcaster theory of social sharing. *Journal of Language and Social Psychology*, 24(4), 382-400.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1), 107-112.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1), 400-407.

Hilbert, M., Vásquez, J, Halpern, D., Valenzuela, S., Arriagada, E. (2016). One Step, Two Step, Network Step? Complementary Perspectives on Communication

Flows in Twittered Citizen Protests. *Social Science Computer Review*, 35 (4), 444 – 461.

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Annual review of psychology*, 63, 55-79.

Hmielowski, J. D., Feldman, L., Myers, T. A., Leiserowitz, A., & Maibach, E. (2014). An attack on science? Media use, trust in scientists, and perceptions of global warming. *Public Understanding of Science*, 23(7), 866-883.

Hochbaum, G. M. (1958). Public participation in medical screening programs: A socio-psychological study (No. 572). US Department of Health, Education, and Welfare, Public Health Service, Bureau of State Services, Division of Special Health Services, Tuberculosis Program.

Höijer, B. (2010). Emotional anchoring and objectification in the media reporting on climate change. *Public Understanding of Science*, 19(6), 717-731.

Hollander, B. A. (2018). Partisanship, individual differences, and news media exposure as predictors of conspiracy beliefs. *Journalism & Mass Communication Quarterly*, 95(3), 691-713.

Hope, L., & Wright, D. (2007). Beyond unusual? Examining the role of attention in the weapon focus effect. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(7), 951-961.

Hough-Telford, C., Kimberlin, D. W., Aban, I., Hitchcock, W. P., Almquist, J., Kratz, R., & O'Connor, K. G. (2016). Vaccine delays, refusals, and patient dismissals: a survey of pediatricians. Pediatrics, e20162127.

Huddy, L., & Khatib, N. (2007). American patriotism, national identity, and political involvement. *American journal of political science*, 51(1), 63-77.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus [Contribution to the theory of ferromagnetism.]. Zeitschrift fur Physik, 31, 253–258.

Janz, N. K., & Becker, M. H. (1984). The health belief model: A decade later. *Health education quarterly*, 11(1), 1-47.

Jennings, F. J. (2018). Where to turn? The influence of information source on belief and behavior. *Journal of Risk Research*, 1-10.

Jervis, R. (2006). Understanding beliefs. *Political psychology*, 27(5), 641-663.

Jiménez, L., & Mendez, C. (1999). Which attention is needed for implicit sequence learning?. *Journal of experimental Psychology: learning, Memory, and cognition*, 25(1), 236.

Jolley, D., & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one*, 9(2), e89177.

Jolley, D., Meleady, R., & Douglas, K. M. (2020). Exposure to intergroup conspiracy theories promotes prejudice which spreads across groups. *British Journal of Psychology*, 111(1), 17-35.

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

Kalin, M., & Sambanis, N. (2018). How to think about social identity. *Annual Review of Political Science*, 21, 239-257.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319, 966–968.

Katz & Lazarsfeld (1955). *Personal Influence*. New York: Free Press.

Kashy, D. A., & Kenny, D. A. (1999). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology*. New York: Cambridge University Press.

Kim, J. W. (2018). They liked and shared: Effects of social media virality metrics on perceptions of message influence and behavioral intentions. *Computers in Human Behavior*, 84, 153-161.

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, 358(6364), 776-780.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, 45, 142–152.

Knobloch-Westerwick, S., & Meng, J. (2009). Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, 36(3), 426-448.

Knobloch, S., Hastall, M., Zillmann, D., & Callison, C. (2003). Imagery effects on the selective reading of Internet newsmagazines. *Communication Research*, 30(1), 3-29.

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, 10, 908–914.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics*, 62(3), 790-816.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. K. (2016). From social to sale: The effects of firm-generated content in social media on customer behavior. *Journal of Marketing*, 80(1), 7-25.

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863.

