UNIVERSITY OF CALIFORNIA

Los Angeles

Belief dynamics in online social networks

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

by

John Hunter Priniski

2024

Belief dynamics in online social networks

by

John Hunter Priniski
Doctor of Philosophy in Psychology
University of California, Los Angeles, 2024
Professor Keith Holyoak, Chair

Machines curate our news narratives to weaponize our minds against us, making commonsense politics impossible because we no longer share a common baseline of facts. Many helplessly accept this alienating and accelerating reality as the new normal. However, technology guided by humanistic principles and cognitive science may restore our vitality. My dissertation integrates psychology experiments on individuals and networked-groups, computational cognitive modeling, large-scale analyses of social media, open-source software development, and citizen science principles to illustrate how online networks and generative artificial intelligence can promote healthier attitudes towards established evidence and each other.

In *Chapter 2*, I analyze networked behavior and narrative agency across a series of online social network experiments (total $N = 660$, $13,200$ interactions). Through a hashtag generation game I manipulated communication patterns within a group to encourage shared or polarized beliefs. Entropy dynamics of generated hashtags and language data revealed that belief and behavioral coherence vary according to neighborhood topology at both local and global levels. Rewards for aligning hashtags also shifted participants' causal language use when writing personal narratives about a disaster event. Given these findings, *Chapter 3* introduces a computational framework rooted in Bayesian decision theory to disentangle how rewards (e.g., engagement metrics embedded in social media) *ought* to influence our beliefs about evidence, and describes how the framework can guide interventions on networked-groups.

For better and worse, people's beliefs are sensitive to online interactions. *Chapter 4* integrates large-scale analyses of internet discourse ($100,000$ interactions), belief-updating experiments (total $N = 2,676$), and interactive data visualizations to (1) identify features of persuasion in naturalistic online interactions, and (2) extend those features into a crowdsourced *data narrative* that countered misconceptions about structural racism in a random sample of Americans (Cohen's $d = .4$). *Chapter 5* describes a state-of-the-art Large Language Model system that models causal beliefs from natural language data, which I applied in previous chapters.

The first and final chapters expand on my vision for the future of cognitive and behavioral science, and sketch a trajectory for human-oriented technology development and academic achievement. As practicing scientists we must systemically reorient our goals if we are to conquer the networks that radicalize and atomize society.

The dissertation of John Hunter Priniski is approved.

Idan Blank

Hongjing Lu

Fred Morstatter

Keith J. Holyoak, Committee Chair

University of California, Los Angeles

2024

For my many loved ones, who make life worth living.

Contents

List of Figures

List of Tables

Preface

My dissertation consists of a hodgepodge of experiments, computational models, data science analyses, and descriptions of open source software. They work together to characterize my thinking about human psychology in online spaces, and how computations in human minds and machines interface to shape our individual and collective agency. The first and final chapters orient readers towards the motivation behind the internal four chapters, which describe the core of my cognitive science research throughout graduate school. These chapters are based on papers published (or to be published) in conferences and journals.

The second chapter was done in collaboration with USC's computer science department, and will be published in this summer's *Proceedings of the Cognitive Science Society*. The third chapter was written with Prachi Solanki (Michigan State University) and Zach Horne (University of Edinburgh), who both did a great deal of the intellectual lifting and writing for this chapter. The paper is currently under review at *Psychological Review*. You can consult my CV for the list of collaborators on that project. The fourth chapter is based on my earliest data science work with Zach Horne, and is a synthesis of two articles published by the Cognitive Science Society. Chapter 5 was presented at the Association for Computational Linguistics as a software demonstration. The first and final chapters are inspired by projects with Mason, Ringo, Nick, and others through Collective Thinking.

I have so many loved ones and friends who have helped me through this half-decade of triumphs and tribulations, insanity, and creativity. Since my first breath, my mom, dad, sisters, brothers, and grandparents have done nothing but love and support me, giving me the confidence to move to California and develop my scientific practice with those I idolized as an undergraduate. My grandmother and mom sacrificed everything for me to have this wonderful life and my grandfather nurtured my creativity in science, engineering, technology, and math. Zach Horne, my first serious mentor and close friend, nurtured my research with a loving brutality that drove me to reach my full potential no matter the costs. I would not have written such a document if it weren't for them three, so you can thank them or blame them.

My closest friends in Los Angeles—Amalia, Mason, Raihyung, Nick, Sara, Katie, Cash, Mary, Ringo, Jason, Carolyn, Fred (and many others)—carried me through this tormenting and tantalizing city. Submitting

this dissertation is bittersweet mainly because I will no longer be around all of you all the time anymore but I look forward to whatever new forms of relationship we may share. I love you all forever and always.

John Hunter Priniski

Education

University of California, Los Angeles, CA, Masters of Arts, Cognitive Psychology, 2020

Arizona State University, Tempe, AZ, Bachelors of Science, Mathematics, 2017, *summa cum laude*

Publications

Priniski, J.H., Linford, B., Krishna, S., Morstatter, F., Brantingham, P.J., & Lu, H. (2024). Social network topology shapes narrative processing and hashtag production. *Proceedings of the 46th Annual Meeting of the Cognitive Science Society.*

Priniski, J.H., Verma, I., & Morstatter, F. (2023). Pipeline for modeling causal beliefs from natural language. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) (pp. 436-443).*

Priniski, J. H., & Holyoak, K.J. (2022). A Darkening Spring: How Preexisting Distrust Shaped COVID-19 Skepticism. *PLOS One, 17*(1).

Adams, C., Bozhidarova, M., Chen, J., Gao, A., Liu, Z., **Priniski, J.H.**, Lin, J., Sonthalia, R., Bertozzi, A. L. & Brantingham, P. J. (2022). Knowledge Graphs of the QAnon Twitter Network. Proceedings of the *2022 IEEE International Conference on Big Data.*

Harandizadeh, B., **Priniski, J.H**., Morstatter, F. (2022). Keyword Assisted Embedded Topic Modeling. *Proceedings of the 15th ACM International Conference on Web Search and Data Mining.*

Wong, E., **Priniski, J.H.**, & Holyoak, K.J. (2022). Cognitive and emotional impact of politically-polarized internet memes about climate change. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*

Priniski, J.H., McClay, M., & Holyoak, K.J. (2021). Rise of QAnon: A Mental Model of Good and Evil Stews in an Echochamber. In T. Fitch, C. Lamm, H. Leder, & Te ßmar-Raible (Eds.) *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Lee, J., **Priniski, J.H**., Valderrama, S., & Holyoak, K.J. (2021). Disgraced professionals: Revelation of immorality decreases evaluations of professional competence and accomplishment. In T. Fitch, C. Lamm, H. Leder, & Te ßmar-Raible (Eds.) *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Priniski, J.H., Mokhberian, N., Harandizadeh, B., Morstatter, F., Lerman, K., Lu., H., Brantingham, J.B. (2021). Mapping Moral Valence of Tweets Following the Killing of George Floyd. In K. McKeown & T. Abdelzaher (Eds.), *Proceedings of the 6th International Workshop on Social Sensing.* AAAI. Menlo Park, CA: USA.

Priniski, J.H. & Holyoak, K.J. (2020). Crowdsourcing to Analyze Belief Systems Underlying Social Issues. In S. Denison, M. Mack, Y. Xu, & B.C. Armstrong (Eds.), *Proceedings of the 42th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Priniski, J.H. & Horne, Z. (2019). Crowdsourcing effective educational interventions. In A.K. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Priniski, J.H. & Horne, Z. (2018). Attitude Change on Reddit's Change My View. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2276-2281). Austin, TX: Cognitive Science Society.

Kwon, K.H., **Priniski, J.H.**, & Chanda, M. (2018). Disentangling User Samples: A Supervised Machine Learning Approach to Proxy-population Mismatch in Twitter Research. *Communication Methods and Measures*, *12*(2), 216-237.

Kwon, K.H., **Priniski, J. H.**, Sarkar, S., Shakarian, J. & Shakarian, P. (2017). Crisis and Collective Problem Solving in Dark Web: An Exploration of a Black Hat Forum. In A. Gruzd, J. Jacobson, & P. Mai (Eds.), *Proceedings of the 8th International Conference on Social Media & Society*. New York City, NY: Association for Computing Machinery.

Under Review

Priniski, J.H., Solanki, P., & Horne, Z. (*under review*). A computational framework for distinguishing motivated reasoning and practical rationality. *Psychological Review.*

Presentations

Priniski, J.H. (2023, Dec). *Behavioral dynamics and coherence shifts in a hashtag coordination game.* UCLA Department of Psychology Cognitive Forum, California, USA.

Priniski, J.H. (2023, July). *Causal language as the key to overcoming misinformation.* Causal Cognition Labratory, University of College London, U.K.

Priniski, J.H. (2022, July). *Modeling causal relations in natural language: why do it, why it's difficult, paths forward.* 1st Edinburgh Computational Cognitive Science Workshop, University of Edinburgh, U.K.

Horne, Z. & **Priniski, J.H.** (2021, Dec). *Correcting misconceptions in the lab and wild. DeepMind*, London, U.K. (Remote).

Priniski, J.H.. (2021, June). *Mapping Moral Valence of Tweets Following the Killing of George Floyd.* 6th International Workshop on Social Sensing (Remote).

Priniski, J.H. (2021, January). *Guiding Intervention Development with Naturalistic Data.* Guest lecture in *Naturalistic data in psychology.* Department of Psychology and Language Sciences, University of Edinburgh, Scotland (Remote).

Priniski, J.H. (2020, July). *Crowdsourcing to Analyze Belief Systems Underlying Social Issues.* 42nd Annual Conference of the Cognitive Science Society, Toronto, CA.

Awards

Edwin W. Pauley Fellowship, UCLA, 2022 – 2023
Dissertation Year Fellowship, UCLA, 2023 – 2024

Teaching

Teacher Assistant
Thinking (Graduate), Department of Psychology, UCLA, Spring 2023
Cognitive Psychology Laboratory, Department of Psychology, UCLA, Winter 2021
Signal Detection Theory, Department of Psychology, UCLA, Fall 2020

Research Mentorship
Chaewon Bak, UCLA, 2020-2021. *After graduation:* PhD student in Health Informatics at UIUC.
Kedar Garzon Gupta, UCLA, 2021-2022. PhD student in Neuroscience at Columbia.
Sonya Kapur, UCLA, 2022-2024. Consultant at Deloitte.
Chloe Ji, UCLA, 2023-2024. MS student in technology management at CMU.
Kaycee Stiemke, UCLA, 2022-2024. PhD student in Computer Science at CMU.

Chapter 1

Introduction

Digital media and online environments influence us through an immensely complex set of processes, with mixed effects on our beliefs, politics, and perception of ourselves and others. On the one hand, online social networks can center historically silenced voices and help political demonstrations transcend geographic space. Social media websites including Twitter, Instagram, and TikTok have enabled some of the largest demonstrations in human history and have advanced our shared understanding of racial, religious, and gender-based violence, class struggle, and the forces of imperial oppression. However, these same technologies nurture radical extremism, conspiracy theories, and identity politics on both sides of the political aisle, which thwarts our ability to tackle eminent threats facing us all. How can a single set of psychological mechanisms, social factors, and algorithms produce both sets of phenomena? How may developments in artificial intelligence, cognitive modeling, and behavioral experimentation encourage transformative, rather than detrimental, attitudes and actions towards science, democracy, and oppressed individuals? My dissertation integrates behavioral experiments, computational cognitive modeling, data science analyses, open-source software development, and principles from citizen science to reconcile the paradoxical impacts of online environments on our lives, and to chart a path for developing human-oriented technologies that empower all people.

On the need to reconsider reasoning norms when interpreting online phenomena

The predominant view among behavioral scientists, particularly social psychologists, characterizes the emergence of political, moral, and scientific misconceptions as the result of biased or lazy thinking. Popular

intervention strategies aim to optimize how people integrate evidence into their beliefs (Nyhan & Reifler, 2010; Nyhan et al., 2014a,b; Nyhan & Reifler, 2015) or to limit the uptake of inaccurate information (Pennycook & Rand, 2019, 2020). For example, *accuracy nudges* assume people lazily interpret the information they read on social media and do not exert the cognitive effort required to assess the veracity of a post, leaving them susceptible to updating their opinions on inaccurate information (Pennycook & Rand, 2019, 2020; Pennycook et al., 2020; Pennycook & Rand, 2021; Pennycook et al., 2021; Pennycook, 2022). Accuracy nudges encourage people to engage more critically with potentially false information by, quite literally, informing them to "consider the accuracy of the information" they are reading. Accuracy nudges are widely studied because they can be easily embedded in real-world social media environments (Pennycook et al., 2021), exhibit statistically significant (albeit small) effects on a variety of hotly debated topics, and can influence people's decisions without limiting their agency (Thaler & Sunstein, 2009). However, it is worth considering that direct replications of specific accuracy nudges have failed to measure significant effects (e.g., reducing COVID-19 misinformation; Roozenbeek et al., 2021), and a recent meta-analysis suggests that behavioral nudges do not produce reliably significant effects when controlling for publication bias (Maier et al., 2022). Furthermore, the cognitive mechanisms posited to underlay their efficacy are poorly specified (at least against the standard of a computational mechanism), which makes it more difficult to predict their effects on untested topics and to systematically build upon established findings.

To be fair, most intervention strategies do not articulate the computational cognitive mechanisms they target to revise beliefs, although there are exceptions (D. Powell et al., 2022; Roozenbeek et al., 2021, 2022; Lewandowsky et al., 2019). Even interventions driven by computational modeling rarely specify what constitutes rational belief updating when both evidence and utility information (e.g., social rewards) influence people's beliefs and decisions. This limitation has concrete implications for how we characterize and correct the onset of misconceptions and related behaviors, particularly those emerging in online environments where evidence and social information are radically curated by news feed algorithms and networked interactions. A reasoner may rationally evaluate the expected utility of decisions when interpreting relevant information (e.g., sharing a proactive post that will get increased engagement), but if the expected utility of a decision outweighs the predicted veracity of information suggesting another, lower-utility action, an individual could act in ways that is locally rational (in that the selected decision satisfies expected utility constraints) but is simultaneously incompatible with the best available evidence and is thus globally irrational. In such cases, is the reasoner exhibiting irrationality and motivated reasoning or are they following optimal decision-strategies? Furthermore, if a decision requires changing (e.g., reducing red meat consumption) an effective intervention strategy may need to target a reasoner's utility-calculus to shift behaviors, not realign mental representations to optimally cohere with causal and probabilistic information in the environment. I unpack

this idea in excruciating detail in the third chapter of my dissertation by establishing a computational framework and simulation results to discern how reward information (e.g., social media metrics) ought to influence our thinking about causal evidence. I return to this idea in the final chapter by discussing how this framework can extend popular intervention strategies in the academic literature to incorporate social reward information widespread in online networks.

Bridging behavioral experiments and natural language processing to determine psychological mechanisms shaping beliefs in online networks

While tightly-controlled experiments carried out on individuals in the laboratory can uncover psychological mechanisms linking information processing to decision-making, the dynamic nature of online environments challenges generalization of experimental findings outside of the lab (J. H. Priniski & Horne, 2018, 2019; Goldstone & Lupyan, 2016). For example, hashtags can frame text information embedded in social media to signal group support, as was evident on Twitter following the murder of George Floyd. Liberals were more likely to view tweets containing #AllLivesMatter as racist towards black people, while conservatives were more likely to view tweets containing #BlackLivesMatter as racist towards white people (M. Powell et al., 2023). The computational framework introduced in the third chapter establishes how social information ought to impact people's behaviors and representation of evidence, and can be applied to understand how beliefs can shift overtime to cohere with behavioral outputs (for example, how coordinating on a hashtag shapes perception of information it tags or how increased utilities for certain behaviors could increase people's credence in evidence supporting those behaviors). However, psychology experiments and measures are necessary to relate the framework's computational predictions to human reasoning and behavior in the laboratory and in real-world online environments.

Beliefs are internal representations that need to be inferred from behavioral outputs, including decisions and use of language. As language data is a lens onto people's representation and beliefs about causal information, researchers have gathered and analyzed language data at scale from naturally-occurring online interactions to understand how psychological processes operate outside of tightly-controlled laboratory contexts. Over the last decade, psychologists have used natural language processing algorithms to automate the analysis and prediction of online behavior embedded massive repositories of *trace data*: data that signals online behavior, including the composition of a Reddit post, sharing a TikTok, or posting an image on Instagram. While a single piece of social media trace data says very little about an individual's psychology,

many millions of pieces of trace data in the aggregate can illuminate both individual level and group level dynamics. However, while natural language analyses of social media are applicable to realistic situations and can extend empirical findings outside of the laboratory, they generally yield correlational evidence that does not isolate causal mechanisms. This can make it difficult to determine the causal factors dictating the spread and uptake of online information over other factors. For example, by correlating a message's predicted morality labels with its engagement metrics, a few high-impact papers argued that morality frames the spread of online information, as predicted by psychological theories of moral reasoning and emotion (Brady et al., 2017, 2020, 2023). These papers claim that moralized content spreads farther than other content because moral cues trigger emotive responses that drive individuals to engage. However, once a message's valence is controlled for, the effect of morality was reduced to essentially zero (Burton et al., 2021). Difficulties in replicating key naturalistic findings in the cognitive and social psychology literature reinforces the necessity of using experiments to determine causal factors. While natural language processing algorithms are useful tools for increasing the naturalism of theories and findings, they are no replacement for traditional psychology experiments guided by computational theories of cognition (Guest & Martin, 2021). Natural language processing systems should isolate representations and be used in tandem with experiments and large-scale analysis of naturalistic behavior to align cognitive mechanisms across domains.

In chapter two, I run a series of experiments with groups of individuals to measure how coordination with networked neighbors impacts participants representation of evidence, and show how coordination utilities have measurable impacts on a group's behaviors and people's representation of narrative-based information. Specifically, in chapter two, I apply a state-of-the-art natural language processing model I developed to measure representational change induced by coordinated behavior in the social network experiments. I also apply this pipeline to demonstrate how causal information sourced from these discussions can be used to correct misconceptions about African Americans in a random sample. I describe the underpinnings of this model in chapter five, and demonstrates how the pipeline can model belief systems surrounding social issues by sourcing causal beliefs about the Covid-19 vaccine from tweets.

Integrating theoretical and methodological advances with experiments to revise misconceptions

The many misfires in the academic literature, coupled with the widely-observed phenomena of trolling, fake news, and misinformation on social media, have lead many to conclude that there is no hope in restoring our attitudes, discourse, and politics. While reward-based interventions can potentially shift people's behaviors

by targeting the shape of their utility calculus, these strategies, by design, do not target people's representation of causal information. This of course could have downstream consequences, as people will not generalize learnt information or systematically realign their perceptions in a way that is coherent with the global set of causal and probabilistic relations in the environment. While the merits of such an intervention strategy is a point of debate, I suspect reward-based interventions will be most useful in cases *after* attempts to realign causal representations have been systematically exhausted. Behavioral experiments are the gold standard for elucidating causal effects, but are costly and can take years to complete, and ensuring an empirically tested intervention scales to online environments is close to impossible without direct access to a social media platform. Complicating matters, even interventions crafted with considerable care can have a minimal effect on targeted beliefs and behaviors in random samples.

However, even if a scientifically-grounded intervention fails to change perceptions about the causal evidence, it is still possible that other untested interventions could be effective. Experiments are not practically feasible to systematically explore and test the entire hypothesis space of possible interventions and exhaust all possibilities to realign representations of causal information. Scalable methodologies in Natural Language Processing can allow researchers to prune the hypothesis space before running an experiment, and develop educational interventions based on information previously vetted in real-world online networks. Chapter 4 integrates data mining, natural language processing, and behavioral experiments to highlight the efficacy of such a methodology based on Reddit's ChangeMyView, an online forum where users post their beliefs with the expectation that others will provide responses in an attempt to change their mind.

We must establish critical approaches to developing AI and online networks

I conclude my dissertation by discussing how these findings and methodologies can be integrated to advance AI and network technology in a more ethical manner. I discuss how practicing scientists, engineers, and concerned citizens can work together to implement simple community-driven solutions that limit the harms scientific and technological development has on historically exploited communities and identities. In a broad sense, the current state of these technologies and algorithms depend on our isolation and antagonism to sustain themselves. Our suffering is their flourishing. The solution is simple because it is not a solution. It is our shared agency that will restore our vitality.

Chapter 2

How online interactions shape narrative agency and shared beliefs

Introduction

The internet has remapped how people interpret and discuss events. Digital technology enables organizing efforts that transcend geographic limitations, leading to some of the largest demonstrations in history (P. Dawson, 2020; G. Yang, 2016). However, these same platforms can create information environments that foster extremism, hate, and anti-democratic ideals. Echochambers, where online communities consume media that confirms their beliefs and identities, are breeding grounds for unreliable information and conspiracy theories (Sasahara et al., 2021a). Empirical research is necessary to understand how networked environments shape belief formation at both individual and group levels, so as to better control the dynamics of information spread, and to possibly mitigate the harm of misinformation and social segregation.

Empirically investigating networked behavior is a challenging task because the narratives arising in online contexts are sprawling and unwieldy (Tangherlini et al., 2020), largely due to the dynamic and complex nature of modern social media environments. For instance, the interactivity of digital media and online social networks allows for a new sense of *narrative agency* (G. Yang, 2016). Through the production and sharing of social media data, users can embed personal narratives and express their point of view on an event or social issue (Boyd et al., 2010). From these low-level interactions, a collective narrative emerges within a group of individuals, who bring their own prior beliefs and goals when characterizing an event or issue, which directly shapes the narrative shared by a network (P. Dawson, 2020).

Hashtags are a potent force for narrative interaction on social media; they allow users to tag personal

narratives and contribute to online discourse by indexing their produced content with proxy topic labels. Previous research has shown that hashtags are concise representations of the narratives (Giaxoglou, 2018; P. Dawson, 2020), and connect spatially disorganized groups according to the content of their narratives and goals (Papacharissi, 2015, 2016; Howard & Hussain, 2013). Across an online network, hashtags categorize social media discourse for effective indexing and search, and can allow interpersonal signaling and sense making between instances (Papacharissi, 2015, 2016).