Lamb, R., King, J. L., & Kling, R. (2003). Informational environments: Organizational contexts of online information use. *Journal of the American Society for Information Science and Technology*, 54, 97-114.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., Ratzan, S. C. (2011). Addressing the vaccine confidence gap. *The Lancet*, 378, 526–535.

Lawton, G. (2015). Beyond belief. *New Scientist*, 226(3015), 28-33.

Leach, C. W., Van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., ... Spears, R. (2008). Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology*, 95, 144-165.

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

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2015). The robust relationship between conspiracism and denial of (climate) science. *Psychological Science*, 26(5), 667-670.

Li, A., Cornelius, S. P., Liu, Y. Y., Wang, L., & Barabási, A. L. (2017). The fundamental advantages of temporal networks. *Science*, 358(6366), 1042-1046.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, 49(2), 433-442.

Liu, B. F., Jin, Y., & Austin, L. L. (2013). The tendency to tell: Understanding publics' communicative responses to crisis information form and source. *Journal of Public Relations Research*, 25(1), 51-67.

Litman, L., Robinson, J. & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442.

Loftus, E. F. (1979). The malleability of human memory: Information introduced after we view an incident can transform memory. *American Scientist*, 67(3), 312-320.

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus". *Law and Human Behavior*, 11(1), 55-62.

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

Lowndes, C. (2019, March 1). How the Hindenburg killed an entire industry. Retrieved from https://www.youtube.com/watch?v=g9bkQ7OiEdQ

Lyles, C.R., Lopez, A., Pasick, R., Sarkar, U., 2013. 5̈ mins of uncomfyness is better than dealing with cancer 4 a lifetime¨ an exploratory qualitative analysis of cervical and breast cancer screening dialogue on Twitter. *J. Cancer Educ.* 28, 127–133.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological review*, 82(4), 276.

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science advances*, 5(8), eaax0754.

Mangels, J. A., Butterfield, B., Lamb, J., Good, C., & Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social cognitive and affective neuroscience*, 1(2), 75-86.

McCauley, C., & Jacques, S. (1979). The popularity of conspiracy theories of presidential assassination: A Bayesian analysis. *Journal of Personality and Social Psychology*, 37(5), 637.

McQuiggan, S. W., Rowe, J. P., Lee, S., & Lester, J. C. (2008). Story-based learning: The impact of narrative on learning experiences and outcomes. In *International Conference on Intelligent Tutoring Systems* (pp. 530-539). Springer, Berlin, Heidelberg.

Mensink, G. J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, 95(4), 434.

Miller, A. (2006). Watching viewers watch TV: Processing live, breaking, and emotional news in a naturalistic setting. *Journalism & Mass Communication Quarterly*, 83(3), 511-529.

Miller, D. T., & Prentice, D. A. (1996). *The construction of social norms and standards.*

Milgram, S. (1967). The small world problem. *Psychology Today*, 2, 60–67.

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, 60(4), 824-844.

Momennejad, I., Duker, A., & Coman, A. (2019). Bridge ties bind collective memories. *Nature communications*, 10(1), 1-8.

Montanaro, D. (2020). FACT CHECK: Coronavirus Is Not The Flu, Despite Trump's Comparison. Retrieved from https://www.npr.org/sections/coronavirus-live-updates/2020/03/24/820797301/factcheck-trump-compares-coronavirus-to-the-flu-but-they-are-not-the-same

Mooney, C. (2012). *The Republican brain: The science of why they deny science–and reality.* John Wiley & Sons.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of personality and social psychology*, 64(5), 723.

Murayama, K., Mityatsu, T., Buchli, D., Storm, B. C. (2014). Forgetting as a consequence of retrieval: a meta-analysis of retrieval-induced forgetting. *Psychol Bull*, 140(5), 1383–1409.

National Consumers League. (2014, April). Survey: One third of American parents mistakenly link vaccines to autism.

Newhagen, J. E. (1998). TV news images that induce anger, fear, and disgust: Effects on approach-avoidance and memory. *Journal of Broadcasting & Electronic Media*, 42(2), 265-276.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19(5), 969-974.