Previous research on hashtags has primarily focused on how they are used in real-world (i.e., "scale-free") networks, by focusing on the linguistic and semantic structure of popular hashtags (Booten, 2016), or modeling the dynamics underlying their spread online (Cunha et al., 2011; Lin et al., 2013). For instance, hashtags fall into one of two categories. *Focal hashtags* tag posts with broad semantic topics to relate posts to broader discussions and movements across an onine network. A second set of *individualistic hashtags* make the distribution of hashtags heavy-tailed, as they co-occur with focal hashtags while allowing users use to signal personal narratives (Booten, 2016). Furthermore, hashtags generally fall into a "winner' or "loser" category (Lin et al., 2013), in that while many hashtags initially compete for popularity, only a small set of hashtags will persist to allow for broader narrative collaboration across the network. It is unclear how endorsed narratives shape the production of hashtags and how hashtag dynamics are sensitive to network structures. To address these important questions, empirical experiments must mirror the interaction structure and media of online environments to effectively account for how groups engage with narratives in real-world networks, and how group behaviors are affected by an individual's representation of the underlying event.

One approach to studying narrative interaction is to run social network experiments, where a group of participants are placed in a social network and interact with one another. This paradigm allows network structure to be manipulated under experimental control. Social network experiments have historically used relatively simple materials to measure how varying social network structure (e.g., node connectivity) influences the adoption of normative behaviors. For example, Centola and Baronchelli (2015) asked participants to coordinate on naming an image of a face with network neighbors. They found that interactions within *homogeneously-mixed networks*—fully-connected networks where each participant is linked to every other participant—support the emergence of normative behaviors (i.e., the full network aligning on a single name), while *spatially-embedded networks*, where each participant is linked only to a handful of neighbors, did not. Here, we extend this well-established network experiment design by using naturalistic narrative materials and interaction behaviors akin to those in real-world online networks. We hypothesize that repeated interactions in localized neighborhoods will allow groups to coordinate more effectively, but neighborhoods spanning a fully-connected network will be more likely to produce dominant behaviors. To test these hy-

potheses, we developed fine-grained measures of behavioral coherence at both the level of local coordination between pairs of participants, and at the level of global convergence across the full network. In addition, we applied an NLP model to compare people's causal language use in personal narratives before and after networked interaction.

Network Experiment 1: Neighborhood topology shapes personal narrative and hashtag dynamics

Participants

We sampled a total of $N = 420$ participants from the Prolific and SONA subject pools, and placed them into one of ten network conditions. Conditions vary the size of a network ($N = 20, 50, 100$) as well as its connectivity (homogeneously-mixed/fully-connected; spatially-embedded/ring-like). We collected six network runs for $N = 20$ (three runs per network structure), and single runs for each structure of $N = 50$ and $N = 100$. Participants $N = 20$ and $N = 50$ conditions were sampled using Prolific. For the $N = 100$ condition, we recruited undergraduates in the Department of Psychology at UCLA through SONA . We posted initial recruitment surveys a week prior to each run in SONA and a few hours prior to each run in Prolific. Participants who received the most points at the end of the experiment received an additional $10 bonus.

Materials

Across all network conditions, participants first read a four-paragraph narrative description of the Fukushima nuclear disaster. The narrative explains how a large earthquake triggered a tsunami that caused damage to a nuclear reactor and resulted in radiation leaks, population displacement, and an energy-saving movement "Setsuden". We selected this narrative based on a pilot study demonstrating that it resulted in the most diverse set of hashtags within a set of tested narratives related to natural and financial disasters. This is likely because the narrative describes a rich set of causal relations (a generative causal chain producing a branching common cause sequence) and included both negative (e.g., displacement, poisoning) and positive effects (e.g., energy saving movement). Fig. 2.1 illustrates the causal structure of the Fukushima disaster narrative.

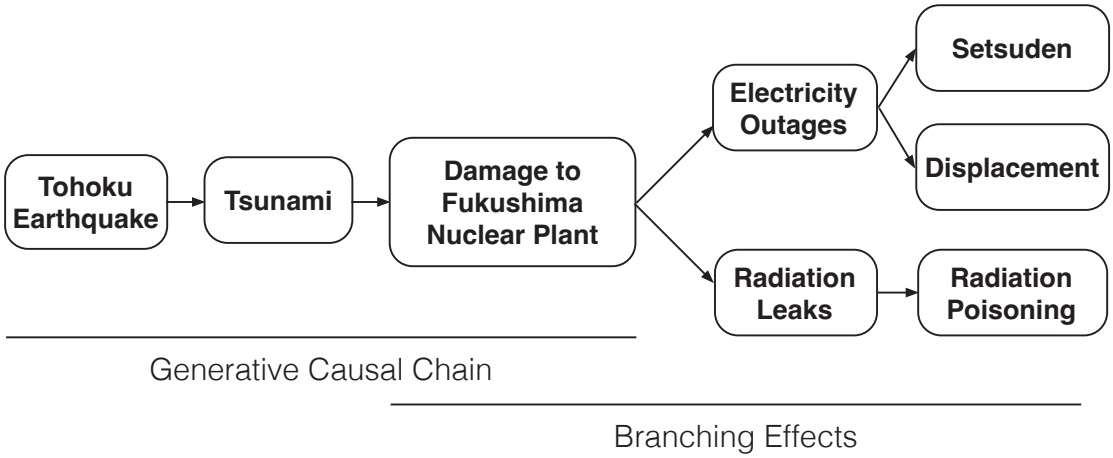

Figure 2.1: **Causal model communicated in the nuclear disaster narrative**. This diagram is just for illustration purposes, participants did not see this diagram. They read a four-paragraph narrative describing how the Tōhoku earthquake triggered a massive tidal wave that damaged the Fukushima Nuclear Power Plant, resulting in electricity outages, radiation leaks and poisoning, human displacement, and *Setsuden*, a national energy-saving holiday.

Experimental Design and Procedure

Network Experiment Method

We used the open-source framework OTree written in Python (D. L. Chen et al., 2016), and hosted experiments on a Linux server. Participants joined the experiment through a Qualtrics survey that directed participants to the network experiment.

Our social network experiment proceeded in three steps. First, we randomly assigned each participant as a player in a network that defined who may interact with whom on a given trial. Second, we assigned interactions between individual participants on each trial. Third, we rewarded participants based on the outcome of their interactions. We can specify this process using graph theory notation. The first step is to initialize a fixed graph $G(N, E)$, defined by a set N nodes representing individual participants connected through an edge set E. We discuss below the specific graph structures used. The second step iterates over T trials. On a given trial $t \in T$, connection (edge) configurations follow mixing participants randomly within a participant's neighborhood. The third step is to identify and reward coordinated behavior. If the response from participant n_i on trial t is r_i^t, then participants n_i and n_j coordinate if $r_i^t = r_j^t$.

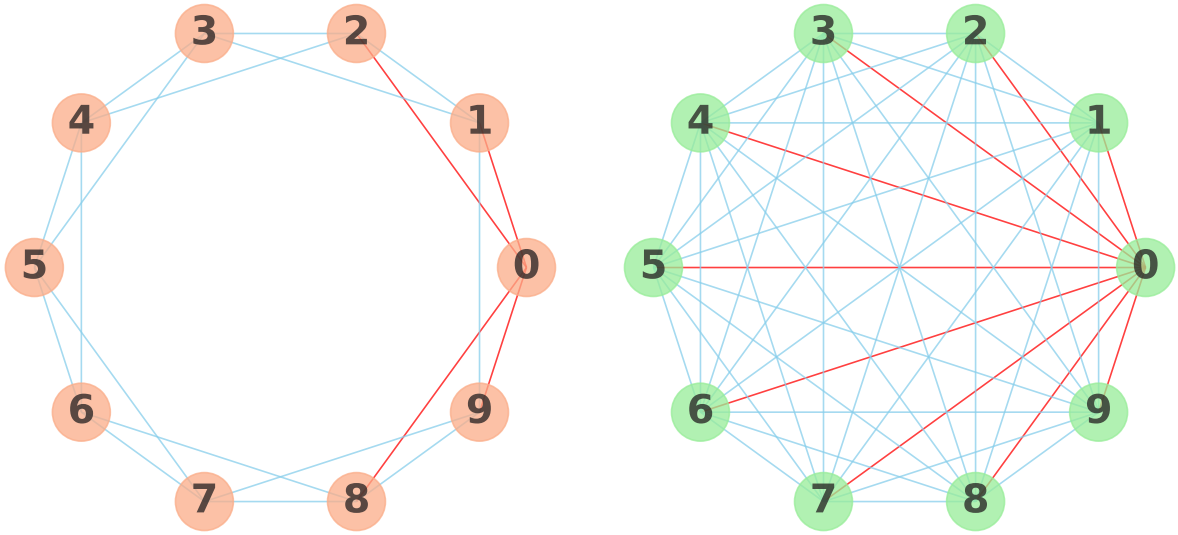

Figure 2.2: **Two neighborhood structures tested in this experiment:** Spatially-embedded (**left**) and homogeneously-mixed (**right**) networks with $N = 10$ nodes. Red edges represent the neighborhoods for a hypothetical node 0 in both networks. As a network's size grows, the diameter of spatial networks grow whereas homogeneous networks maintain a diameter of 1.

Procedure

The experiment consisted of three blocks: a pre-interaction block, a networked interaction block (described above), and a post-interaction block. This three-block design allowed us to assess behavioral dynamics during the networked interaction block, in addition to examining whether networked interaction can shift beliefs from pre- to post-interaction blocks.

In the **pre-interaction block**, participants read a four paragraph narrative describing the Fukushima nuclear disaster, and then were asked to write a "tweet" (within a 140 word limit) and ten hashtags characterizing the events described in the narrative.

In the **networked interaction block**, participants joined a network experiment through real-time interaction on the online platform. Participants were assigned to one of six experimental conditions based on the size of the network ($N = 20; 50; 100$) and network structure (spatially-embedded and homogeneously-mixed; see Fig. 2.2). Regardless of network size, nodes in spatial networks have a consistent neighborhood size $k = 4$, meaning each participant would interact with four other participants in the entire experiment. Neighborhood size in homogeneous networks is $N - 1$, as each participant can interact with any of the remaining participants. A consequence is that the network diameter (i.e., the largest geodesic distance in the connected network) was consistently 1 in all tested homogeneous networks, but grows as a function of size in spatial networks. Both of these features of network topology (i.e., size and diameter) uniquely influence

the emergence of shared behavior in online networks (Anagnostopoulos et al., 2012).

The networked interaction block consisted of 40 trials, where participants interacted with their partners based on the edge structure in the assigned network. On each trial, participants were instructed to write a single hashtag describing the narrative they read in the pre-interaction block. After participants submitted their hashtag response, they were then presented with a new page showing their own hashtag response, their partner's hashtag response, whether they received a point for matching responses with their partner, and their cumulative reward point.

Following networked interactions, participants entered a **post-interaction block** where they wrote one more "tweet" for the same narrative and another ten hashtags describing the event. Before completing the experiment, they provided demographic information.

Results and Discussion

Global connections support emergence of dominant responses

We fit a Bayesian generalized linear model (GLM) to predict how the two network structures (spatially-embedded vs homogeneously-mixed structure) support the emergence of a dominant hashtag response. We assume that the proportion of participants who produced the dominant hashtag on trial t follows a Beta distribution, a commonly used distribution to predict proportion values (McElreath, 2016); we used uninformative priors (i.e., $\mathcal{N}(0, 10)$) over regression coefficients. Specifically, the GLM model predicted the proportion value as a function of trial number (i.e., $Trial$) interacted with network structure ($Spatial$ vs $Homogeneous$), while controlling for network $size$. Here, we simply predict the proportion of players in a network producing the dominant response on a given trial, which could be different hashtags across different trials for a single run.

As shown in Fig. 2.3, shared responses emerge from networked interactions in both network structures ($\beta_{Trial} = 0.04$, 95% CI $[0.04, 0.05]$), however dominant responses emerge more easily in homogeneously-mixed networks than spatially-embedded networks ($\beta_{Trial \times Spatial} = -0.01$, 95% CI $[-0.02, -0.00]$). We found that within the confines of a given experimental run (i.e., 40 interaction trials), the shared responses emerge more quickly in smaller network sizes than larger ones ($\beta_{Size} = -0.01$, 95% CI $[-0.01, -0.00]$) (note that this effect represents the additive effect of increasing network size by one). These results suggest that network structure and size are important characteristics to determine the adoption of shared beliefs, and replicates previous findings from behavioral economics and the computational social sciences (Golub & Jackson, 2010; Centola & Baronchelli, 2015).

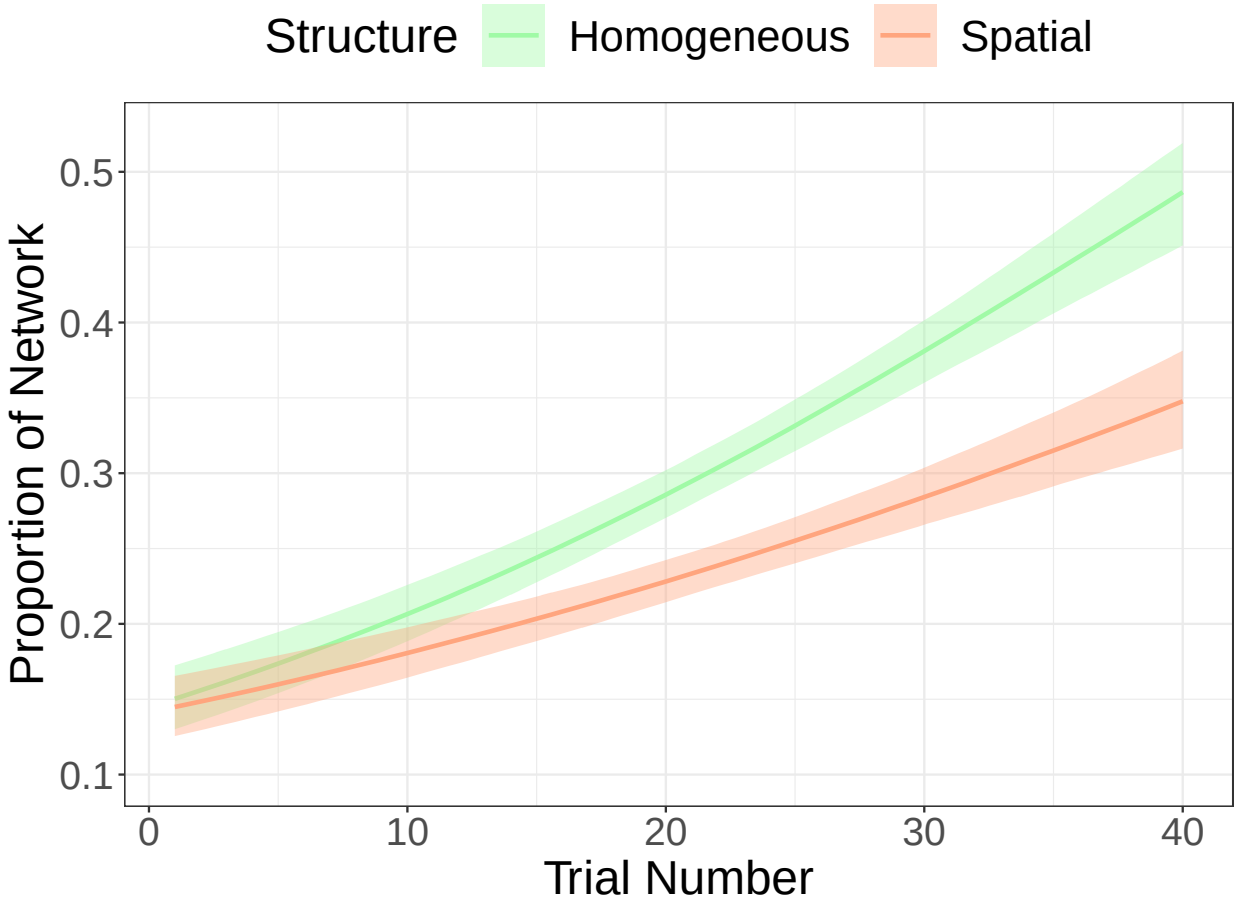

Figure 2.3: **Proportion of participants generating a dominant hashtag as a function of trials across network structures.** Curves represent marginal effects of the linear model, and error bars represent 95% Bayesian credible intervals. Homogeneously-mixed networks show faster emergence of dominant responses due to information aggregation across network connections.

Network topology affects entropy dynamics of response distribution

Responses from all players in a network produce a distribution of generated hashtags on each trial. The variability of hashtags over time captures the degree of coherence of individual responses across all participants in a network. Hence, entropy is a concise measure of response variability across a network (Avolio et al., 2019; Hallett et al., 2016); the lower the entropy, the more similar the set of responses from all players. We computed the change in entropy of the response distributions across each network run, and fit a Gaussian generalized linear model to predict a network's entropy as a function of time and structure (and their interaction), while controlling for a network's size.

As shown in Fig. 2.4, the GLM model shows that the starting entropy is similar across network structures ($\beta_0 = 2.59$, 95% CI $[2.47, 2.71]$). Furthermore, the entropy of hashtag responses decreases over time ($\beta_{Trial} = 0.04$, 95% CI $[-0.04, -0.03]$), however entropy decreases at a faster rate in homogeneously-mixed networks than spatially-embedded networks ($\beta_{Spatial} = 0.01$, 95% CI $[0.01, 0.02]$). Because larger networks have more participants contributing responses, we also found a positive effect of network size on entropy ($\beta_{Size} = 0.01$, 95% CI $[0.01, 0.01]$). While a single dominant response may struggle to emerge in spatially-embedded networks, responses are still cohering within local neighborhoods, which in turn decreases the entropy (less variability) of hashtag responses, but at a slower rate than homogeneously-mixed networks. Because separate local-neighborhood groups can align on different hashtags in spatially-embedded networks, this finding is consistent with echochambers found in online networks. We discuss this idea in more detail at the end of this section (see Fig. 2.6).

Local connections support coordination within neighborhoods

Connections across a network support the emergence of a dominant belief. However, increasing the connectivity of a node decreases the amount of times that any given pair of players can coordinate. In the spatially-embedded, partner players coordinate ten times across forty trials, regardless of total network size; whereas in homogeneously-mixed networks repeated interactions significantly decrease as a function of network size (i.e., number of players in the network). To assess how coordination dynamics varied across network structures, we fit a Bernoulli generalized linear model to predict if a pair of participants coordinated on trial t interacted with network structure over time (controlling for the size of the network).

As shown in Fig. 2.5, participants in both neighborhood structures learned to coordinate with their networked neighbors as trials progressed ($\beta_{Trial} = .04$, 95% CI [0.00, 0.04]), however those in spatially-embedded networks coordinated more effectively than they did in homogeneously-mixed networks ($\beta_{Spatial} = .02$, 95% CI [0.01, 0.03]).

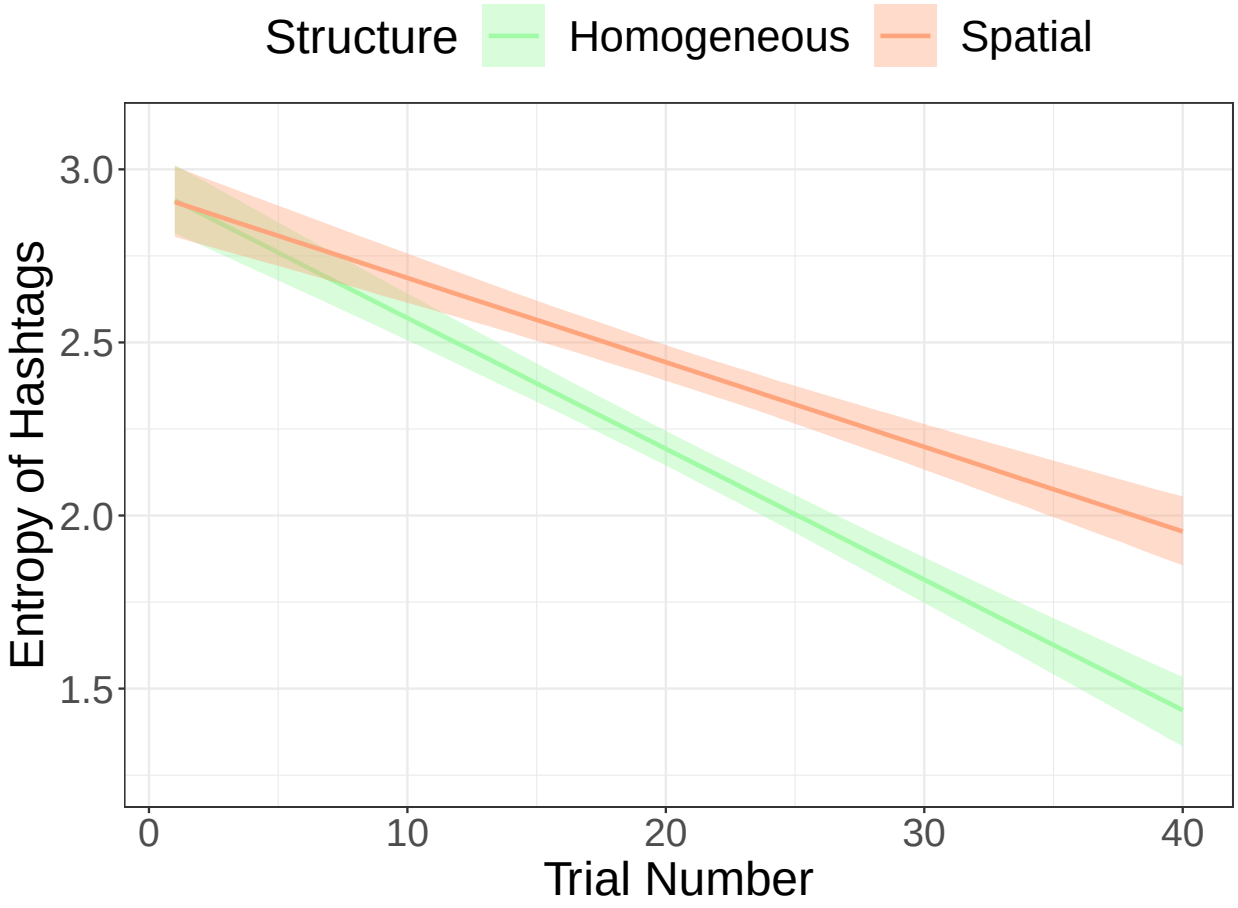

Figure 2.4: **Entropy of hashtag responses decrease over time across network structures.** Lower values imply greater coherence of responses across participants because generated hashtags are more similar. However, entropy of responses decrease more rapidly in homogeneously-mixed networks due wider communication of information.