Newman, E. L., & Norman, K. A. (2010). Moderate excitation leads to weakening of perceptual representations. *Cerebral Cortex*, 20(11), 2760–2770.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, 14(9), 1105.

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

Nyhof, M., & Barrett, J. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of cognition and culture*, 1(1), 69-100.

Oliver, J. E., & Wood, T. (2014). Medical conspiracy theories and health behaviors in the United States. *JAMA internal medicine*, 174(5), 817-818.

Osterholm MT, Moore KA, Kelley NS, Brosseau LM, Wong G, Murphy FA, et al. (2015). Transmission of Ebola viruses: what we know and what we do not know. *MBio*. 6.

Ozubko, J. D., & Fugelsang, J. (2011). Remembering makes evidence compelling: Retrieval from memory can give rise to the illusion of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 270.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 80.

Paluck, E. L., & Green, D. P. (2009). Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda. *American political Science review*, 622-644.

Pasek, J., Stark, T. H., Krosnick, J. A., & Tompson, T. (2015). What motivates a conspiracy theory? Birther beliefs, partisanship, liberal-conservative ideology, and anti-Black attitudes. *Electoral Studies*, 40, 482-489.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865-1880.

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*

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. (2020). Predictors of attitudes and misperceptions about COVID-19 in Canada, the UK, and the USA.

Petty, R. E., & Briñol, P. (2015). Emotion and persuasion: Cognitive and metacognitive processes impact attitudes. *Cognition and Emotion*, 29(1), 1-26.

Pew Research Center. (2017). News Use Across Social Media Platforms 2017. Retrieved from https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Pew Research Center. (2020). Survey of U.S. adults conducted March 10-16 2020. Retrieved from https://www.pewresearch.org/fact-tank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-was-made-in-a-lab/

Poland, G. A., & Spier, R. (2010). Fear, misinformation, and innumerates: How the Wakefield paper, the press, and advocacy groups damaged the public health. *Vaccine*, 28, 2361-2362.

Poppenk, J., & Norman, K. A. (2014). Briefly cuing memories leads to suppression of their neural representations. *Journal of Neuroscience*, 34, 8010–8020.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34, 243–281.

Porter, E., Wood, T. J., & Kirby, D. (2018). Sex trafficking, Russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news. *Journal of Experimental Political Science*, 5(2), 159-164.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76(2), 270-279.

Radnitz, S., & Underwood, P. (2017). Is belief in conspiracy theories pathological? A survey experiment on the cognitive roots of extreme suspicion. *British Journal of Political Science*, 47(1), 113-129.

Ranney, M., Cheng, F., Garcia de Osuna, J., & Nelson, J. (2001). Numerically driven inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates. In *Annual Meeting of the Society for Judgment & Decision Making.*

Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, 8(1), 49-75.

Ratzan, S. C. (2010). Editorial: Setting the record straight: Vaccines, autism, and The Lancet. *Journal of Health Communication*, 15, 237-239.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning, 2: Current research and theory* (pp. 64–69). Appleton-Century-Crofts.

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5), 570-580.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4), 296-320.

Saad, L. (2013). Americans' top critique of GOP: "unwilling to compromise". Gallup, April, 1.

Saker, L., Lee, K., Cannito, B., Gilmore, A., & Campbell-Lendrum, D. H. (2004). Globalization and infectious diseases: a review of the linkages (No. TDR/STR/SEB/ST/04.2). World Health Organization.

Schildkraut, D.J. (2010). *Americanism in the Twenty-First Century: Public Opinion in the Age of Immigration.* Cambridge University Press.

Schul, Y., & Mazursky, D. (1990). Conditions facilitating successful discounting in consumer decision making. *Journal of Consumer Research*, 16, 442–451.

Schwarz, N., Sanna, L., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, 39, 127–161.

Scott, S., & Duncan, C. J. (2001). *Biology of plagues: evidence from historical populations.* Cambridge University Press.