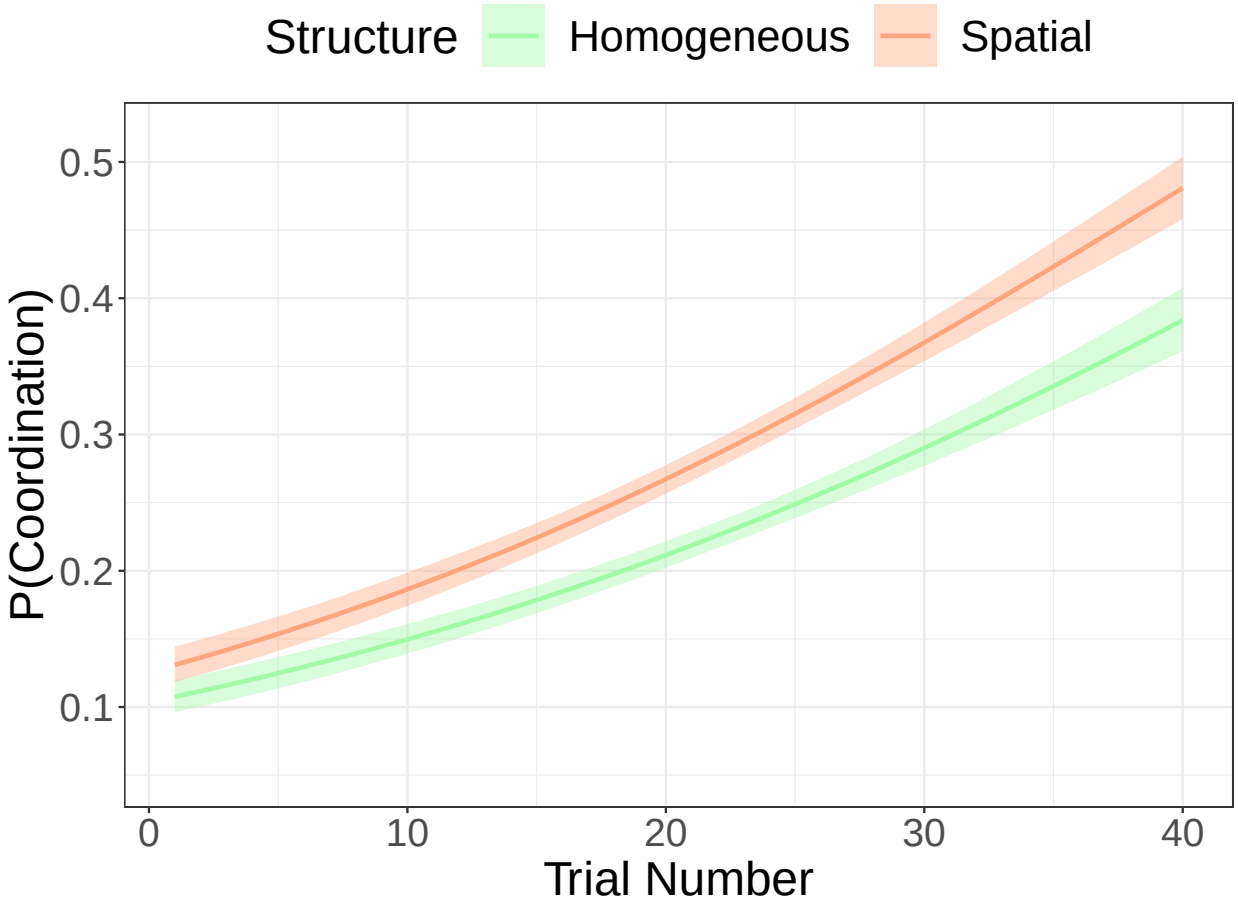

Figure 2.5: **Emergence of coordinated behavior across network structures.** Participants can coordinate more effectively with their network neighbors in spatial networks than in homogeneous networks, because smaller neighborhood sizes allow for more repeated interactions and quicker learning of one's environment.

Local coordination can result in separable neighborhoods coordinating on different hashtags. Fig. 2.6 depicts the full array of responses for a single run of both $N = 20$ network structures. Different local groups align on different hashtags in the spatially-embedded network. For example, 16-18 adopted #NuclearDisaster, while 2-6 and 8-14 aligned on #Nuclear; and nodes 19, 20, and 1 aligned on #Setsuden. The emergence of separate, localized groups coordinating on different hashtags likely hinders a dominant response from emerging in these networks, and limits the decrease of entropy as shown in the earlier analysis of global network coherence. The other $N = 20$ runs and network sizes $N = 50, 100$ display similar results. Participants in localized clusters received high rewards for coordinating within their partners in the local clusters (as indicated by the node size in the networks to the left of the color maps). Although participants were not coordinating as effectively in homogeneously-mixed networks, network topology still supported the emergence of a dominant response globally.

This result suggests that a latent form of information aggregation leads to the emergence of dominant hashtag, rather than being directly due to participants coordinating effectively in the local neighborhood (Golub & Jackson, 2010). Furthermore, note how in Fig. 2.6, participants 7 (top) and 17 (bottom) both received zero reward for coordinating, and continually generated new hashtags over course of the experiment. Participants are rewarded by adopting shared behaviors that encourage a shift towards consensus, and those that don't learn to exploit environmental regularities are not rewarded.

Emergence of social learning dynamics in networked environments

How might social learning strategies emerge in relation to the space of options to search and environment uncertainty? On a given trial, a participant could either copy what they said on the last trial (self-consistency), copy what their partner provided (external consistence), copy a reward (if they and their partner provided the same response), or not copy and generate a new hashtag (at least in relation to the previous trial). Simulation studies suggest that copying a successful individual is an effective learning strategy in noisy environments (Barkoczi & Galesic, 2016), but what social learning strategies may underlay effective coordination when coordination utilities are not observable?

To predict the number of participants adopting a given strategy on a given trial, we fit a categorical generalized linear model to predicted the counts of each strategy as a function of time, network structure, size. Models did not reveal a significant effect of structure on decision strategy sampling, so we average across all network conditions. This model revealed, as shown in Figure 2.7, that participants are far more likely to not copy themselves or others, in effect generating their own new hashtag; however this rate decreases from 80% at beginning of experiment to about 25% of population in final chapters. Of the participants choosing to copy the last trial, the majority of the time participants chose to copy them selves (15 to 40% rise). By

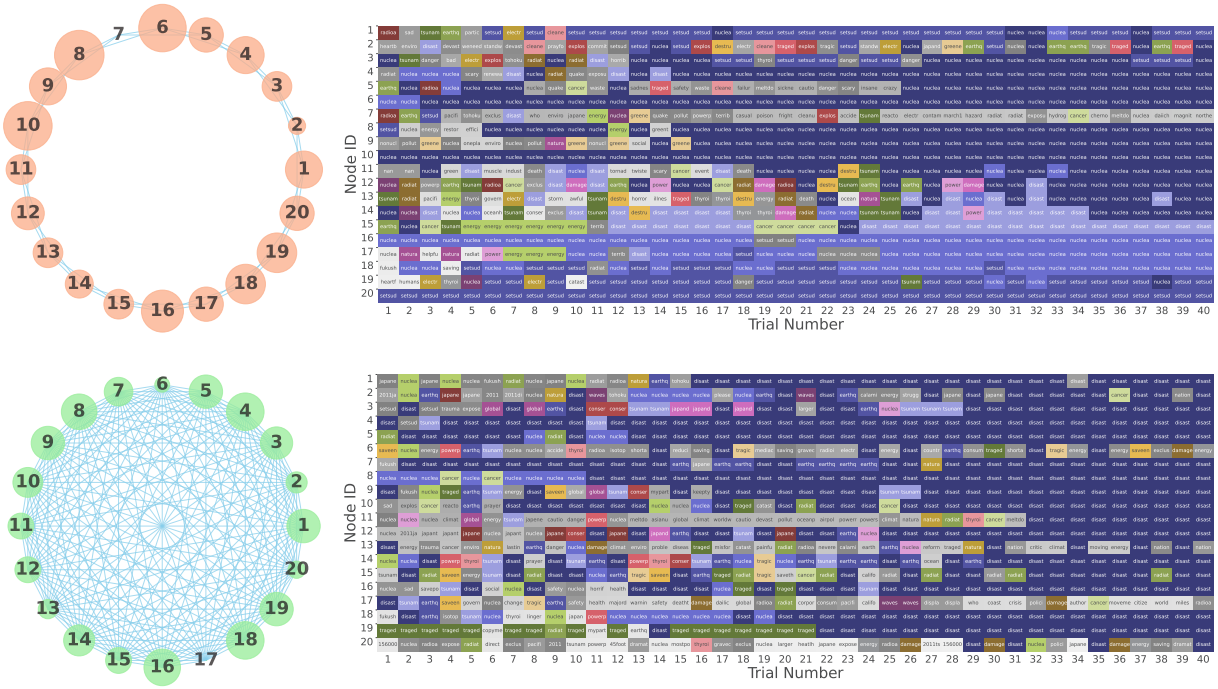

Figure 2.6: **Rewards and colormaps of hashtag responses across a single $N = 20$ run.** Top panel shows results for a spatially-embedded network, the bottom panel is from a homogeneously-mixed network. **Left:** Network structure with player nodes sized by participants' final rewards for coordinating (range (top) 0-25; (bottom) 0-19). **Right:** Colormap of individual responses, rows represent individual participants' set of responses, columns represent trials. Cells encode the first five letters of the generated hashtag.

the end of the experiment, participants are nearly re-sampling their previous hashtag as often as generating a new one.

To predict which strategies are most effective for coordinating with one's neighbors over the course of networked interaction, I fit a Bernoulli generalized linear model to predict if a pair coordinated as a function of decision type and network covariates. As shown in Figure 2.8 , I found that copying responses is crucial for coordinating. Copying a rewarded trial ($\beta_{CopyReward}$ = 2.54, 95% CI [2.20, 2.89]) and a partner's last response ($\beta_{CopyPartner}$ = 1.23, 95% CI [0.90, 1.57]) were the best strategies for coordinating compared to copying one's own response (self-consistency) ($\beta_{CopySelf}$ = .33, 95% CI [0.07, 0.60]) (these estimates are relative to not copying which is the reference condition). Additionally, each of the copy-based decisions has a positive interaction with time, which suggests that consistency pays off more often as time progresses, with the largest interaction effect for copying a partner's: $\beta_{CopySelf\times T}$ = .01, 95% CI [0.00, 0.02]); $\beta_{CopyPartner\times T}$ = .03, 95% CI [0.02, 0.05]); $\beta_{CopyReward\times T}$ = .02, 95% CI [0.00, 0.01]). These results are striking because by far the dominant strategy among populations across runs is to not copy, which is the least effective for coordinating. While copying a reward trial is the best decision to coordinate,

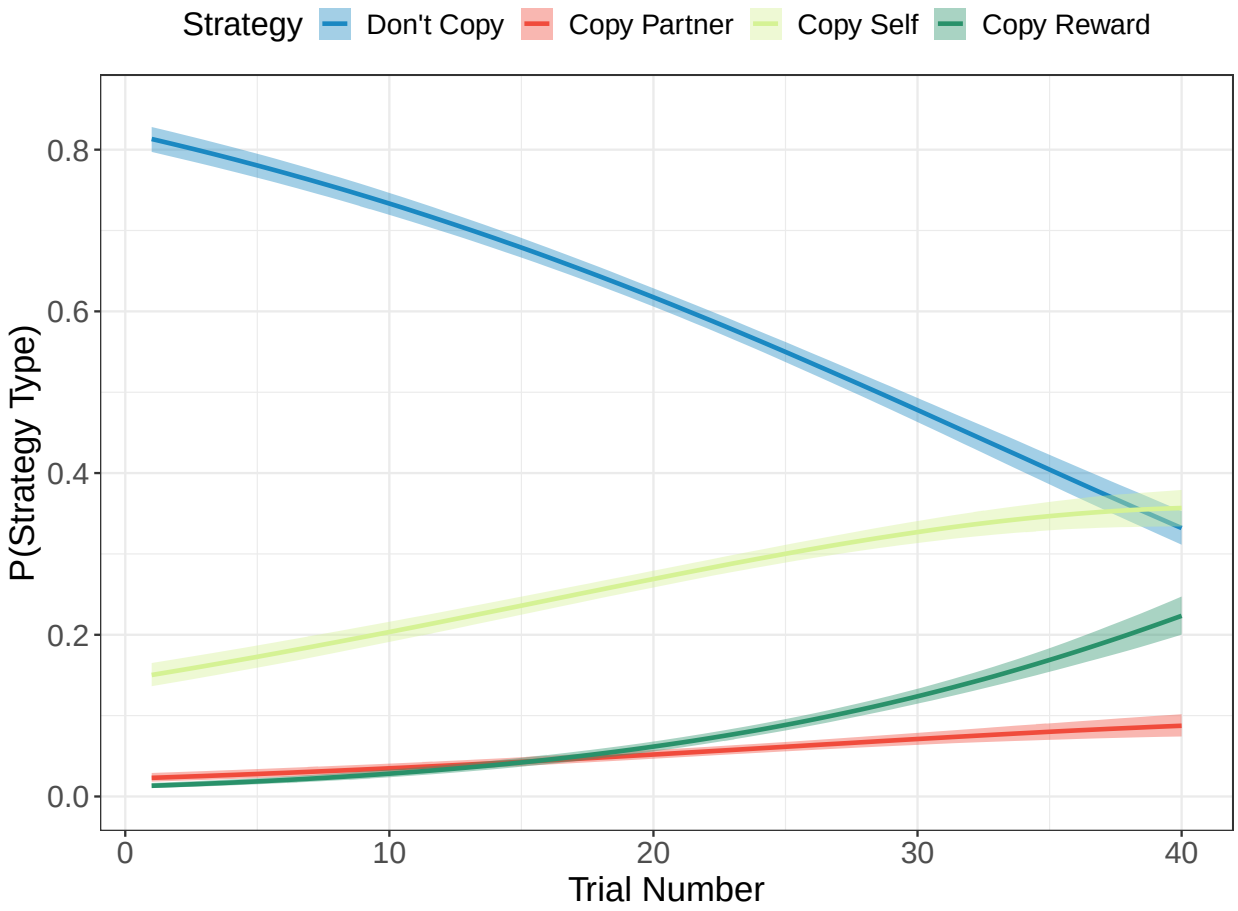

Figure 2.7: **Strategy sampling dynamics over the course of experiments.** At the beginning of the experiment, participants are most likely to generate new hashtags (i.e., explore responses), and monotonically shift towards copying themselves, rewarded trials, or their partner's responses.

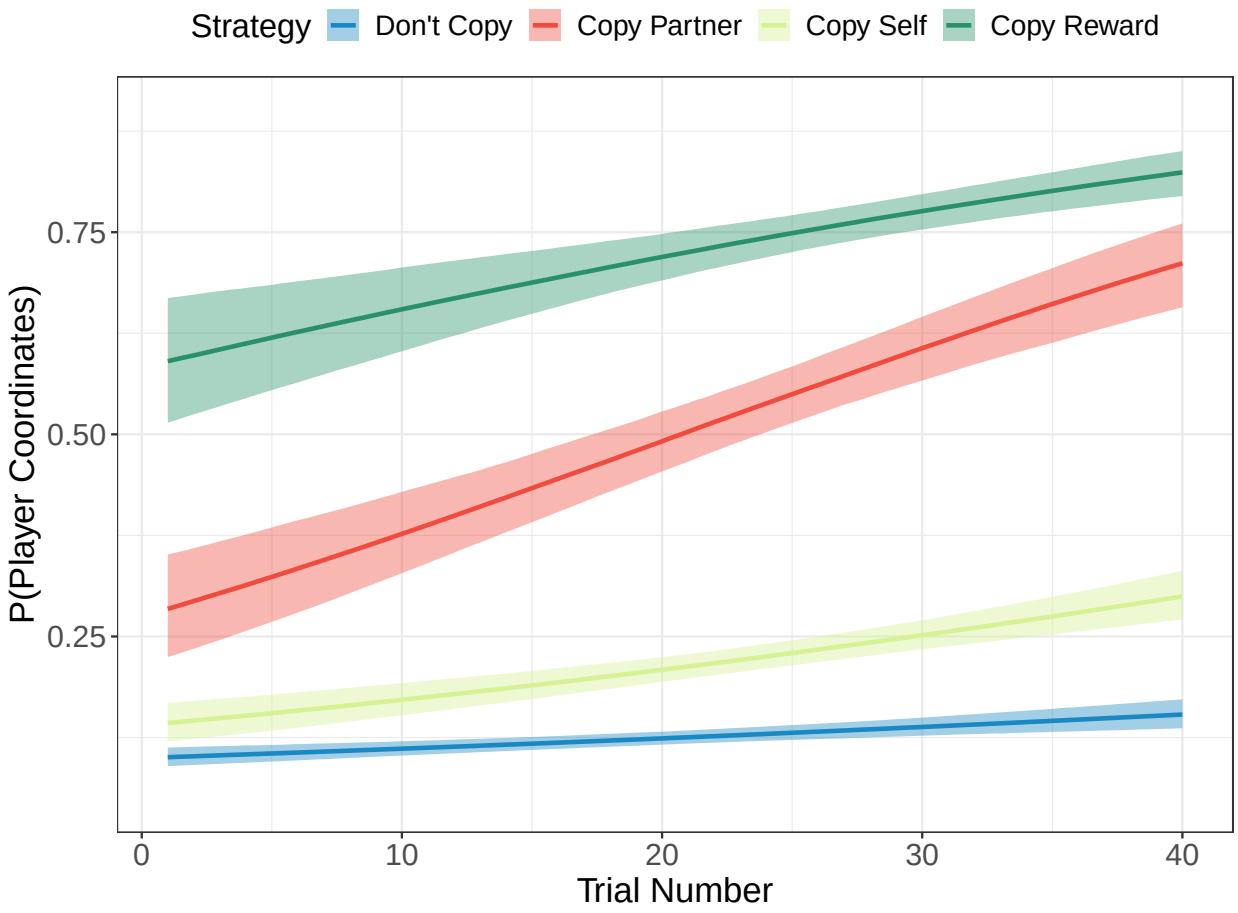

Figure 2.8: **Dynamics of effective strategies.** The probability of a player coordinating given a sampled strategy. Participants are most likely to coordinate if they copy rewarded trials or their partner's last response (i.e., exploiting environmental regularities), than if they copy their own response (i.e., self-consistency).

fewer instances exist at the beginning of the experiment, and emerge later despite a healthy amount of individual differences in strategies across a group (see Figure 2.9). Therefore, repeating a partner's response is a reasonable strategy to adopt earlier on, and to switch to rewarded responses after more environmental consistency emerges within the networked group. In future work, computational models can help distinguish when switching strategies is optimal given environmental regularities, and can be applied to predict group level outcomes as a function of the variance on individual-level strategies composing a group.

Probing causal representation change following networked interactions

Before and after networked interaction, participants wrote "tweets" describing the narrative. Natural language processing (NLP) methods applied to these documents can illuminate which parts of the narrative people focused on when describing the events in a brief format such as tweets. Because causal relations are

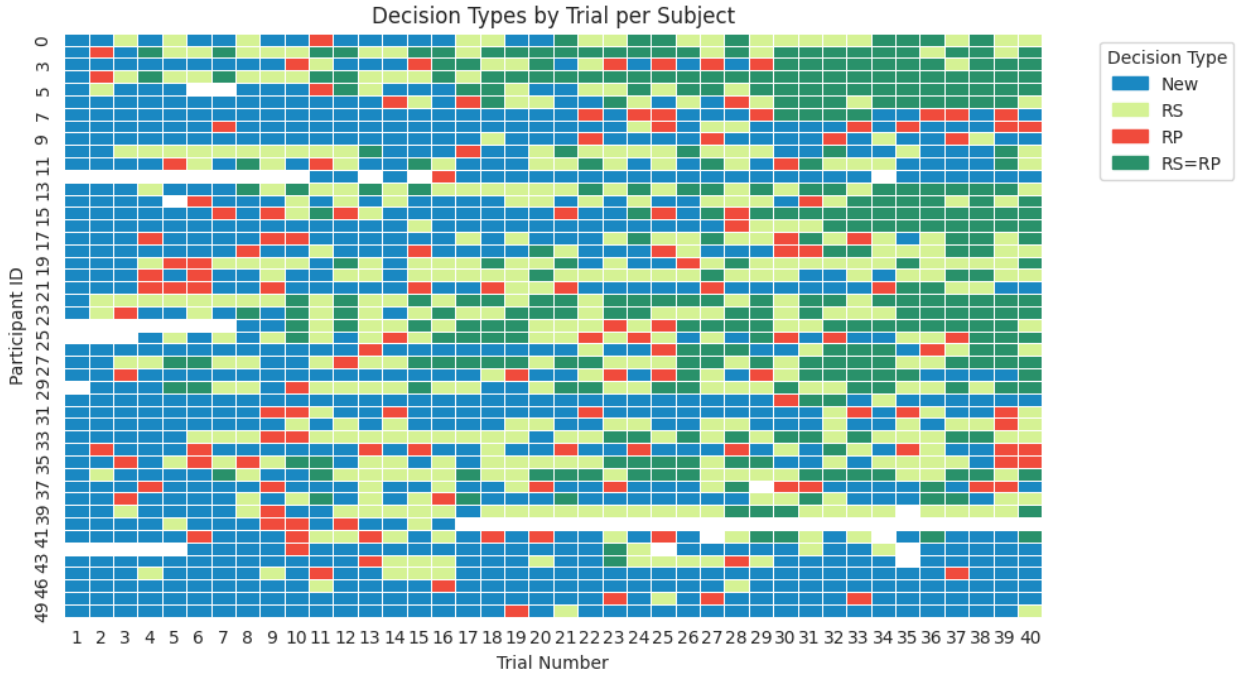

Figure 2.9: **Social learning over time in a single experimental run (N= 50 spatially-embedded)** Each participant's response on a given trial colored by strategy type. Most players begin by sampling new hashtags before finding consistency of some form (light green or red cells), participants with many blue responses did not exhibit response consistency. Participants with many green cells base their responses on rewarded trials, suggesting they effectively learned a normative response within the networked group. White cells represent empty responses.

central to narrative representation (Zwaan, Magliano, & Graesser, 1995; Morrow et al., 1989), we analyzed the causal claims that participants made in the tweet documents.