Shariff, A. F., & Rhemtulla, M. (2012). Divergent effects of beliefs in heaven and hell on national crime rates. *PloS one*, 7(6), e39048.

Shi, Z., Wang, A. L., Emery, L. F., Sheerin, K. M., & Romer, D. (2016). The importance of relevant emotional arousal in the efficacy of pictorial health warnings for cigarettes. *Nicotine & Tobacco Research*, 19(6), 750-755.

Shoshani, A., & Slone, M. (2008). The drama of media coverage of terrorism: Emotional and attitudinal impact on the audience. *Studies in conflict & terrorism*, 31(7), 627-640.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91(4), 531-553.

Shermer, M. (2011). The believing brain: From ghosts and gods to politics and conspiracies. How we construct beliefs and reinforce them as truths. Macmillan.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

Slater, M. D., & Rouner, D. (1996). How message evaluation and source attributes may influence credibility assessment and belief change. *Journalism & Mass Communication Quarterly*, 73(4), 974-991.

Slater, M. D., Buller, D. B., Waters, E., Archibeque, M., & LeBlanc, M. (2003). A test of conversational and testimonial messages versus didactic presentations of nutrition information. *Journal of nutrition education and behavior*, 35(5), 255-259.

Smallpage, S. M., Enders, A. M., & Uscinski, J. E. (2017). The partisan contours of conspiracy theory beliefs. *Research & Politics*, 4(4), 2053168017746554.

So, J., Prestin, A., Lee, L., Wang, Y., Yen, J., Chou, W.Y.S., 2016. What do people like to "share" about obesity? A content analysis of frequent retweets about obesity on Twitter. *Health Commun.* 31, 193–206.

Sperber, D. (1996). *Explaining culture: A naturalistic approach* (Vol. 323). Oxford: Blackwell.

Starbird, K. (2019). Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766), 449.

Stempel, C., Hargrove, T., & Stempel III, G. H. (2007). Media use, social structure, and belief in 9/11 conspiracy theories. *Journalism & Mass Communication Quarterly*, 84(2), 353-372.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2012). On the durability of retrieval-induced forgetting. *Journal of Cognitive Psychology*, 24, 617-629.

Swire, B., & Ecker, U. K. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. *Misinformation and mass audiences*, 195-211.

Tang, J., Musolesi, M., Mascolo, C., & Latora, V. (2009). Temporal distance metrics for social network analysis. In *Proceedings of the 2nd ACM workshop on Online social networks* (pp. 31-36).

Thiriot, S. (2018). Word-of-mouth dynamics with information seeking: Information is not (only) epidemics. *Physica A: Statistical Mechanics and its Applications*, 492, 418-430.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*.

Toner K., Leary M.R., Asher M.W., Jongman-Sereno K.P. (2013). Feeling superior is a bipartisan issue: extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, 24(12):2454-2462.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207-232.

Van Bavel, J.J., ... Vlasceanu, M., ..., Boggio, P.S. (2020) National identity predicts public health support during a global pandemic. Retrieved from: https://doi.org/10.31234/osf.io/ydt95

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

van Oostendorp, H. (1996). Updating situation models derived from newspaper articles. *Medienpsychologie*, 8, 21–33.

van Oostendorp, H., & Bonebakker, C. (1999). Difficulties in updating mental representations during reading news reports. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading.* 319–339

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

Vlasceanu, M., & Coman, A. (2018). Mnemonic accessibility affects statement believability: The effect of listening to others selectively practicing beliefs. *Cognition*, 180, 238-245.

Vlasceanu, M., Enz, K., & Coman, A. (2018). Cognition in a social context: a social-interactionist approach to emergent phenomena. *Current Directions in Psychological Science*, 27(5), 369-377.