Priniski et al. (2023) developed a Large Language Model (LLM) system that identifies and extracts causal claims expressed in text documents. The model identifies spans of words serving as inputs to explicitly stated causal relations. Both the cause and effect events, and the underlying causal relation (i.e., a causal trigger) must be explicitly stated for the algorithm to identify the causal claim. The extracted claims are then co-referenced based on BERT embeddings of the identified entities (Devlin et al., 2019), to produce clusters of semantically similar topics, termed "causal topics". The model additionally encodes the direction of the stated causal relationships linking any two clusters.

Networked interaction shifts causal language expressed in personal narratives

To assess if participants expressed more causal language following networked interaction, we fit a hurdle Poisson model to predict the number of causal claims a participant produced at a given phase of interaction. The hurdle Poisson model consists of a logistic classification step for identifying tweets with no identified causal claims, and then a Poisson distribution estimates the counts for remaining documents. Because some participants may be more prone to generate causal reports than others, we fit the model with a random intercept for subject. Effects are to be interpreted as cumulative log odds which describes the expected number of claims in each interaction phase and network condition.

The hurdle parameter predicts that around 54% of documents contained zero causal claims ($hu = .54$, 95% CI [0.51, 0.57]). The intercept equals the expected log count of causal claims before networked interaction in a homogeneously-connected network ($\beta_0 = .19$, 95% CI [-0.04, 0.39]), which equates to a mean of 1.19 claims. Interestingly, after networked interaction, the expected count increases to around 1.61, which is more than pre-interaction counts ($\beta_{Postinteraction} = .30$, 95% CI [0.08, 0.52]). Participants in spatially-embedded networks produced slightly fewer claims than those in homogeneously-mixed networks, however there was not a credible difference ($\beta_{Spatial} = -.12$, 95% CI [-0.34, 0.09]).

Finer-grained analyses of the content of participants' causal claims can reveal what causal topics are most salient in the narrative, and how network structure may shift an individual's representation of the narrative's causal content. We compared differences in claims made after and before interaction to highlight causal relations that may have been enhanced by interaction. As shown in Fig. 2.10, the model identified fourteen causal clusters, plus a non-topic category of claims that couldn't be clustered based on their embeddings. All causal events expressed in the narrative appear as clusters in the corpus (for reference see Fig. 2.1), reflecting a tendency to use causal claims when summarizing events, even in short-formatted messages such as tweets. Because causal relations are directional (causes produce effects), Fig. 2.10 shows the direction of

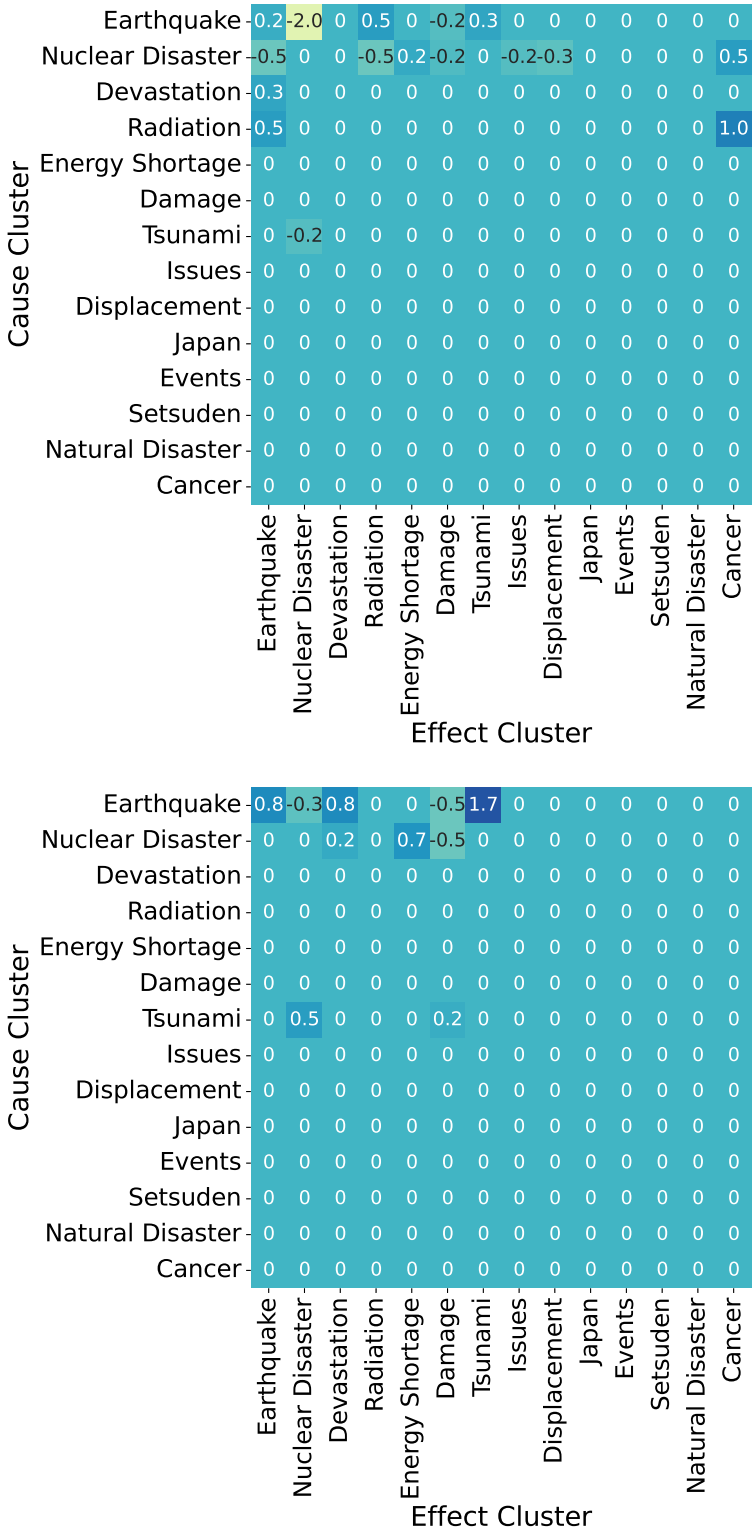

Figure 2.10: **Shift in causal language following networked interaction. Top:** spatially-embedded networks; **bottom:** homogeneously-mixed networks. Cells represent average difference scores of claims instantiating each causal topic across network structures. Cell i, j represents documents claiming that topic i caused j. Positive values indicate more documents expressed that causal relationship after interaction, negative values indicate less causal claims in the tweet documents after network interaction.

the relationships linking any two topics. Rows denote clusters used as a cause and columns denote clusters used as an effect. Cell values are the averaged differences across networked conditions in the number of claims expressing that causal relation after interaction relative to before (i.e, post - pre-interaction). Due to the global aggregation of information in homogeneously-mixed networks, participants in these networks generate tweet messages including a smaller set of causal relations centering around the initial causal chain in the narrative. For example, for the groups with homogeneously-mixed interactions, the causal link *earthquake* $\rightarrow$ *tsunami* had increased by 1.7 documents after interaction, and the subsequent causal link *tsunami* $\rightarrow$ *nuclear disaster* had increased by .5. However, the enhancement of this causal chain is much weaker for the groups with spatially-embedded interactions; the difference scores are .3 and $-.2$, respectively.

Network Experiment 2: Replication of networked interaction on personal narrative generation

In the first social network experiment, we manipulated a network's neighborhood structure to affect behavioral dynamics during interaction. We found that while locally-connected networks facilitated more coordination between pairs of nodes in the network, dominant behaviors emerged more quickly in globally-connected networks. Furthermore, we performed an exploratory analysis of pre- and post-interaction behavior and found that networked interactions induced a coherence shift in a group's behavior. In homogeneously-mixed networks this shift was such that personal narratives about the event became more similar to one another after the interaction than before; while in spatially-embedded networks, due to the localized neighborhood structure producing separable clusters, induced a more scattered shift in causal language use. Furthermore, in homogeneous networks, participants aligned around the initial generative causal chain in the narrative, while those in spatial networks did not. In this experiment, we seek to replicate this shift with a different sample of participants.

Method

We conducted 12 experimental runs with a constant network size of $N = 20$ participants, consisting of six locally-connected networks and six globally-connected networks, where partner pairing was conducted according to the rotational matching algorithm described earlier. The objective of this experiment is to replicate the shift in causal language use measured in the first experiment using a separate sample of participants, and doesn't focus on interpreting dynamics during interactions. Half of the runs consisted of face-naming interaction, while the other half of participants were required to produce hashtags during interaction. In the

face-naming condition, participants were asked to name an image of a face during networked interactions; hashtag coordination condition, participants were tasked with producing hashtags. This allowed us to test the effect of interaction media content on individuals' representation of the processed narrative.

Procedure

The procedure followed the same basic structure as the previous experiment, which consisted of pre-interaction, interaction, and post-interaction phases. In the interaction phases, participants were placed into one of the content conditions and network topologies (face + hashtag; local vs. global connections). The pre-interaction and post-interaction phases of the experiment were identical to before: all participants read the Fukushima nuclear disaster narrative and wrote tweets; then, all participants wrote another tweet and ten hashtags describing the event. Half of the participants coordinated over the name of the face, and half wrote hashtags. This allowed us to test the effect of interaction content on people's reproduction of narrative content after networked interaction.

Results

Replication of causal language findings following networked hashtag interaction

As shown in Fig. 2.11, the language model identified causal claims from nearly half of the participants, with a total of 234 claims in pre-interaction, and 267 in post-interaction tweets. In both blocks, the distribution of the number of extracted claims per subject appears to follow a Poisson distribution. However, the distribution from post-interaction block showed a larger mean and a higher total than from the pre-interaction block. Future work will assess the source of this shift (e.g., decision uncertainty, memory error, or social interactions).

The content of participants' causal claims can reveal what causal topics are most salient in the narrative. We compared differences in claims made following and before interaction to highlight what causal topics may have been enhanced by interaction. As shown in Tab. 2.1 and Fig. 2.12 and 2.13, the model identified eight causal topics (including, but not shown in Tab. 2.1, a non-topic category of claims that couldn't be clustered based on their embeddings). All causal events expressed in the narrative appear as causal topics in the corpus, reflecting a tendency to use causal claims when summarizing events, even in short-formatted messages such as tweets. Because causal relations are directional (causes produce effects), Fig. 2.12 and 2.13 shows the directionality of the relationships linking any two clusters. Rows denote clusters used as a cause, and columns denote clusters used as an effect. Cell values are the averaged differences across networked conditions in the number of claims expressing that causal relation after interaction relative to before. By comparing the distribution of differences across the two structure's heat maps, participants in

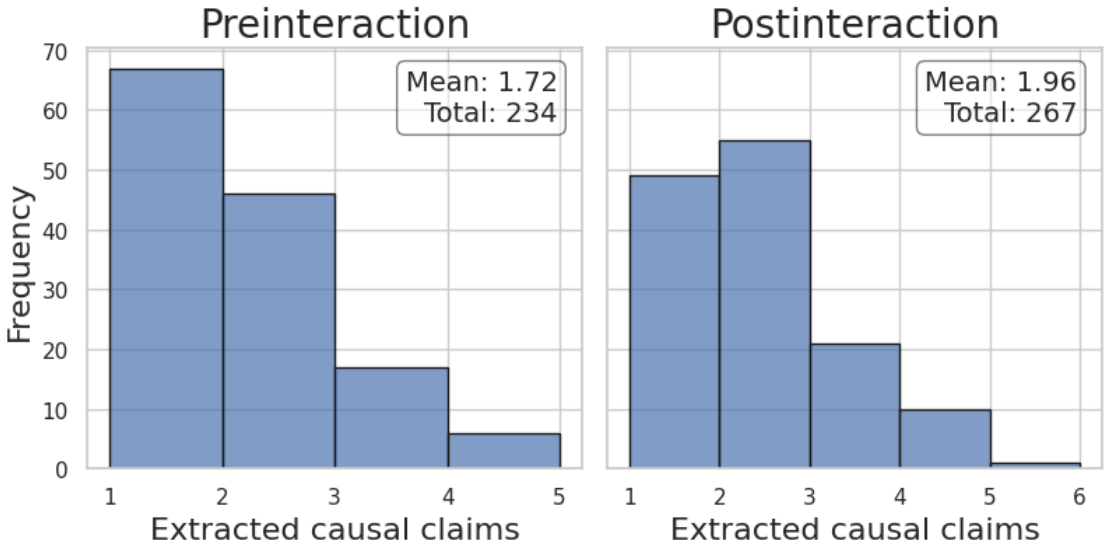

Figure 2.11: **Distribution of the number of extracted causal claims per subject.** Distribution is zero-inflated (zeros not shown), as approximately half of participants didn't produce an identified causal claim.

globally-connected networks exhibited a shift towards expressing the causal chain, while participants in the locally-connected networks expressed a variety of different relations. This is seen by comparing the activated cells that identify the most enhanced causal relations after interaction, in Fig. 2.13 these cells relate to both components of the generative causal chain in the narrative (see Fig. 2.1), in Fig. 2.12, they do not.

ID	Cluster Label	Entities
0	Disaster Event	nuclear disaster, fukushima plant, damage, power, fukushima nuclear, nuclear accident
1	Earthquake	earthquake, tsunami, tohoku, powerful, massive, tohoku, effects, largest, huge
2	Radiation	radioactive, radiation, isotopes, radioactive isotopes, discharge, particles, radioactivity
3	Energy	energy, electricity, shortage, energy crisis, conservation, saving, movement
4	Tsunami	tsunami, 2011 tsunami, resulting tsunami, triggered tsunami, catastrophic
5	Displacement	displacement, displaced, tragic, relocation, relocate, significant, mass, human
6	Change	loss, decrease, increase, change, 70, changes, dramatic decrease

Table 2.1: **Causal topics of "tweets" in Experiment 1.** Clusters of causal entities extracted from "tweets" by a causal language analysis pipeline (J. Priniski et al., 2023).

In the globally-connected networks, the causal link earthquake → tsunami had a difference score of 2 (cell [1,4]), while the causal link tsunami → nuclear disaster had a difference score of 2.2 (cell[4, 0]). While in the locally-connected network, these cells had difference scores of .5 and a 7, respectively.

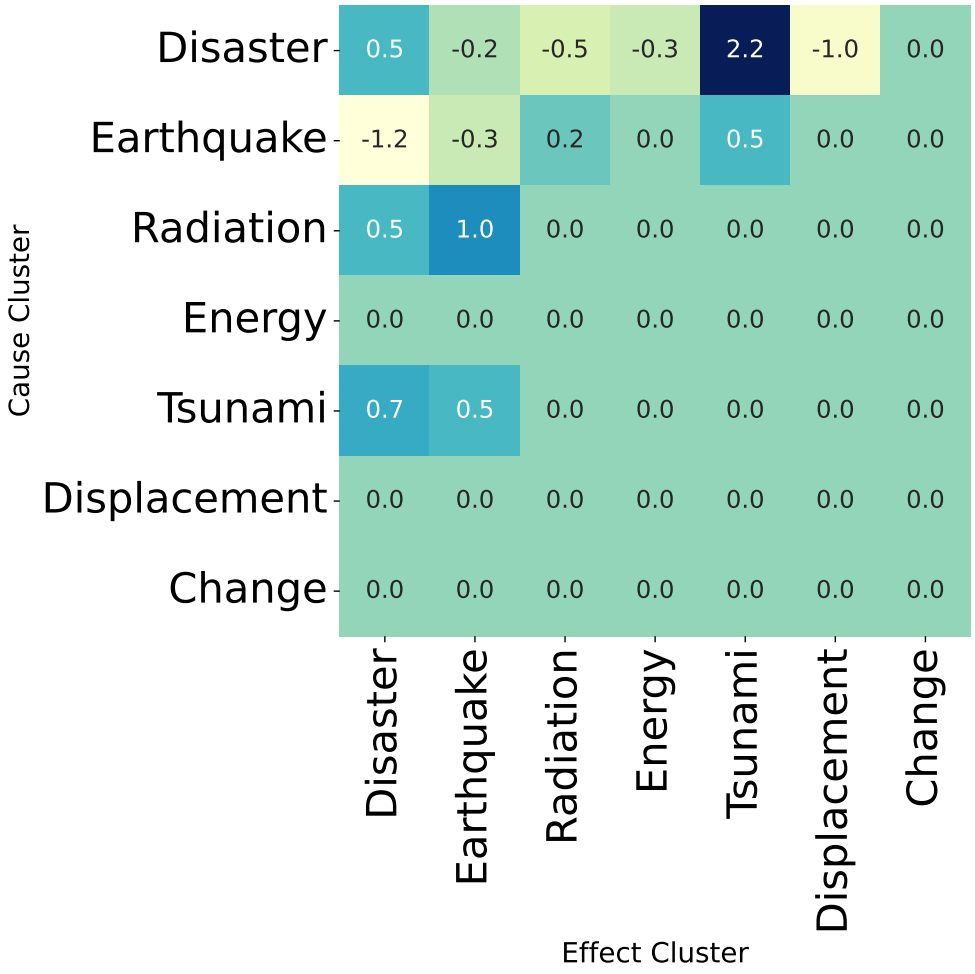

Figure 2.12: **Shift in expressed causal relations following networked interaction in locally-connected network.** Cells represent difference scores of claims instantiating each causal cluster across network conditions. Cell i, j represents documents claiming that cluster i caused j. Large positive values indicate more documents expressed that causal relationship after interaction, negative values indicate more did so before interaction.

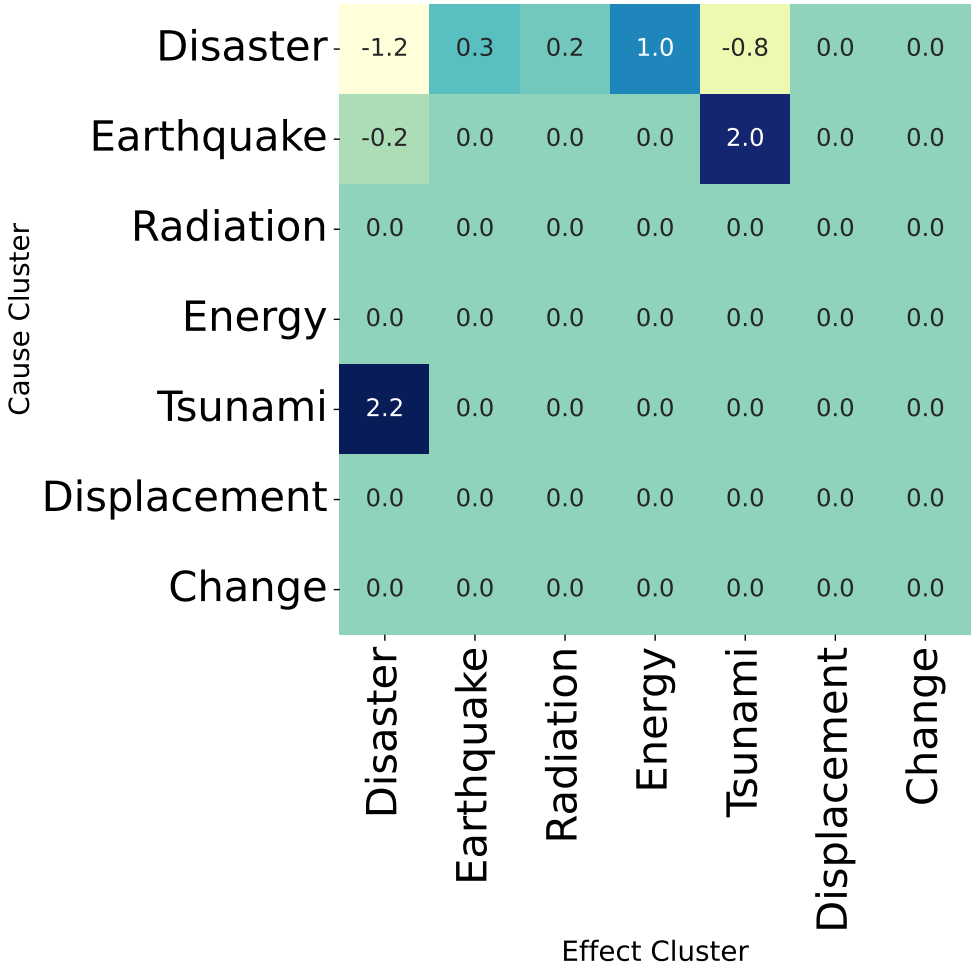

Figure 2.13: **Shift in expressed causal relations following networked interaction in globally-connected network.** Cells represent difference scores of claims instantiating each causal cluster across network conditions. Cell i, j represents documents claiming that cluster i caused j. Large positive values indicate more documents expressed that causal relationship after interaction, negative values indicate more did so before interaction. Note for the globally-connected networks, causal language is enhanced for the generative causal chain in the narrative, and not for locally-connected networks.

Discussion

We examined how the similarity between a full set of hashtags across a network, and the similarity between pairs of coordinating nodes is sensitive to a network's neighborhood topology. We replicated the finding that global connections facilitate the emergence of dominant behaviors, while repeated interactions within localized neighborhoods promotes higher rewards based on coordination, with separable communities producing different hashtag responses.

These findings have direct implications for understanding the onset of echochambers and belief polarization in online social networks. High reward values in the spatial networks proxy social reinforcement for stating one's beliefs online. People will increase their confidence in their stated beliefs when those in their neighborhood reaffirm their responses, despite different beliefs reported from others in different communities. After separate clusters of people begin coordinating on different responses, it will become more difficult to find common ground once beliefs are solidified with one's groups (D. M. Kahan et al., 2012). Our results suggest that one way to increase global agreement within a online network is to structure neighborhoods as to encourage interactions across a wider-array of nodes in the network. Even if participants don't directly coordinate their beliefs with their direct contacts, access to this information could have an aggregate impact on the global consensus of the network.

In addition to analyzing hashtag reporting during networked interaction, we measured shifts in causal language use before and after networked interaction. We found that more causal topics emerged after network interactions, especially more in homogeneous structure than in spatial structure. These results suggest that network interactions have the potential to deepen people's understanding of causality in complex events. Future NLP analyses of "tweets" should parse a wider-array of semantic relations to model shifts in *situation models*, the memory representations people build when processing text-based narratives (Morrow et al., 1989; Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998). This effort could elucidate how networked interactions impact people's narrative representations, and illuminate mechanisms for encouraging healthier discourse online.

Chapter 3

Computational framework distinguishing practical rationality and motivated reasoning

Introduction

There is widespread consensus in the scientific community that climate change is exacerbated by human activities, but a recent Gallup poll indicates that almost 35% of Americans do not agree with this claim (Brenan & Saad, 2018). Evolutionary theory is among the most well-supported theories in all of science, yet more than 40% of US citizens believe in creationism (Gallup, 2014). Meanwhile, large-scale studies have shown that vaccines are not linked to autism (Jain et al., 2015), and still 10% of the United States population believes that the side effects of vaccines are more dangerous than the diseases they prevent (Gallup, 2014). Why do people hold these and other beliefs when they are inconsistent with established empirical evidence? Four decades of research have found that psychological, political, cultural, and sociological factors shape how people form and revise their beliefs (e.g., Alker & Poppen, 1973; Emler et al., 1983; Fishkin et al., 1973; Hickling et al., 2001; Killen & Stangor, 2001; Schult & Wellman, 1997; Shweder et al., 1987; Shweder & Sullivan, 1993). A cross-cutting theme in this literature is that people hold onto their beliefs even in the face of evidence by ignoring or reinterpreting information in a way that supports what they think (e.g., Babcock & Loewenstein, 1997; E. Dawson et al., 2002; Ditto et al., 2018; Gilovich, 1983; Hastorf & Cantril, 1954; Kunda, 1990; Zuckerman, 1979; Jost et al., 2022).

The apparent effects of motivation to maintain one's beliefs are pervasive (e.g., Kunda, 1990; Klaczynski, 2000; Lord et al., 1979; West & Kenny, 2011) and may even be inherent to decision-making processes. Researchers have observed motivated reasoning in political psychology (e.g., D. Kahan, 2013; Kunda, 1990; Taber & Lodge, 2006; Taber et al., 2009), in attitudes about climate change (e.g., Dixon et al., 2019; Hart & Nisbet, 2012), and in science literacy (e.g., Pasek, 2018; Drummond & Fischhoff, 2017; Druckman, 2015). Even practicing scientists who are aware of the effects of motivation on cognition are not immune to their influence (e.g., Simmons et al., 2011; E. C. Yu et al., 2014).

Social scientists often study motivated reasoning by conducting case studies guided by verbal theories (i.e., theories which make only qualitative predictions; Alker & Poppen, 1973; Emler et al., 1983; Fishkin et al., 1973; Hickling et al., 2001; Killen & Stangor, 2001; Schult & Wellman, 1997; Shweder & Sullivan, 1993; Babcock & Loewenstein, 1997; E. Dawson et al., 2002; Ditto et al., 2018; Gilovich, 1983; Hastorf & Cantril, 1954; Zuckerman, 1979), rather than performing comparisons against computational models that allow for quantification of the factors that ought to impact how people update their beliefs (e.g. Cook & Lewandowsky, 2016; Jern et al., 2014; Pilditch et al., 2022). Verbal theories and the often implicit definitions of motivation that accompany them come at the cost of introducing ambiguity; it is difficult to determine the cognitive mechanisms that produce people's seemingly irrational behavior. An example can illustrate this point. Many political and fiscal conservatives are skeptical that human activities impact Earth's climate, despite near uniform agreement among climate scientists worldwide. One possible explanation of this rejection is that accepting the reality of human-caused climate change would entail imposing increased limits on free-market capitalism, an unwanted outcome to many conservatives (Cook & Lewandowsky, 2016). In turn—so the argument goes—conservatives are motivated to reject or reinterpret evidence of human-caused climate change (for evidence this occurs in other domains, see A. G. Levy et al., 2022)

However, other explanations are available. It is possible a climate skeptic has not accessed the relevant facts in forming their beliefs, instead basing their views on unreliable sources. Conditional on these inaccurate prior beliefs, people may reason rationally based on the information at hand (similar phenomena have been observed in other domains, see Jern et al., 2014). This pattern of behavior could, on its face, appear to be evidence of motivated reasoning, but the mechanisms underlying the output would be quite different than what researchers have typically suggested. The error would not be in how people make inferences, but rather in the *inputs* to their inferential machinery (e.g., the accuracy of their prior beliefs; veracity of the information they are reasoning about). There are now well-established alternative hypotheses that explain effects previously assumed to be evidence of motivated reasoning as the result of rational reasoning processes (e.g. D. Kahan, 2013; Jern et al., 2014), but in much of the research on attitude and belief change, the conceptualization of motivated reasoning-like effects as evidence of a cognitive defect persists (e.g., Dixon et

al., 2019; Pennycook & Rand, 2019).

A central problem in much of the research on motivated reasoning is measurement. Prior studies of motivated reasoning have often focused on examining how evidence impacts beliefs that are inextricably linked to aspects of one's identity (e.g., beliefs about politics, religion, or morality). These topics provide a more naturalistic test of the impact of motivation on reasoning, but certain aspects of these topics can obscure the underlying cognitive mechanisms. For example, *what kind* of information is objectively relevant to believing in human-caused climate change, *how much* evidence should a reasoner accommodate to do so, and most importantly, how would one even begin to *quantify* this evidence to measure its direct effect on a reasoner's beliefs? The very nature of evidence related to climate change (such as consensus among scientists, mathematical models, severe weather events) will make it difficult to measure the extent to which motivation impacts people's beliefs about climate change because it is unclear how much each piece of evidence should impact what someone should believe. Or to consider another example, if we provide people with evidence that vaccines are safe and effective in the form of summary statistics from large-scale trials, how much evidence do these statistics actually convey and, objectively, how ought participants update their beliefs in light of this evidence?

In both cases, even if one could quantify the direct impact of evidence on beliefs, a principled mathematical benchmark needs to be assumed to properly understand the extent to which motivation biases people's reasoning, a difficult task when relying on verbal theories. Some researchers now argue that once a benchmark is defined, and human reasoning is tested against optimal models of belief updating instantiating this benchmark, it appears that people approximate these optimal models more than psychologists have often assumed (e.g. Dasgupta et al., 2020). For example, Jern and colleagues (2014) argue that the normative standard for reasoning under uncertainty is a kind of probabilistic inference, and therefore should be subject to the axioms of probability theory. Surprisingly, they find that when comparing participants' performance to this normative model, seemingly irrational decisions—such as belief polarization—conform to (approximately) Bayesian reasoning (also see, Little, 2021). Work in several other domains has reached similar conclusions over the last two decades (Austerweil & Griffiths, 2011; Griffiths & Tenenbaum, 2006; Tenenbaum et al., 2006; Vul et al., 2014; Zimper & Ludwig, 2009; Dasgupta et al., 2020; Jin et al., 2022; Wallace, 2020; A. J. Yu & Cohen, 2008). Whether one accepts their interpretation of the data, it highlights the need to delineate the factors that ought to impact the beliefs people form and be more specific about what constitutes rational behavior in these contexts.

A further assumption that is overlooked not only in much of the original research on motivated reasoning but also newer Bayesian explanations of heuristics and biases, is a somewhat technical but nonetheless important point. Namely, early research on motivated reasoning and contemporary Bayesian interpretations

often assume that the rationality of the cognitive processes (e.g., how new data and base-rate information are combined) that deliver an output (e.g., a posterior belief) dictate whether this output needs to be corrected. To speak generally, social psychologists' interest in motivated reasoning is not limited to understanding the operations of the mind; rather, motivated reasoning is studied *because* people appear to believe many things that are inconsistent with the facts and their behavior has substantial societal implications – whether for vaccination uptake (e.g. Horne et al., 2015), climate change policy support (e.g. Lewandowsky, Oberauer, & Gignac, 2013), or social justice (e.g. Kraus & Tan, 2015). These beliefs demand correction as a matter of public policy. Assuming that the source of the underlying problem is a bug in our inferential machinery is thus a plausible starting place.

Papers offering alternative Bayesian explanations (Dasgupta et al., 2020; Lieder & Griffiths, 2020) seem to make a similar assumption about the connection between underlying processes, output, and the need for correction, but the conclusions they draw from these assumptions are exactly contrary to those of many practicing social psychologists. Researchers examining the preceding questions from a Bayesian perspective suggest that if the underlying processes producing polarization, or motivated reasoning-like effects, are the result of optimal Bayesian inference, there is not much to be done – people are reasoning as optimally as they plausibly could (but see Cook & Lewandowsky, 2016). In fact, evaluating and attempting to correct the outputs of people's inferential machinery and evaluating the rationality of the generative processes are *separate issues*. Scientists are not (and should not) be interested in motivated reasoning just as a psychological question, but because misconceptions (whatever their cause) can impact society, requiring policy-makers to act (Cook & Lewandowsky, 2016; Pilditch et al., 2022). The ability to distinguish the rationality of the generative process and evaluate how to correct its outputs depends on our ability to measure key psychological constructs and to use these measurements in a computational model, or so I will argue.

In this chapter, I describe three broad factors—prior information, evidence, and utility—that should normatively affect how people update their beliefs when confronted with new evidence. I define a computational framework for distinguishing Bayesian updating, motivated reasoning, and practical rationality, which I detail in the next section to demonstrate how these factors are integrated during belief updating. This model serves as a computational framework for reevaluating prior research on motivated reasoning and allows us to devise a strategy for measuring and locating the impact of motivation on belief formation and updating. I conclude by discussing the open questions and implications of this framework.

Conceptual foundations

In this section, I aim to sharpen how I talk about rationality and reasoning in the context of research on motivated reasoning. Specifically, I will suggest that distinguishing theoretical rationality from practical rationality helps us clarify how we talk and think about motivated reasoning.

Researchers studying motivated reasoning often appear to assume different accounts of rationality—of how people ought to reason—but a thorough discussion of just what these accounts entail is not always stated in many empirical papers. For example, many researchers seem to assume that so long as people are reasoning in ways consistent with the axioms of probability theory, we can say they are reasoning rationally (e.g. Jern et al., 2014), a kind of *probabilism* (e.g. Pettigrew, 2019). Or, on a more recent account, once we account for the fact that they have limited resources (e.g., time), people reason more rationally than they initially may appear to (e.g. Lieder & Griffiths, 2020). Here, we'll take a broader perspective, one which is rooted in established distinctions in epistemology. A core distinction we'll focus on is between theoretical and practical rationality. From this broader perspective, we'll see how making decisions which accommodate the axioms of probability theory, but also taking into account directional goals, could be practically rational. Before turning to this philosophical literature, we'll first consider an influential account of motivated cognition.

Kunda's (1990) account of motivated reasoning and the distinction between theoretical and practical rationality

What is the link between rationality and motivation? On one influential account, motivation is goal-orientation and *all* reasoning is motivated by goals (Kunda, 1990). The kind of goal representation that impacts a belief distinguishes rational motivated reasoning from defective motivated reasoning. For example, according to Kunda (1990) *accuracy goals* shape interpretations of evidence to cohere with states of the world. In contrast, *directional goals* shape inferences to cohere with conclusions one may wish to draw about states of the world. On this account, accuracy goals ought to motivate how we update our beliefs, but directional goals ought not to.

I argue that two broad factors ought to shape what people believe. One factor involves doxastic considerations (e.g., one's priors, evidence, and the like). A second factor involves practical considerations, in this case, an assessment of the utility associated with believing a particular proposition, including an assessment of the utility of one's belief being true and an assessment of the *consequences* of holding that belief (Radcliffe, 1999; Von Neumann & Morgenstern, 2007; Schoenfield, 2018). I'll give some reasons why I think this shortly, but a further bit of jargon is necessary. I distinguish between two types of doxastic states, first-order and second-order beliefs, which track different internal representations and integrate different sets

of information. *First-order beliefs* are constructed from only evidentiary considerations. They are beliefs as people commonly understand them, such as the belief that the distance between Los Angeles and New York City is less than 3,000 miles. When beliefs of this sort fail to represent the way the world really is based on the evidence at hand, the belief is false. To our knowledge, motivated reasoning as it is typically construed is a strictly *first-order* account because researchers assume that directional goals shape people's beliefs as they are being constructed, perhaps because people sample from the evidence inappropriately or because they improperly integrate new evidence and their prior beliefs. On this view, researchers have *assumed* that irrational motivated reasoning occurs in any cases where directional motivations impact what one believes (for discussion of this account, see Little, 2021).

However, directional goals may influence beliefs in other ways, or so I will argue — for instance, by realigning doxastic states with practical considerations after a posterior is initially computed. *Second-order beliefs*, as I define them, are belief states that incorporate not only doxastic considerations, but practical considerations as well (see Jachimowicz et al., 2018; Cialdini et al., 1991). I operationalize these practical considerations as goal representations, or more precisely, as a utility-calculus defined over counterfactual world states resulting from an action. Our view is that goal representations can shape second-order beliefs during a second-order inference step: after people have properly integrated evidence and determined the utility of taking certain actions. This is a second-order influence on doxastic states, which is why I call them second-order beliefs. Thus, second-order beliefs can incorporate directional influences, in turn diverging from a belief formation process that only relies on probabilistic and causal information from the environment.

An example will make the point clear. A Republican may be uncertain that human activities affect the climate based on the evidence they are aware of, but decide to report believing human activities do not impact the climate after recognizing that many others in their community report the same belief. This could be for evidential reasons – knowing other Republicans don't believe some proposition could have evidential value. But this isn't the only possibility; Bill, a Republican, could recognize that just because their fellow Republicans think something does not mean it's true. However, they know disagreeing with their fellow Republicans could be socially costly. Consequently, when the evidence is uncertain, they may report they don't believe in human caused climate change based on this utility calculation.

Is it ever rational for one's beliefs to incorporate a directional influence in the way we've just described? In Kunda's (1990) terminology, it is rational for reasoning to be motivated by accuracy, but directional goals are fundamentally at odds with rationality. Indeed, one implicit assumption in much of the research on motivated reasoning is that the utility of holding a belief—which is directional—is *irrelevant* to what one ought to believe (Little, 2021). What makes motivated reasoning pernicious is the fact that utilities directionally impact people's beliefs in the first place.

Contrary to this assumption, Bayesian epistemologists have argued that there are situations where, when the evidence is logically compatible with multiple hypotheses, a decision problem arises that requires people to assent or dissent to a hypothesis on the grounds of expected utility rather than evidence alone (Maher, 1993). I unpack this idea below.

Theoretical and practical rationality

We can understand how rationality and motivated reasoning relate, as well as how these issues contrast with the account I develop below, by briefly reviewing work on the distinction between theoretical and practical rationality (Wallace, 2020).

Theoretical rationality concerns the relationship between the acceptance of a proposition and its truth value. This kind of rationality concerns the evidence for and truth of propositions absent the consequences of believing them. *Practical rationality* has an end that is not truth per se. Rather, the aim of practical rationality concerns the value of taking actions (where actions need not be literal physical acts; Wallace, 2020). This kind of rationality concerns actions being good or worthwhile (where goodness is not necessarily morally valenced). Thus, theoretical rationality concerns changes in sets of beliefs, and practical rationality concerns reasons that give rise to actions, including intentions and *reporting* one's beliefs (Harman, 1986; Bratman, 1987; Wallace, 2020). Theoretical rationality describes what one should believe, and practical rationality describes what one should do. Thus, for theoretical rationality, it is irrational to update one's beliefs in a way that is incompatible with the evidence once we recognize an inconsistency. In contrast, philosophers typically suggest that practically irrational behavior is close to a kind of weakness of will (Wallace, 2020). For example, I form a plan to exercise more, I ascribe high utility to exercise, but I fail to execute on my plan – this is practically irrational behavior.

Thus, the primary concerns of theoretical and practical rationality are distinct. Theoretical rationality concerns doxastic states or *credences* and their fit with the facts about the world. Credences can be thought of as complex, fine-grained beliefs about the world – they vary in degree of certainty and roughly correlate with the subjective probability that some proposition is true (Jackson, 2019). Practical rationality concerns *intentions or acts*, which incorporate doxastic information about the world but will also include an aim to realize some plan for acting optimally, broadly defined.[1]

It is tempting to assume that people should only shift their beliefs to reflect the evidence they're presented with and that's it. I am going to cast doubt on this idea. To do this, we'll focus on how participants respond in a typical experiment examining whether people are engaged in motivated reasoning. For example,

[1] I should also note that we've drawn a sharp distinction between theoretical and practical rationality, but this distinction may not be so clear. Our beliefs may *always* include some action-taking component.

psychologists measure whether participants' beliefs are improperly influenced by information (e.g. Nyhan & Reifler, 2010; Nyhan et al., 2014a; Ditto & Lopez, 1992; Ditto et al., 1998). But to properly assess potential cases of motivated reasoning, psychologists must also assess whether a reasoner's goals cause a mismatch between their beliefs and the facts at hand (e.g., Ditto & Lopez, 1992).

Practically rational behavior can be directionally motivated

In typical experiments on motivated reasoning, psychologists measure *belief reports* which they hope reflect people's underlying beliefs directly. But belief reports are not strictly internal representations. Rather, belief reports are second-order decisions that extend belief representations to jointly encode an associated action-plan, and thus entail evaluation of the utility of taking certain actions. Second-order beliefs, including belief reports, can thus diverge from credences if people attribute higher positive utility to one belief over another. In Maher's (1993) terminology, reasoners encode a cognitive utility function that assigns utilities to the consequences of rationally accepting hypotheses in ways that may diverge from purely probabilistic information. A consequence of this fact is that people's belief reports not only involve their internal credences, they also involve goal-based considerations facilitating optimal actions.

Given this fact, I argue that researchers need to instead examine how people's belief reports match what would be considered practically rational behavior given both doxastic considerations and their goals. What would this look like? For example, Jones initially thinks he has enough petrol in his car to reach his destination (a first-order credence), because he filled his tank most of the way yesterday. However, Jones realizes it is particularly important that he has enough fuel because there are no petrol stations for the next 100 miles, so he decides to stop for petrol to play it safe (a second-order belief with an accuracy goal). Researchers evaluating Jones' belief report "I need to get petrol before departing" should evaluate not just the evidence Jones had (i.e., a mostly full tank) but also how his goals impact his decision. Failing to consider Jones' accuracy goal and only focusing on his evidence would make it appear that Jones' belief report is irrational.

Researchers often assume belief reports directly reflect internal credences. As a consequence, when participants respond to scale items measuring their attitudes, their responses may be influenced by utility information in ways the researcher doesn't account for. In some ways, this is a well-established issue in survey research; there are a litany of biases which can impact the interpretation of participants' responses, including desirability biases and other similar demand characteristics. I draw a separate distinction because, as I will suggest, belief reports could also produce sustained changes in credences via coherence mechanisms; this is of course exactly the opposite pattern as task demands, which are transient effects by definition. As

I discuss in more detail below, models that acknowledge the relevance of the utility of holding a belief – models of practical rationality – can help identify the impact of utilities on people's credences and belief reports.

Unpacking ambiguity in the rationality of belief reports

To better understand this distinction and how making it can be useful in understanding belief reporting in real-world contexts, consider a common situation on social media. People learn about social, political, and moral issues online, and at least 50% of Americans receive some of their news from social media (Walker & Matsa, 2021). Social media discussions often mix first-order information (e.g., facts, statements about the issues) with second-order information (e.g., engagement metrics of friends, network information, etc.).[2] This complicates the picture of how people's beliefs are influenced in modern information environments, particularly for topics psychologists have typically studied.

Consider a situation in which Smith is uncertain about the truth of a provocative hypothesis, but believes that publicly endorsing that hypothesis could yield a positive social reward. For instance, Smith believes that the evidence for the hypothesis $\mathcal{H}$ "increased firearm regulation would **not** have stopped the mass shooting in Homesville, U.S.A." is completely uncertain. But Smith predicts that taking this stance on social media in light of a tragic shooting could lead to increased engagement with their content and profile: something they attribute positive utility to. These second-order considerations result in a high utility for stating $\mathcal{H}$ given their internal credence is uncertain, which is enough reason for Smith to post There is zero-chance that gun control would have stopped the Homesville mass shooting. There is experimental evidence this occurs: people see utilities in forming beliefs, and anticipate the consequences of those beliefs (Golman & Loewenstein, 2018; Falk & Zimmermann, 2016; Jachimowicz et al., 2018; A. G. Levy et al., 2022).

How strongly does Smith believe what they said? One possibility is that the predicted increase in online engagement caused them to believe that the evidence for H was indeed very strong. This is a clear case of directionally-motivated reasoning as it is traditionally conceived. This social media user has sampled from evidence in a way so as to discount evidence against support of gun control and in line with the reward (here, more likes, upvotes, and retweets).

A second possibility is that Smith believes the evidence for gun control is uncertain, but after assessing the evidence and weighing the utility of reporting the belief, he has *formed the intention to believe* that gun control is not effective at reducing violence, and acted in accordance with that intention by posting his skepticism about gun control legislation. In this second case, it could be *practically rational* to decide to

[2]This information is second-order because it encodes what other people think about the facts at hand. A common way to construe second-order beliefs in the motivated reasoning literature is as socially normative beliefs based on other people's beliefs and behaviors (Jachimowicz et al., 2018; Bicchieri, 2005).

believe $\mathcal{H}$, given Smith's priors (i.e., mixed evidence about the effectiveness of gun control laws) and utility-calculus (e.g., social media engagement is more valuable than encouraging pro-gun regulation attitudes) (Maher, 1993; Briggs, 2019). In this case, we would need to assess the rationality of their behavior by considering both the evidence at hand and the utility of reporting their belief.