Vlasceanu, M., Morais, M.J., Duker, A., & Coman, A. (2020). The Synchronization of Collective Beliefs: From Dyadic Interactions to Network Convergence. *Journal of Experimental Psychology: Applied.* Advance online publication. http://dx.doi.org/10.1037/xap0000265

Vlasceanu, M., Goebel, J., Coman, A. (2020). The Emotion-Induced Belief Amplification Effect. *Proceedings of the Annual Meeting of the Cognitive Science Society*

Vlasceanu, M., & Coman, A. (2020). The Effects of Dyadic Conversations on Coronavirus-Related Belief Change. https://doi.org/10.31234/osf.io/9zyk2

Vlasceanu, M., & Coman, A. (2020). Information Sources Differentially Trigger Coronavirus-Related Belief Change. https://doi.org/10.31234/osf.io/5xkst

Vlasceanu, M., & Coman, A. (2020). The Impact of Social Norms on Belief Update. https://doi.org/10.31234/osf.io/gsem6

Vlasceanu, M., & Coman, A. (2020). Network Structure Impacts the Synchronization of Collective Beliefs. https://doi.org/10.31234/osf.io/7rq4g

Vlasceanu, M., Morais, M.J., & Coman, A. (2021). The Effect of Prediction Error on Belief Update Across the Political Spectrum. *Psychological Science*

Vogel, D. R., Dickson, G. W., & Lehman, J. A. (1986). Persuasion and the role of visual presentation support: The UM/3M study. Minneapolis: Management

Information Systems Research Center, School of Management, University of Minnesota.

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

Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant—and memory helps to keep it that way!. *Review of General Psychology*, 7(2), 203-210.

Watts, D. J. (2004). *Six degrees: The science of a connected age*. WW Norton & Company.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440.

Wegner, D. M., Wenzlaff, R., Kerker, R. M., & Beattie, A. E. (1981). Incrimination through innuendo: Can media questions become public answers?. *Journal of Personality and Social Psychology*, 40(5), 822.

Wee, S. (2013). Development and Initial Validation of the Willingness to Compromise Scale. *Journal of Career Assessment*, 21(4), 487-501.

Wheeler, C., Green, M. C., & Brock, T. C. (1999). Fictional narratives change beliefs: Replications of Prentice, Gerrig, and Bailis (1997) with mixed corroboration. *Psychonomic Bulletin & Review*, 6(1), 136-141.

White, K. R., Kinney, D., Danek, R. H., Smith, B., & Harben, C. (2020). The Resistance to Change-Beliefs Scale: Validation of a New Measure of Conservative Ideology. *Personality and Social Psychology Bulletin*, 46(1), 20-35.

Wilkes-Gibbs, D., & Clark, H. H. Coordinating beliefs in conversation. *Journal of memory and language*, 31(2), 183-194. (1992).

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

Wilkes, A. L., & Reynolds, D. J. (1999). On certain limitations accompanying readers' interpretations of corrections in episodic text. *The Quarterly Journal of Experimental Psychology*, 52A, 165–183.

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

World Economic Forum (2013). Outlook on the Global Agenda. Geneva: World Economic Forum

World Health Organization (2020). Coronavirus disease 2019 (COVID-19): Situation Report – 59. 19 March 2020.

Zhang, J., Le, G., Larochelle, D., Pasick, R., Sawaya, G. F., Sarkar, U., & Centola, D. (2019). Facts or stories? How to use social media for cervical cancer prevention: A multi-method study of the effects of sender type and content type on increased message sharing. *Preventive medicine*, 126, 105751.

Zhong, B. L., Luo, W., Li, H. M., Zhang, Q. Q., Liu, X. G., Li, W. T., & Li, Y. (2020). Knowledge, attitudes, and practices towards COVID-19 among Chinese residents during the rapid rise period of the COVID-19 outbreak: a quick online cross-sectional survey. *International journal of biological sciences*, 16(10), 1745.

Zillmann, D., Knobloch, S., & Yu, H. S. (2001). Effects of photographs on the selective reading of news reports. *Media Psychology*, 3(4), 301-324.