This second possibility highlights an important distinction. What is practically rational for Smith is not what is morally or scientifically correct given the facts of the world. Practically rationally behavior can still entail someone has a false belief about the world that requires correction, but will likely require a *different form* of correction than what researchers may commonly assume. In this case, intervening on the shape of their utility calculus might be more effective because the problem could be in their assignment of utilities to a belief report.

The central point here and the models I develop in the next section of the chapter is to highlight that we cannot experimentally distinguish between biased evidence sampling and practically rational behavior when we have only measured belief reports; we need to quantify the intervening evidence, participants' understanding of the evidence, and the utilities a person assigns to reporting a belief. This is the focus of the remainder of the paper.

Computations underlying motivated reasoning and practical rationality

I distinguish three factors that should normatively influence how people update and report their beliefs. The first factor is the evidence on which people update their beliefs. The evidence people have and how it is weighted could take a number of forms. For example, people believe things because of the testimony of their friends or experts. Likewise, they believe things based on their own observations about the world. Reasoners assign some weight to each kind of evidence – the extent to which they think it supports a hypothesis. For example, some evidence, like anecdotes from friends, might be perceived as relatively weak evidence compared to expert testimony. Of course, trust in expert testimony will dictate the weight evidence is likely to be assigned (Imundo & Rapp, 2022).

The second factor is what people believe, and the information available to them, *before* considering new evidence – their priors. This could be in the form of known base-rate information, or a mere hunch about the credibility of a hypothesis. We'd expect some of our priors to be very weak, perhaps even flat, whilst others to exert a strong effect on what we believe even after we confront new data. For example, our prior that the number of atoms in the universe is an even number might be completely uncertain, but our prior that

extrasensory perception is not real might be extremely strong. In the former case, even a little bit of new information could shift our credence in the proposition, but a lot more data would be needed to materially impact our belief in extrasensory perception.

The third factor is the utility of holding a belief. This is the least clearly defined. Here, I will assume that a reasoner who is motivated to believe a hypothesis associates some *utility* with the prospect of holding that belief. For instance, utility could be associated with the social benefits attained by displaying in-group support of climate change skepticism (a directional goal; D. Kahan et al., 2017; Sidanius & Pratto, 1999). People may also attribute high utility to maintaining coherence between their beliefs (an accuracy goal). People may perceive some hypotheses to have extremely large utility (e.g., eternal salvation) or little to no utility at all (e.g., believing the number of atoms in the universe is an even number).

I'll now discuss how these factors feature in the computational mechanisms supporting motivated reasoning and practical rationality. I first describe the typical mechanisms for generating first-order beliefs, namely, using Bayesian updating. Models of both motivated reasoning and practical rationality perform Bayesian updating to update credences and generate belief reports. What distinguishes these models — and the types of reasoning they instantiate — is that the motivated reasoning model computes credences via Bayesian updating using a directional prior — a prior which encodes utility information about hypotheses being true. In contrast, the practical reasoning model entails a two-step process that separates priors from the utility: the model first computes posteriors via Bayesian updating before integrating utility information to produce a second-order belief report (see Figure 3.1).

First-order processes

Bayesian updating: Computing first-order beliefs with normative priors

Bayesian updating is the standard model for computing first-order beliefs from data. Both the motivated reasoning model and the Bayesian decision-theoretic model generate beliefs via Bayesian updating. People hold beliefs about hypotheses, and these hypotheses include a subjective likelihood attributed to each hypothesis being true. A hypothesis could be a logical, numerical, or natural language proposition that explains some feature of a (logical, statistical, or real-world) system. A single hypothesis, h_i, comes from a larger set of hypotheses called a hypothesis space, $\mathcal{H}$. The hypothesis space expresses a set of hypotheses which partition a set of *possible worlds* as either true (i.e., actual) or false, where "possible worlds" just means different ways the world could be. Let $h_i \in \mathcal{H}$ be a hypothesis from a hypothesis space.

People deliberate about hypotheses given data, which needs to be reflected in the framework. We need to computationally define how evidence is integrated with prior beliefs. Bayesian updating is a normative

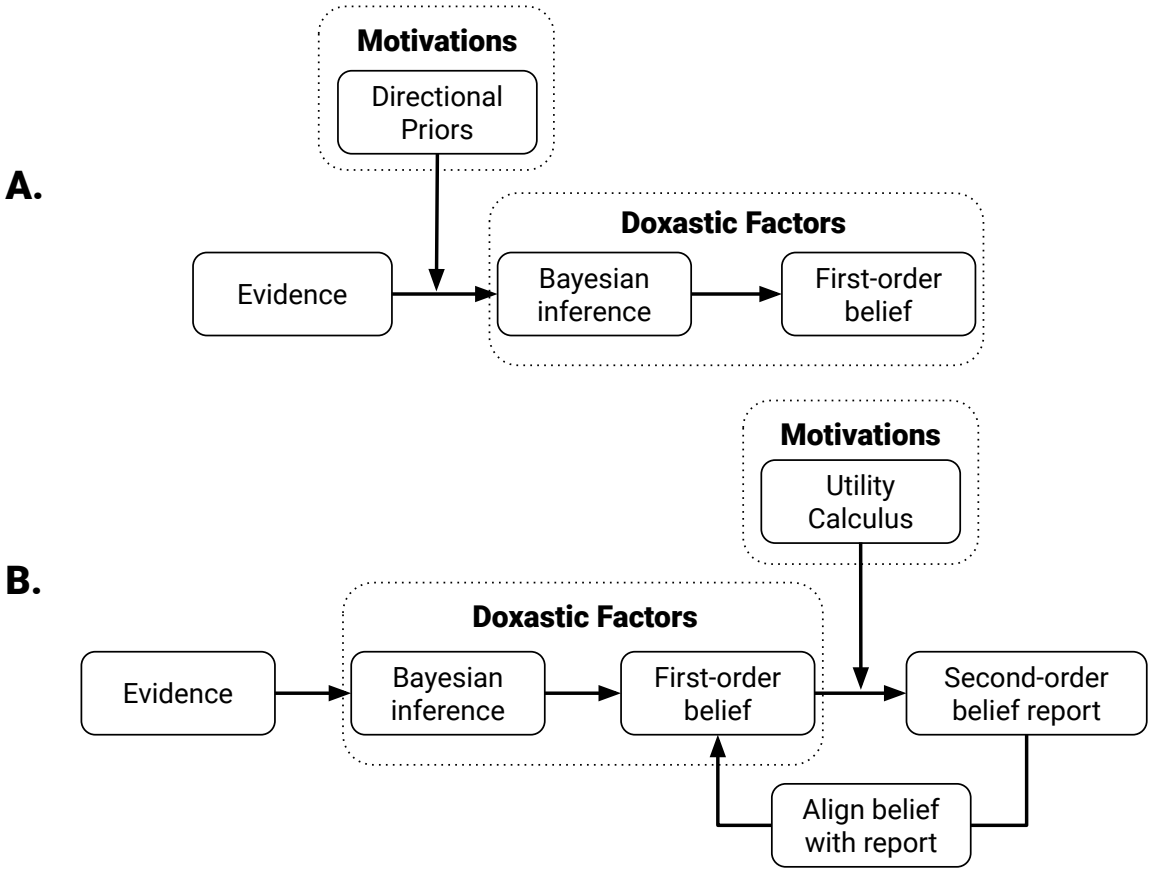

Figure 3.1: Flowchart showing two ways motivations can shape belief reporting. Diagram A is the motivated reasoning account. This figure captures the direct influence of motivations on first-order belief construction. Previous research doesn't specify how this happens, but I introduce directional priors to quantify this impact. Diagram B is the practically rational account. It shows how utilities can exert influence on normative first-order doxastic representations to generate a second-order belief. Internal credences (i.e., first-order beliefs) are then aligned to cohere with the reported belief.

framework to do this. Bayes' Theorem describes how to update the perceived probability of a certain hypothesis, h_i, given new data, d. It is composed of three parts: a prior $\mathbb{P}(h_i)$, a likelihood $\mathbb{P}(d|h_i)$, and a normalizing constant given the observed data $\mathbb{P}(d)$. When a prior is not conditioned on utility information, we say that it is a *normative prior*. Normative priors are the basis of the Bayesian decision model which I describe at the end of this section. The motivated reasoning model uses non-normative priors because they are dependent on utility information.

Let $d \in \mathcal{D}$ be a representation of some data d sampled from a data space $\mathcal{D}$. For instance, if this summer is warmer, d, than previous years $\mathcal{D}$, a reasoner would need to update their beliefs about how hot summers are likely to be. Data is evidence that can be integrated to generate doxastic states. We can express the updated probability in their hypothesis h_i given d using Bayes' Theorem:

$$\mathbb{P}(h_i|d) = \frac{\mathbb{P}(d|h_i)\mathbb{P}(h_i)}{\mathbb{P}(d)} \tag{3.1}$$

where $\mathbb{P}(d)$ is a normalizing constant, that scales the computed posterior values so they sum to one and form a probability measure. Modelers often disregard the probability of the data and leave the posterior unscaled because doing so leaves the relative credences unchanged and simplifies the computations. The posterior calculation can therefore also be simplified to:

$$\mathbb{P}(h_i|d) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i) \tag{3.2}$$

Bayesian updating is the normative process (that is, it satisfies the axioms of probability theory) for integrating information into belief states. I discuss how Bayes' Theorem does this in more detail in the appendix. It is uncontroversial that beliefs updated using this method are theoretically rational (Vineberg, 2022). I'll denote doxastic states that are only updated on the data as *first-order beliefs*. First-order beliefs are outputs of the Bayesian updating module from the computations in Equation 3.1 above.

Motivated reasoning: Directional priors affect evidence integration

In the Bayesian decision model that I will detail after this section, normative priors—priors which are not impacted by motivation—dictate how evidence is integrated to form a first-order belief. This is in contrast to the common account of motivated reasoning, which assumes that people are theoretically irrational when motivations shift how they compute their first-order beliefs. I develop a motivated reasoning model where goals bias the construction of first-order beliefs by directing prior credences to sample data for hypotheses ascribed higher utility. I define *directional priors* as prior credences conditioned on utilities. Directional priors scale the prior evidence for hypotheses by their utilities. I use the notation $\mathbb{P}^*$ to define directional

priors, where one's prior credence for a hypothesis h_i is conditioned on the output of a cognitive utility function $U(h_i)$:

$$\mathbb{P}^*(h_i) = \mathbb{P}(h_i|U(h_i)) \tag{3.3}$$

Directional priors influence evidence integration and produce directionally motivated beliefs, so we can interpret Equation 3.1 in light of Equation 3.3, by substituting the normative prior $\mathbb{P}(h_i)$ with the utility-conditioned, directional prior $\mathbb{P}(h_i|U(h_i))$. Evidence updates a posterior by multiplying the prior and the likelihood. The likelihood quantifies the amount of information a piece of data has, conditioned on a hypothesis being true; in other words, how much learning should occur given the observed data under *each* hypothesis. Priors, on the other hand, encode the *a priori* probability a reasoner attributes to each hypothesis being true. The likelihood cannot be interpreted independently of the prior because the prior dictates the extent to which the likelihood function will shape a first-order, posterior belief. Credences sampled from $\mathbb{P}^*$ are theoretically irrational because directional utilities should not shape first-order beliefs (Kunda, 1990). When they do, first-order beliefs contain additional information than the causal and statistical regularities of the environment, and could be out of line with how the world actually is (at least when a reasoner attributes high utility to directional rather than accuracy goals). Note that revisionist accounts of rationality may disagree with this claim, but this point is beyond the scope of the paper (Lieder & Griffiths, 2020).

From a modeling perspective, we need to specify a $\mathbb{P}^*$ that best suits a given domain. To illustrate, in the next section I define, implement, and simulate predictions from Bayesian models with optimal (normative) and motivated (directional) priors. I show how $\mathbb{P}^*$ can be defined as a β distribution based on a utility function assigning positive values to possible simulation outcomes. In this motivated reasoning model, an expectation is equal to the expected relative utility of a hypothesis. I simulate predictions with varying utilities to demonstrate how they impact the model's credences.

Thereafter, I provide a toy example of a study which could distinguish motivated reasoning from practical rationality, where I measure the parameters discussed above. I relate two models (the motivated reasoning model and a practically rational reasoning model) to this toy study to show how this structure may be able to distinguish motivated reasoning from practical rationality.

Second-order processes

Computing the expected utility of beliefs

I now describe an expected utility account of how second-order belief states can be derived from a posterior computed using normative priors (i.e., the Bayesian decision framework which I use to build a practically

rational reasoning model). I will first describe how expected utilities can be computed for hypotheses and then describe how to incorporate Bayesian inference in this model.

I'll write the expected utility of reporting h_i as:

$$EU(h_i) = \mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{3.4}$$

where $\mathbb{P}(h_i)$ is the probability that a given hypothesis h_i is true, and $\mathbb{U}(h_i)$ is the utility the reasoner attributes to believing h_i. For instance, adopting a new belief can be sensitive to consequences (Williams, 2021), such as maintaining or severing ties to one's social group (D. Kahan, 2013). Social consequences broadly, and the maintenance of the ties we form to our social groups particularly, have utilities. Consider a reasoner forming a belief about climate change. The consequences of believing in human-caused climate change is the summation of the consequences of believing that proposition is true. What does that mean? Some consequences could be increased business regulation, gasoline taxes, and being discouraged from air travel unless it is necessary. There are further consequences, of course, such as the social relationships we maintain as a result of taking a side on a controversial issue. I provide quantitative details for how researchers can operationalize utilities and the consequences of hypotheses in the appendix.

When people choose between hypotheses with varying utilities, our framework assumes that they will prefer hypotheses with higher utility to hypotheses with lower utility. Therefore, the preference of an outcome depends on the values of other possible outcomes (Von Neumann & Morgenstern, 2007); all else being equal, people prefer outcomes which yield the greatest relative expected utility. The framework expresses the utilities associated with hypotheses as relative expected utilities (the expected utility of a hypothesis *relative* to the other hypotheses in the hypothesis space). I use the function $z(h_i)$ to define the *relative utility* for a hypothesis h_i given the utilities of the other hypotheses. The relative utility can be thought of as a scaled utility value, where utilities for all outcomes are normed to be between the values of zero and one. (I define a relative utility function because it will be helpful when explaining computations in the cognitive models.)

There are various ways a relative utility function can be computed as the shape of this equation depends on the preference space (e.g., Regenwetter et al., 2011) and other psychological mechanisms (e.g., loss aversion; Tversky & Kahneman, 1991). Here, I describe the simplest relative utility function, which is essentially a value of the summation of outcomes for each hypothesis scaled by the summation of utilities for all other outcomes for all other hypotheses. I am not saying this is the computational mechanism for preference formation, but I use this equation simply to illustrate our framework. I define the relative utility of a hypothesis, $z(h_i)$ as follows, where the expected utility of h_i is scaled by the remaining hypotheses in the space:

$$z(h_i) = \frac{EU(h_i)}{\sum_{h_j \in \mathcal{H}} EU(h_j)} \tag{3.5}$$

I provide quantitative details for how relative utilities can be calculated in the appendix.

In its current form, the framework shows how utility information can be incorporated into the selection of hypotheses. When a hypothesis is selected given this process, it's called a second-order belief. We can interpret this process as a second-order sampling of a posterior distribution that is scaled by a utility calculus. Utility information can lead to divergences in credences and belief reports (when conditioned on identical data) in one of two ways: (1) When reasoners set different utility values to different outcomes and (2) when reasoners infer different likelihoods of obtaining consequences.

Consequently, the hypothesis $\hat{h}$ (the sampled second-order belief) is that which maximizes the expected relative utility calculation in Equation 3.5:

$$\hat{h} = \arg\max_{h_i \in \mathcal{H}} z(h_i) \tag{3.6}$$

Here, arg max simply means return the hypothesis that maximizes the relative utility function z. The expected utility of a hypothesis is by definition a function of the likelihood of that hypothesis being true, scaled by the utility of its outcomes being true. Therefore, as Equation 3.6 states, the hypothesis with the largest relative expected utility is the one that maximizes these two constraints (i.e, the likelihood of the hypothesis scaled by its utility). To be precise, the actual belief a reasoner reports is the degree of belief (subjective probability of truth) they assign to the maximizing hypothesis $\hat{h}$ being true, $\mathbb{P}(\hat{h})$.

Incorporating Bayesian inference when computing second-order beliefs

By integrating Bayesian updating in our expected utility framework, we get a computational model of Bayesian decision making that I use to construct a practically rational reasoner, with the three factors stated above (i.e., the likelihood, the prior, and the utility):

$$EU(h_i|d) = \mathbb{P}(h_i|d)\,\mathbb{U}(h_i) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{3.7}$$

Equation 3.7 is an extension of Equation 3.4 in which $\mathbb{P}(h_i)$ is replaced with the posterior of h_i given d. Therefore, the expected utility of a hypothesis is updated via Bayes' Theorem to incorporate learning new data. I describe the computations of this process in more detail in the appendix.

People will report the belief which maximizes utility in this equation, given a utility calculation and an integration of data where the reasoner reports the belief $\hat{h}$ given data. We can calculate that value with this

expression:

$$\hat{h} = \arg\max_{h_i \in \mathcal{H}} z(h_i|d) \tag{3.8}$$

This function maximizes the relative utility over the hypothesis space, given observing a new piece of data d. This is how a reported belief is calculated in our framework.

An implication of our proposal is that reasoners weigh the Bayes factor of a hypothesis against the relative utility of an opposing hypothesis. In effect, people weigh the evidence of a hypothesis against its utility. This result has important implications (and makes testable predictions) for how we understand reasoning in everyday situations, for instance, how people form beliefs about climate change, vaccines, and the like. Consider a case where a reasoner has to decide what to believe between two hypotheses h_1 and h_2 given data d and utilities of the consequences of c_{h_1} and c_{h_2}. In this framework, $EU(h_1) = \mathbb{P}(h_1|d)\,\mathbb{U}(c_{h_1})$ and $EU(h_2) = \mathbb{P}(h_2|d)\,\mathbb{U}(c_{h_1})$. And our central claim is that if $EU(h_1) > EU(h_2)$ then the reasoner will decide on h_1; their belief report will be h_1. This is because the reported belief $\hat{h}$ is the hypothesis h_i with the largest value. I discuss this result in the model simulation section.

How do motivation-representations affect beliefs?

As shown in Figure 3.1, a reasoner evaluates evidence for or against their beliefs in generating a belief report. One way directional motivations could impact belief updating is by shaping evidence representations in the $\beta-$Bernoulli model. This is the view tacitly assumed by much of the research on motivated reasoning. In contrast to this view, directional motivations can alter one's decisions *after* integrating the evidence to construct a first-order belief. Motivations shape second-order belief representations after a Bayesian model optimally transforms evidence to a belief. When engaged in motivated reasoning, first-order beliefs encode directional information. In the Bayesian decision model instantiating practically rational reasoning, first-order beliefs need not encode directional information — this is a key difference between the models.

These two paths of influence have different implications from the standpoint of rational updating: Motivated reasoning could be at odds with theoretical rationality because directional utilities could produce a mismatch between the world and people's credences. Alternatively, it may be a belief report is actually a decision that is the result of utility-sensitive Bayesian updating, and this decision is practically rational because it is uncontroversial that *decisions* should take into account both doxastic and utility considerations (Von Neumann & Morgenstern, 2007; Maher, 1993; Wallace, 2020). Both of these cases are realizable in the computational models I described above.

In Table 3.1, I summarize the assumptions and consequences of the Bayesian decision-theoretic framework

instantiating a practically rational reasoner.

Table 3.1: Assumptions and consequences of using the proposed computational framework to distinguish motivated reasoning and practical rationality.

Assumption	**Consequence**
The factors that shape people's beliefs include priors, likelihoods, and utilities.	Practical reasoning models generates belief reports as a function of these three factors.
Utility is computed separately from posteriors given data.	Representations of uncertainty are separated from utility calculations.
Directional motivations can impact evidence representations directly or indirectly.	Directional motivations can violate norms of theoretical reason or conform with norms of practical reason.

Simulations in a toy experiment

How can we distinguish motivated reasoning from practical rationality in an experimental context? One way to measure the impact of directional goals on the sampling of evidence and the construction of a second-order belief is to (1) sequentially present evidence for a hypothesis and (2) ask participants about their memory for the amount of evidence which supports one of two hypotheses, where a distinguishing feature of one hypothesis is it yields a reward. There are established designs in psychophysics where the experimenter systematically varies decision thresholds by assigning different reward-values (i.e., utilities) to the outcomes (e.g., Wickens, 2001). In this design, the experimenter induces a directional goal. A virtue of inducing a directional goal is that participants will not have predetermined expectations about the likelihood of the data under each hypothesis, so the effects of participants' priors will be more easily quantified. That is, inducing a directional goal would make it easier to assume that the prior distribution is approximately flat.

In the design described above, researchers could distinguish practical rationality from motivated reasoning by measuring how much evidence participants represent for a hypothesis and use that value to predict their subsequent belief report. For example, if participants' memories for the evidence is accurate (or not systematically biased towards a directional goal), but their decisions nonetheless indicate they've incorporated the utility of holding the hypotheses with the larger reward, this could indicate they are reasoning in a practically rational way. In Table 3.2 I list a series of features of experimental designs which both capture typical ways people form beliefs about the world and would be necessary to incorporate to distinguish motivated reasoning from practical rationality.

In the remainder of this section, I spell out this proposed toy experiment further. The purpose of this experiment is to highlight how the Bayesian decision-theoretic framework differentiates motivated reasoning

Table 3.2: Key features of belief formation in naturalistic settings.

- People accumulate evidence about a hypothesis before they make a decision. This means one's memory of the evidence can impact what they believe.
- People might perceive evidence as being consistent or inconsistent with what they want to believe. Evidence can also be ambiguous under hypotheses they are considering. Either way, people can make decisions despite their level of uncertainty.
- Individual pieces of evidence are not necessarily definitive. That is, they may support a hypothesis, but rarely outright confirm or reject a hypothesis.
- People do not receive immediate feedback about whether their judgements are correct, but they do anticipate the consequences of assent or dissent towards a hypothesis.
- For many beliefs of interest to people, there is some utility associated with endorsing a hypothesis.
- People's directional goal for an outcome could be relatively weaker or stronger (i.e., the utility associated with a given outcome could vary), which suggests the need to investigate how different rewards impact the accuracy of people's beliefs.

and practical rationality in a constrained setting. Our goal is not to provide empirical evidence that—as a matter of fact—people are *reliably* practically rational rather than engaged in motivated reasoning, or make any other similarly broad claims. This question can only be answered with respect to a specific context once researchers have measured people's priors, their utility functions, quantified the evidence, and compared their behavior against a well-defined mathematical benchmark.

Experimental setup

I built two computational models that simulate people's credences in situations where utilities and evidence can impact their beliefs.[3] The *motivated reasoning* model uses directional priors, which encode utility information for the hypotheses. The *Bayesian decision* framework instantiating a practically rational reasoner incorporates utilities after sampling credence values in an additional belief reporting step based on an expected utility calculation. I run these models through a toy motivated reasoning experiment. I discuss how each model can yield a similar decision even though the underlying representations differ when producing this decision.

As we've discussed, psychologists often measure participants' prior beliefs towards a hypothesis h_i before they're presented with evidence. Alternatively, participants can be presented with base-rate information so the prior distribution is easier to quantify. In the toy experiment we'll describe, base-rate information supports a uniform prior over hypotheses. I sample from a uniform distribution because in this toy experiment, the models have no prior expectations about the distributions of the evidence under each hypothesis. Our

[3] I implement the models in Python, and host the code as Google Colab notebook at https://colab.research.google.com/drive/1BXdKnfZCKCUxFA9UR77oSXkfxmMaOuvx.

models are then presented with a series of facts (or evidence), some of which support a hypothesis h_i and some of which support an alternative hypothesis. I treat the number of facts (i.e., evidence count) supporting h_i as a discrete random variable, ev_{world}, sampled uniformly from values ranging from zero to six. I represent this mathematically as $ev_{world} \sim unif[0:6]$.

In studies of motivated reasoning, participants are motivated to believe one hypothesis over another. For example, they are motivated to believe that their political party has the right economic policy (Caddick & Rottman, 2021). I quantify this motivation using a utility function, which maps world states (e.g., the actuality of a hypothesis h_i) to utility values. I begin by setting the utility of h_i being true equal to three, $\mathbb{U}(h_i) = 3$, and the utility of $\neg h_i$ equal to one, $\mathbb{U}(\neg h_i) = 1$. The numbers themselves are arbitrary; the point is only that the perceived utility of one hypothesis is higher than the utility of its negation. Consequently, in the present toy experiment, motivation is manipulated independently of the prior distribution – a feature that is atypical of most experiments examining motivated reasoning and polarization (e.g. Lord et al., 1979; Nyhan et al., 2014a; Ditto & Lopez, 1992), but a feature which enables us to quantify the unique impact of both the prior distribution and the utility function.

Assume a reasoner is deciding between reporting they believe a hypothesis H or not H (denoted as $\neg H$). Both hypotheses have some probability of being correct. If the reasoner reports the correct hypothesis, they get a reward. The probability and reward structure for these hypotheses follows:

$$\mathcal{H} : (p_H = 0.5, u_H = 3)$$
$$\neg\mathcal{H} : (p_{\neg H} = 0.5, u_{\neg H} = 1)$$

Assume the reasoner encounters a discrete count of evidence d_i for $\mathcal{H}$ being true. What should their posterior credence in $\mathcal{H}$ be? Rational choice theory says agents should maximize their expected utility, but what are the agent's internal credences? The two accounts I discuss differ in how d_i impacts credence in $\mathcal{H}$, and provides competing answers. The Bayesian decision framework instantiating a practically rational reasoner incorporates utilities after first-order beliefs are constructed in a second-order inference step. For this example, I define the Bayesian decision prior in $\mathcal{H}$ as the normative prior, or simply, p_H. This is a prior that accurately tracks base-rate information provided in an experiment. In the motivated reasoning model, priors are affected by the perceived utility of a hypothesis.

There are many ways to model how utilities could influence the prior distribution, but I will consider the simplest possible case, where $\mathbb{P}^*$ equals the expected relative utility, the normative prior is uniform, and utilities are discrete quantities:

$$\mathbb{P}^*(\mathcal{H}) = \frac{p_H u_H}{p_H u_H + p_{\neg H} u_{\neg H}} = \frac{0.5 \times 3}{0.5 \times 3 + 0.5 \times 1} = \frac{3}{4} \tag{3.9}$$

The directional prior in $\neg\mathcal{H}$ is simply $1-\mathbb{P}(\mathcal{H})$, or its relative utility $z(\neg\mathcal{H})$. In this simple case, increasing the utility of reporting $\mathcal{H}$ linearly updates – or directs – the prior credence in H.

The expectation of $\beta(a, b)$ is $a/(a+b)$. Consequently, setting $a_{prior} = u_H$ and $b_{prior} = u_{\neg H}$ allows us to define a probability distribution over credences with the desired expected value. This modeling choice allows us to not only include uncertainty when sampling credence, but also demonstrates a path for developing statistical models that can be fit to behavioral data. Directly equating a and b to utilities is only possible when utilities are integers (the β-distribution takes integers as inputs) and when the normative prior is uniform. I will not cover here how one might solve systems of equations for finding a and b values for other utilities and priors.

Although this model may be limited, it serves as a useful illustrative example. To help gain an intuition for why instantiating a directional prior as a β-distribution may make sense, consider the shape of the distribution for different combinations of a and b. When $a = b = 1$, the utilities are equal, and the β is uniformly distributed. The expected value of the distribution is $\frac{1}{2}$, which is the relative utility when both hypotheses have utility equal to one. When $a < b$, $u_H > u_{\neg H}$, and credences favoring $\mathcal{H}$ are more probable. Conversely, when $a > b$, $u_H < u_{\neg H}$, and credences favoring $\neg\mathcal{H}$ are more probable. In the Bayesian decision model $a_{prior} = b_{prior} = 1$.

Because the β distribution is conjugate to the Binomial distribution (this distribution encodes the number of facts, or evidence, in favor of a hypothesis), we can update beliefs using Bayes' rule analytically given a discrete count of evidence i (out of n trials) for $\mathcal{H}$. By updating the parameters of the β distribution to $a_{posterior} = i$ and $b_{posterior} = n - i + 1$, we can then sample a posterior credence (first-order belief).

Results from toy experiment

In Figure 3.2 and Figure 3.3, I simulate predictions for the motivated reasoning model and the Bayesian decision framework instantiating a practically rational reasoner. These simulations show that the motivated reasoning model has a boost in credence for $\mathcal{H}$ for each evidence count. This is the impact of directional priors on sampling beliefs in favor of a higher-utility hypothesis. In contrast, the practically rational reasoner has a normative credence function because utilities affect second-order beliefs (i.e., after a first-order credence is constructed). Stated another way, utilities do not impact credences directly for a practically rational reasoner. Divergences from this line shows how much a utility impacts a credence at a given evidence count.

In behavioral tasks, psychologists measure second-order belief reports, which represent a combination of both a credence and a decision to assent to a hypothesis (Maher, 1993). The Bayesian decision-theoretic framework instantiating a practically rational reasoner describes how internal credences produce belief re-

ports. In Figure 3.3, I plot the probability of deciding on H (the higher utility hypothesis), which shows that the utility of the hypothesis affects the model's decision to choose it. Even though the probability of choosing a hypothesis varies as a function of utility, as shown in Figure 3.2, the internal credence—the posterior probability representing the model's first-order belief—is unaffected. These simulations show that a practically rational reasoner can generate credences which are theoretically rational while executing decisions that can, on the surface, look like prototypical instances of motivated reasoning.

Thus, I see that the model's credences can diverge from its belief reports, a situation which can be practically rational when the evidence for the hypotheses are uncertain, but a reasoner needs to make a decision (Maher, 1993).

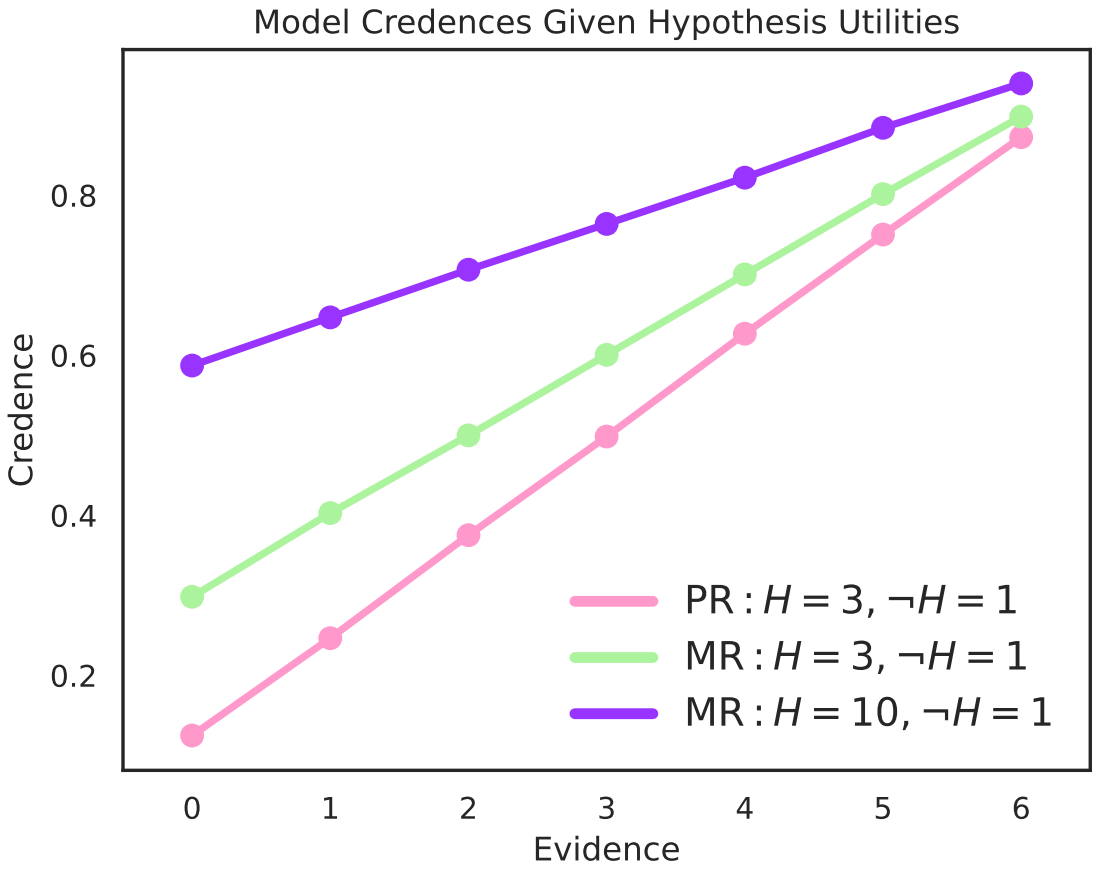

Figure 3.2: Credences sampled from three models. Two motivated reasoning (MR) models and one practical reasoning (PR) model. The motivated reasoning models sample credences in favor of higher utility hypotheses when compared to the (normative) practical reasoning model. In the motivated reasoning model, as the $u(\mathcal{H})$ grows, so does its credence.

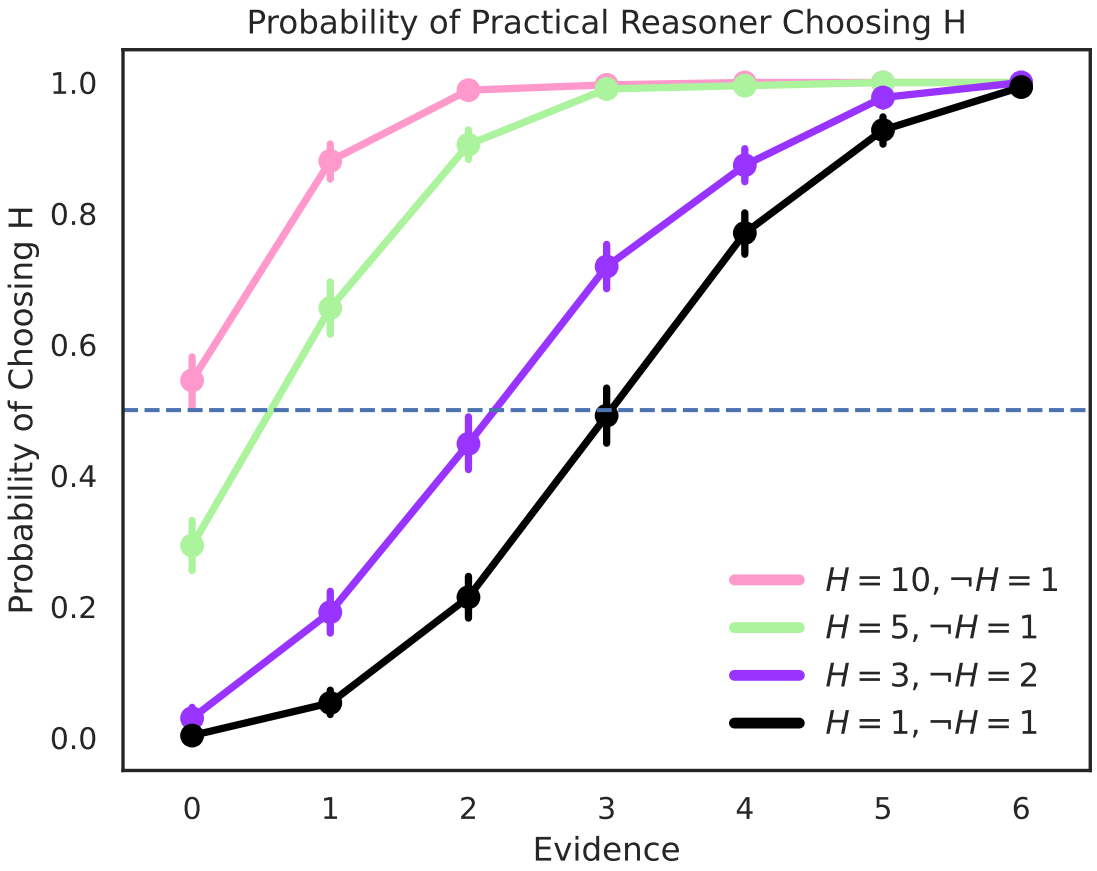

Figure 3.3: The probability the practical reasoning model chooses the higher-valued hypothesis given a utility and evidence threshold. When the evidence is uncertain and utilities are equal (black line), a practical reasoner will choose hypotheses at random. However, when one hypothesis has higher utility (purple, green, and pink lines), the model selects the higher-valued hypothesis with a probability proportional to its relative utility.

Evaluating the measurement and interpretation of data in prior empirical research on motivated reasoning

What are the concrete implications of this model for understanding existing research on motivated reasoning? One tacit claim in the enormous and varied literature on motivated reasoning is that motivations exert direct effects on people's ability to represent information for or against their beliefs, and as a consequence, people fail to update their beliefs as they should (e.g., Kunda, 1990; Hart & Nisbet, 2012; Little, 2021; Taber & Lodge, 2006). It is worth noting that much of the research on motivated reasoning has been conducted in social psychology, which primarily relies on verbal theories which examine the relation between the inputs (data) and outputs (belief reports), while treating the processing that links the two as a difficult-to-quantify black box (but see Jansen et al., 2021; Cook & Lewandowsky, 2016; Little, 2021; Jern et al., 2014; Austerweil & Griffiths, 2011). These links are sometimes instantiated in mediation or structural equation models, and researchers make informed conjectures on the basis of these models, but it is fair to say the conjectures rarely rise to the level of making predictions that could be compared against a computational model. We can use our framework to discuss how their inferences about the effects of motivation on belief updating have not been adequately tested, in part, because what has been measured in their experimental designs precludes the possibility of doing so (Little, 2021). I will consider a series of experiments where participants were presented a fictitious health outcome that had positive or negative health implications depending on the experimental manipulation (Ditto & Lopez, 1992; Ditto et al., 1998, 2003). Our point here is not to criticize these experiments specifically, but rather to make a general point about measurement in much of the research on motivated reasoning.

Ditto and colleagues (1992; 1998; 2003) told participants they were testing their saliva for a TAA enzyme, which was supposedly linked to favorable or unfavorable health outcomes. In the TAA-negative condition, participants were told that having this enzyme would make it less likely for them to develop pancreatitis (a favorable health outcome), whereas in the TAA-positive condition participants were told they were more susceptible to pancreatitis (an unfavorable health outcome). The experimental procedure included asking participants to put their saliva on a color-changing strip of paper which tested for the presence of a TAA deficiency.

Participants received base-rate information on the prevalence of TAA in the general population. Participants were informed of the base-rate information either before or after they self-administered the enzyme test. In the high probability condition, 1 in 3 (33%) people had the TAA enzyme whereas in the low probability condition, 1 in 20 (5%) had the enzyme. The authors found that the base-rate information (which they interpreted as the prior probability) impacted participants' beliefs in expected directions — those in

the high probability condition reported greater likelihood of having the enzyme and vice versa. However, participants who believed TAA enzymes to be deleterious to their health reported that it was less likely they had the enzyme than participants who believed TAA enzymes would lead to favorable health outcomes. This finding suggested the impact of directional motivation on people's beliefs because participants interpreted the outcome of the TAA-deficiency test to maintain the belief that they are, in fact, healthy.

How can we understand the authors' interpretations of these results as evidence of motivated reasoning, and are there plausible alternative explanations under our modeling framework? First, although participants were provided with base-rate information about the TAA enzyme, the sample was primarily undergraduate students. While this is not a unique problem to this series of studies, it presents a unique problem in the current experiment. For example, undergraduate students are mostly healthy and thus might correctly balk at the purported evidence that is inconsistent with what they know about their health. Namely, they are young and thus very likely to be healthy. Consequently, it would be surprising to find out they were secretly unhealthy. In effect, although base-rate information has been communicated to participants, this data is not the *entirety* of the prior which would inform how one would expect participants to interpret data they are presented with. A vast web of auxiliary beliefs compose people's prior distributions. These auxiliary beliefs can rationally produce seemingly biased rejections or reinterpretations of information (Gershman, 2018; Jern et al., 2014). Failing to measure beliefs about causally related hypotheses can make it appear that participants are engaging in motivated reasoning when they are not (D. Powell et al., 2022). The authors would also need to measure the participants' perceptions of their own health, in addition with the base-rate information, to quantify the extent to which the participants' posterior reflects the evidence they are provided, base-rate information, and the more complete prior distribution.

Just as more complete prior information was unmeasured in this original study, utilities may influence how participants select between hypotheses in this experiment as well. A feature of our framework is that it quantifies the possible consequences of adopting a belief and highlights if a failure to update is a result of goal-based considerations. For simplicity, I will assume that the amount of evidence provided by the test itself is quantifiable, and so the remaining open question is how the *utility* of holding a certain belief could alter the interpretation of the results of the studies.

It is clear in these experiments that the utility functions over being in good health or possibly developing pancreatitis have gone unmeasured. This fact alone makes it difficult to determine the extent to which the utility has been inappropriately incorporated into the decision to discount the test result. Belief reports are decisions. To evaluate these decisions, we need to compare participants' perceived utilities of the outcomes scaled by the posterior distribution against a mathematical benchmark. Can we infer these utilities? People differ in the utility they ascribe to being healthy: Some people smoke and some do not, despite the evidence

that smoking is a strong predictor of poor health outcomes. Some people exercise and eat nutritious food, and some do not. The point is the utilities about *how healthy one cares to be* will vary, which reflects the utility function over that outcome. Under our framework, a reasoner's goals must be measured to understand how and whether utilities are changing the way evidence is sampled or whether it is rationally integrated into a belief report.

Let us suppose further that participants' prior beliefs were more completely measured, and participants' utilities over health outcomes were also measured. How would one establish that participants were engaging in motivated reasoning, or alternatively, reasoning practically rationally?

We need to disentangle the impact of health goals on the integration of evidence. Referring back to Figure 3.1, practically rational behavior integrates utility information *after* evidence is sampled into a first-order belief. Consequently, utility-based considerations should *not* be measured by asking participants if they think they have the disease, but rather be measured by assessing their interpretation of the evidential impact of the test results. For example: 'Given this test result, do you think there is now more evidence that you have this disease?' If participants were engaged in motivated reasoning, the motivation to believe in positive health outcomes would asymmetrically change how they sample from the evidence – thus, they would respond to this question by either ignoring evidence of poor health or otherwise fail to update their representations of the evidence appropriately. In contrast, a practically rational reasoner could interpret the evidence in a manner that aligns more closely with the test results—regardless of the valence of the evidence—and this representation need not line up with their belief report. For example, it would be perfectly consistent for them to think that there is additional evidence they have the disease (credence) without believing they have the disease (belief report).

Now we can see how the utility of forming a belief could produce a practically rational belief report which diverges from one's representation of the evidence and how that would differ from motivated reasoning. We can imagine that practically rational reasoners could ascribe negative utility to learning they are unhealthy (e.g., realizing a consequence of thinking you are unhealthy suggests that you might need to act to fix this issue). This utility could impact them in two ways. First, it could impact how they sample from the evidence; a directional utility would impact their representation of the evidence — this is motivated reasoning. Alternatively, their new representation of the evidence could reflect the information they received (as indicated by a question about the evidence rather than only their belief report). And even if this led them to be *uncertain* about whether the evidence indicated they were unhealthy, they could still generate a belief report that they are healthy. This is because they might assign a negative utility to forming the belief they are unhealthy, and given that the evidence is uncertain, they could decide to believe they are healthy (Maher, 1993). Referring back to Figure 3.3, the key implication of our framework is that when the available

evidence is uncertain (or equally points to multiple hypothesis), and people consider the relative utility of different outcomes, people will choose to believe the hypothesis with the most positive outcome – this is practically rational behavior. If participants were engaged in motivated reasoning, their interpretation of the evidence would not include the information they received and likewise their belief report, and representation of the evidence, would reflect this fact.

Our aim here is not to single out any particular set of research, but rather demonstrate how measurement of specific information is necessary to understand the underlying mechanisms producing a belief report. We've argued psychologists need to measure people's prior beliefs, how people weigh the evidence presented (and this needs to be quantified), and the utility they assign to holding certain beliefs. It is likely people engage in motivated reasoning that lead them to sample relevant evidence inappropriately, and the above studies may *suggest* this is true as well. However, to more firmly establish this claim, researchers need to measure and precisely quantify factors to distinguish cases of motivated reasoning, cases of ordinary Bayesian updating, and cases of practical rationality.

Discussion

I have described a computational modeling framework grounded in Bayesian decision theory for distinguishing cases of practical rationality and motivated reasoning. This framework focuses on how to measure the impact of priors, evidence, and utility on credences and belief reports. I report simulation results from a toy experiment to demonstrate how our framework can highlight how first-order information and second-order utilities can differentially impact people's posteriors. This framework highlights several key features of belief formation in naturalistic settings and makes both key predictions and raises important questions for future research on motivated reasoning and practical rationality.

Applied implications of distinguishing motivated reasoning and practically rationality

Recasting existing psychological measures as measures of utility

I have argued that psychologists need to explicitly measure the perceived utility of accuracy and directionality to distinguish between motivated reasoning and practical rationality. Existing psychological constructs within the motivated reasoning literature could be measures of utility. For example, psychologists have measured participants' Need for Cognition in studies of motivated reasoning (Caddick & Feist, 2021; Arceneaux & Vander Wielen, 2013; Caddick, 2016; Nir, 2011), which one could recast within our framework as the utility

of forming accurate beliefs by engaging in effortful cognitive tasks (Petty et al., 2009; Cacioppo & Petty, 1982). Similarly, Need for Closure and Personal Need for Structure scales (Neuberg et al., 1997; Webster & Kruglanski, 1994; Moskowitz, 1993; Neuberg & Newsom, 1993; Sinatra et al., 2014; Kundra & Sinclair, 1999; Ask & Granhag, 2005; Dijksterhuis et al., 1996; Kruglanski et al., 2006) could be recast as measures of the directional utility of getting quick answers. Thus, researchers interested in applying our framework may be able to use existing, validated measures to measure utility functions which can impact second-order belief reports.

Using existing research to meta-analyse measurement of previous reports of motivated reasoning

Social psychologists often rely on verbal theories and fail to explicitly compare their predictions against a computational model. To disentangle cases of motivated reasoning from practical rationality, one must measure people's priors, evidence, and utility associated with directional and accuracy goals. A measurement meta-analysis of past research on motivated reasoning work could highlight what past research could plausibly tell us the extent and context in which people engage in motivated reasoning, as we've defined it. A good starting point would be to operationalize each of these constructs: For example, a prior might be defined as pretest beliefs and the likelihood could be quantified as numeric weights, such as statistical data or percentages assigned to each piece of information presented within the experiment. Utility associated with accuracy and directional goals could be conceptualized (Kunda, 1990) and measured (Luce, 1991; Hershey & Schoemaker, 1985) as in prior research in psychology and behavioral economics. A detailed review would help assess whether the current literature can test differences between practical rationality and motivated reasoning.

Using computational modeling to develop second-order normative interventions

Even when a belief report may be practically rational, it does not guarantee that the belief or behavior does not require intervention. Different intervention tactics are required to overcome these goals. For example, *confirmation bias* leads people to accept information that confirms their beliefs while inappropriately rejecting disconfirming information (Klayman, 1995; Nickerson, 1998). Alternatively, a belief-behavior mismatch occurs when behaviors don't align with internal beliefs (Tesser, 1992; Tumulty, 2014). Someone may demonstrate a belief-behavior mismatch by vocalizing support for policies that combat climate change, while excessively eating foods that are energy- and water-intensive to produce. These two cases may well require different interventions, even if they are the product of theoretically or practically rational Bayesian updating.

Psychologists developing interventions often only measure first-order credences, focusing on how goal-based motivation biases sampling during the construction of an initial credence. However, in cases where first-order interventions fail to shift people's beliefs, it is still possible to shift people's beliefs by focusing on how second-order utility impact belief reports (J. H. Priniski & Horne, 2019; D. Kahan & Braman, 2006; D. Kahan et al., 2017). The modeling framework we've detailed can distinguish between first-order and second-order interventions, where the former aims to revise priors, evidence, and thus credences and the latter aims to shift the utility people anticipate of forming a belief. For example, a nudge that encourages participants to encode information more accurately could be appropriate when the source of problematic, though rational behavior is the result of evidence (Pennycook & Rand, 2019; Bago et al., 2020). In contrast, second-order intervention on anticipated utility, such as how teaching the real-world application of some concepts has been shown to drive learning (Soicher & Becker-Blease, 2020; Hulleman & Harackiewicz, 2009), may be called for when the evidence is thought to be uncertain but people have incomplete information about the utility of forming a given belief. These are distinct situations that implicate different sets of cognitive processes, which may demand altogether different kinds of intervention – some based on defects in one's inferential machinery (first-order interventions) and others due to influences of directional goals (second-order interventions).

Theoretical implications of distinguishing motivated reasoning and practical rationality

Weighing accuracy and directional utility

When the evidence is logically compatible with different hypotheses, the acceptence or rejection of a hypothesis could result from an expected utility calculation (Maher, 1993). However, as Kunda (1990) argues, the motivations in question can take a variety of forms, and may not always lead a reasoner to form inaccurate beliefs. In situations where it's practically rational for directional goals to influence one's belief reports, how do people assign and weigh utilities? For example, in Ditto and colleagues' work (1992; 1998; 2003), participants have a directional goal of believing they are healthy despite getting information which suggests otherwise. In some senses, it would be practically rational for participants' utilities to encode accurate, rather than directional beliefs about their own health (though, as I discussed at the end of the previous section, people also think there is utility in believing one is healthy even when the evidence is uncertain). This raises the question of *how much* utility people ascribe to forming true beliefs and how people weigh this calculation against directional goals. Practically rational behavior could involve a utility function over accuracy *and* directionality, and these would both need to be measured to determine how participants behave

relative to a well-defined mathematical benchmark.

One possibility is that once I take the utility of directional *and* accuracy goals into account, it still follows that people's belief reports should be driven by accuracy goals rather than directional goals – it will depend on the utility people assign to both of these types of goals. It is an empirical question but we'd conjecture for many beliefs, participants are not particularly worried about being inaccurate because they do not care about the truth of a given proposition per se, for example, about some highly specific carbon pricing policy (Falk & Zimmermann, 2016; D. Kahan & Braman, 2006). Instead, participants are invested in the direction of a hypothesis insofar as they perceive the consequences of forming that belief for their future behavior or recognize how holding that belief would change how people perceive them (D. Kahan & Braman, 2006). For instance, their accuracy about highly specific carbon pricing policy is less important to them than the perceived cost of endorsing this belief for their future behavior.

People also have strong convictions; they have beliefs in which they *are* invested (Skitka, 2010). For example, people have strong views about many general claims, like the claim human-caused climate change is real. However, many of their beliefs are unlikely to take this form (D. Kahan & Braman, 2006); this is speculative because it's gone unmeasured to-date, but these situations are likely the norm rather than the exception. Given these considerations, future work must attempt to quantify utilities assigned to different goal states because that will determine a person-specific mathematical benchmark to compare their belief reports and behavior against.

Credences versus belief reports

A fundamental assumption in our discussion of first and second-order beliefs concerns the possibility that people's internal credences and belief reports can diverge. What is the empirical evidence for this claim? In fact, there is rich literature which is specifically focused on these and related questions (Orne, 2017). For example, the creation of one of the most famous paradigms in psychology—the Implicit Associations Test—rests on the idea that internal credences about topics like racism, sexism, and the like can diverge from people's belief reports. It is a secondary question whether tests like the Implicit Associations Test accurately capture a meaningful construct. But, the fact that these and other more implicit measures exist suggests that psychologists are well-aware first-order and second-order beliefs could diverge for a host of reasons (e.g., Greenwald et al., 2020; Orne, 2017). To consider another example, it is well-known that political polling is not so straightforward that pollsters can simply ask who one plans to vote for without considering any other factors; people opt out of studies when they are asked about certain topics (non-response bias; Groves & Peytcheva, 2008), or might withhold stating their attitude about a polarizing topic for fear of the consequences of taking a stance (e.g., Harrison & Startin, 2013). This doesn't mean that second-order

beliefs are always strategic, deceptive, or insincere. They could result from a deliberative process of weighing their certainty about the evidence, their investment in the truth of the proposition, and the perceived utility of forming the belief, and these computations could happen either explicitly or implicitly. For instance, second-order beliefs could result from metacognitive readjustments of a first-order belief constructed using lower-level perceptual and evidence quantification processes (e.g., Maniscalco & Lau, 2012). In this sense, a second-order belief could be explicitly held while a first-order belief may not be.

The current situation we've outlined might lead one to think that I am skeptical credences could *ever* be measured in a way to distinguish motivated reasoning from practically rational behavior. This is not a claim we're committed to. Rather, in many cases belief reports surely reflect people's credences, particularly in cases where people perceive no consequence in forming a belief one way or another. For example, suppose that a participant is asked whether they think the distance between New York City and Los Angeles is greater than 3,000 miles. We'd expect whatever the report here to directly reflect their credence. However, the situation is altogether different for many of the typical topics social psychologists have aimed to study. Cases where social psychologists test for motivated reasoning are often the very cases where belief reports are *least likely* to reflect credences directly because participants can anticipate the consequences of forming a belief about that topic (Falk & Zimmermann, 2016); the topic matters to them and so they consider the consequences in their deliberation. This means that researchers must go beyond measuring belief reports simpliciter. Researchers must measure the utility participants assign to holding some beliefs, and participants' understanding of the evidence for those beliefs.

How coherence mechanisms turn belief reports into evidence

Belief reports are a decision to report a belief in a way that takes into account of both posterior probability of a hypothesis and the utility of forming and reporting that hypothesis. This suggests that a belief report itself may not be a stable attitude. Instead, it is the product of perception of the environment at a particular time and context – so why consider it a belief of any stripe? One question then is whether second-order belief reports are beliefs, or if they could ever produce stable attitudes rather than strategic or insincere responses.

Two related questions are (1) if and how a belief report produces a stable credence or attitude and (2) whether initially practically rational behavior could lead to a stable credence at odds with the evidence. First, it's both likely belief reports can produce stable credences and there is clear evidence of how this could happen. Researchers have demonstrated that once people make a decision, they subsequently adjust their beliefs to cohere with the decision they've made (e.g., Simon et al., 2001, 2015). For example, when people choose one option over another, their evaluation of each option's characteristics shift to cohere with their

choice in that the characteristics of the selected option are more positively evaluated, and the characteristics of the non-selected option are more negatively evaluated (Carlson & Russo, 2001; Holyoak & Simon, 1999; Izuma et al., 2010; D. Lee & Coricelli, 2020; D. Lee & Daunizeau, 2020; D. G. Lee & Holyoak, 2021; Enisman et al., 2021). While choice-induced coherence shifts were first assumed to result from a post-hoc reduction in dissonance, where the reasoner reevaluates their beliefs about the characteristics *after* making a choice (Festinger, 1957; Harmon-Jones & Mills, 2019), recent evidence suggests that these shifts my occur before and during the deliberative processes, and constitute an essential step for effectively selecting a choice (D. G. Lee & Holyoak, 2023, 2021; Voigt et al., 2019). There are complex mechanisms which link first-order and second-order beliefs but one can simply imagine how this would work. Suppose someone reports believing that climate change is real even when they think the evidence is uncertain. Later, they think about how they reporting believing this, but may not remember the cumulative reasons why they did. The mere fact that they reported however is evidence that they believe it *and* that the evidence supported them believing it. In this way, a belief report could be construed as evidence that entails them forming a first-order credence – the evidence is no longer perceived to be uncertain.

The mechanism we've hypothesized above then allows us to address how practically rational belief reports could produce credences at odds with the evidence without any new, external data. To return to our prior example, one can imagine that while a reasoner may initially be well-calibrated to the evidence, their decision to report a belief based on the utility of believing it could shift their representations of the evidence. In this way, initially practical rational behavior could produce internal credences which no longer reflect facts about the world, even by their own lights.

However, the idea that people's decisions can impact their stable credences could have beneficial consequences as well. For example, imagine we've implemented a utility focused intervention to shift climate change behaviors. If people's credences are shifted because they aim to maintain coherence between their behavior and their beliefs, then a utility intervention could eventually lead people to form stable beliefs that accurately reflect facts about the world (Jachimowicz et al., 2018; Simon et al., 2015).

Utility and rationality

How do people form utility representations that impact their belief reports? Using a common assumption from the reinforcement learning literature, where values of world-states are often assumed and used as input for an agent, I proposed that a predefined utility function (e.g., a specified mapping of states to reward values) guides reasoning. Still, our framework does not discuss how utility representations are computed in the first place (for a discussion of related issues, see Bostrom, 2009). It is possible that directional goals influence utility representations just as they can impact how evidence is sampled, arguably making the effect

of utility on credences an instance of motivated reasoning.

Along similar lines, practical *irrationality* is a failure to execute an action plan consistent with one's utility function and knowledge of the world. In contrast, theoretical irrationality is a mismatch between one's beliefs and the available evidence describing the world really is. When can we conclude a *belief report* is practically irrational? People might have utilities that are incorrectly joined with information they believe or have utility functions which deviate from rational-actor models (Loewenstein & Molnar, 2018; Tversky & Kahneman, 1991). They could act in ways which are incompatible with the utilities themselves (like when I plan to exercise but fail to) or fail to update their utilities given new information (Edwards, 1954; Kalis et al., 2008; Wiggins, 1978; Stetzka & Winter, 2021). All of these possibilities can be accounted for under our framework, distinguishing it from revisionist accounts of theoretical rationality (like the idea of resource rationality proposed by Lieder & Griffiths, 2020). Our framework takes an off-the-shelf definition of theoretical rationality because the question of whether to revise this conception of rationality is not the focus of the paper. But our framework is able to incorporate revisionist accounts of theoretical rationality.

Using results from game theory to understand second-order belief reports

People may report a second-order belief *because* it is perceived to be widely-held by in-group members (e.g., families, political or religious groups, nationality). Reporting beliefs that align with one's group requires coordinating with others (Kashima et al., 2021). What mechanisms underlie coordination and how may their operation guide second-order belief revision? Game-theoretic models extend rational choice theories to study interactive decision-making, where optimal decision strategies hinge not only on one's own understanding of the world, but also on one's beliefs about other people's beliefs and desires (Colman, 2003). For example, imagine two reasoners aiming to coordinate on reporting a specific belief. If the belief is understood to signal group ties, and both reasoners perceive their partner as an in-group member, then they stand to receive a greater reward if they coordinate than if they don't. Consequently, one can ask the question, what probability and payoff matrices make people switch between reporting different beliefs, and what type of information may shift these values to incentivize people to report more accurate beliefs? The expected value of a decision depends on the inferred probability of another making a certain decision (e.g., reporting H) in addition to the utility of that decision (e.g., how much reward will they get if they were to report H). This gives a modeler two representations to manipulate: a probability matrix about how another reasoner will act given a current state and a utility matrix encoding rewards for different decisions. Using a game-theory approach may guide better how social sampling and utility calculations shape second-order belief reporting (e.g., Brown et al., 2022). For example, mathematical models of the stag game can describe how social structures (Skyrms, 2004; Bai et al., 2022) and norms (Bicchieri, 2005) emerge, as well as how people

infer another person's motivations and belief-states during strategic interaction (van Baar et al., 2022). Rational decision-making in these games can have negative consequences (Hahn, 2022). Future work should explore ways to merge game theory with our framework to understand the interplay between the top-down (coordination) and bottom-up (evidence integration) processes guiding second-order belief reporting.

Conclusion

The present research provides a computational framework for distinguishing motivated reasoning and practical rationality. I first distinguished theoretical and practical rationality, with the aim of sharpening how motivated reasoning is discussed and ground this discussion in well-understood distinctions in epistemology. I then developed a framework for evaluating motivated reasoning and practical rationality in a Bayesian decision framework with a toy experiment. The model simulations highlighted several key features of experimental designs for distinguishing motivated reasoning from practical rationality. The payoff of these developments are many, including our ability to reevaluate where motivated reasoning has occurred, the sources of rational, but still problematic beliefs, the relationship between these sources and intervention development, and the cognitive mechanics of how belief reports and credences interact. Altogether, this work highlights a new perspective on motivated reasoning, one which distinguishes it from practical rationality.

Chapter 4

Correcting misconceptions by crowdsourcing educational information from online networks

Introduction

It can be difficult to find common ground with people we disagree with. People's beliefs about polarizing issues are often deeply entrenched and evidence that counters these beliefs generally does not lead people to change their minds. This intransigence comes at a cost: Polarization is a growing problem in the United States and widespread misinformation and misconceptions about, for example, climate change only exacerbate polarization, posing considerable challenges to society. What's more, even in situations when very few people hold a misinformed belief——such as believing that vaccines cause autism——the consequences can still have a widespread negative effect in society; this is evident from the recent resurgence of measles borne from parents refusing to vaccinate their children, citing fears that vaccines cause autism (Jain et al., 2015).

To effectively educate the public, researchers have attempted to confront belief polarization and resistance to evidence by experimentally testing whether educational interventions can induce rational belief updating (Horne et al., 2015; Lai et al., 2014; Nyhan & Reifler, 2010; Nyhan et al., 2014b; Turetsky & Sanderson, 2017). Ideally, people would always properly update their beliefs in accordance with the evidence. However, many interventions developed by scientists are ineffective (e.g., Nyhan et al., 2014a), leading researchers to conclude that people cannot change their beliefs about issues such as climate change, vaccination, or

immigration.

There are several psychological explanations that might explain why educational interventions are often ineffective. First, people interpret evidence to confirm their previously-held beliefs (Klayman, 1995; Nickerson, 1998), and our strongly-held beliefs——such as political and moral beliefs——are deeply rooted in our views of ourselves (Strohminger & Nichols, 2014; Carney et al., 2008), and thus are particularly resistant to change (D. M. Kahan et al., 2012). Second, even when people assimilate evidence, they do so imperfectly, requiring much more evidence than seems epistemically warranted (e.g., J. H. Priniski & Horne, 2018). Even massive education campaigns seem to yield only minor changes in public opinion and behavior (e.g., Fiore et al., 1990). Together, these results have led many researchers to either conclude that meaningful belief change is, in a practical sense, infeasible or that something other than education and evidence is needed to overcome strongly-held beliefs.

However, when an educational intervention fails to change people's misconceptions, this does not entail that other educational interventions (even similar interventions) would fail as well. It is an empirical question whether an untested intervention would turn out to be efficacious. Indeed, researchers have successfully developed effective educational interventions. For instance, Lewandowsky, Gignac, & Vaughan (2013) found that making people aware of the scientific consensus surrounding climate change using icon arrays positively affected people's beliefs. More recently, researchers have found that educational interventions can change vaccine intentions (Horne et al., 2015), correct mental health misperceptions (Turetsky & Sanderson, 2017), and address implicit racial biases, though these changes may be transient (Lai et al., 2014). However, beyond combing the academic literature, researchers have little to go on in predicting whether a given untested intervention will succeed or fail. Moreover, educational interventions are rarely tested outside of the lab, which allows for the possibility that effective educational interventions developed in the lab will fail to generalize beyond tightly controlled settings (J. H. Priniski & Horne, 2018, 2019; J. H. Priniski & Holyoak, 2020). To complicate matters further, the hypothesis space of possible interventions is very large (read, infinite). Consequently, it is not feasible for any given lab or even a group of labs to systematically explore the entire hypothesis space of educational interventions to determine whether a possible intervention could change people's beliefs about a given topic. A methodological advance is needed to avoid a protracted search through the intervention hypothesis space

We propose a two-part method for developing educational interventions. First, use Natural Language Processing to determine which factors change beliefs in naturalistic domains. Second, use successful persuasive arguments culled from online discussions (for example, from the Reddit forum Change My View) and test their generalizability using random samples in controlled labratory conditions. We propose that developing interventions based on existing arguments that have proven to be effective in naturalistic environments

Table 4.1: Example discussion topics and argument responses on Change My View.

Discussion topic	**Example response to original post**
Unpaid internships should be illegal	It's sad that unpaid internships have the effect of freezing out talented people who can't afford a few months' living expenses without generating income. However, they play an important role in any developed economy...
There is no moral justification for eating meat in a first world country	It really depends on what your morals are. If one simply does not see it as being immoral to kill an animal for its meat, then killing the animal does not conflict with that individual's morality. Unless we were to believe in some objective sense of morality, in which case a lot of animals would be considered immoral for killing other animals to eat...

provides a compelling starting place for the development of effective educational interventions.

Reddit's Change My View

Change My View is a popular Reddit forum where users post their views on issues ranging from gun control to opinions about movies. Redditors posting in this community understand that others will attempt to change their view by providing arguments opposing their beliefs by providing arguments opposing their perspective (see 4.1; also see Table S1 in Supplemental Materials, found at https://osf.io/v54ut/). As one would expect, some arguments are more persuasive than others and thus the variance in argument quality found on the forum provides a naturalistic resource for examining the features of effective arguments.

As a naturalistic data source, Change My View has provided several insights into how belief change occurs outside of the lab. Researchers have examined the logical qualities of effective arguments on the forum (e.g., use of classical modes of persuasion: ethos, logos, pathos, Hidey & McKeown, 2018). Research on Change My View has extended beyond social psychology. Computer scientists have developed computational models that extract features of argumentation, such as predicting the probability an argument is effective given linguistic features (Tan et al., 2016; Chakrabarty et al., 2020) or machine classifying "parts" of beliefs most amenable to change (Jo et al., 2018).

As one would expect, some user-generated arguments are more persuasive than others and thus provide a naturalistic dataset for examining the factors that predict attitude change outside of the lab. However, rather than attempting to extract nuanced but not easily generalizable linguistic properties of convincing arguments in online communities (Hidey & McKeown, 2018; Jo et al., 2018; Tan et al., 2016), our work looked at global factors of attitude change across entire discussion threads taking place online (naturalistic study), before culling effective arguments to test if their effects generalize outside of Reddit. We did this with

the hope that by identifying these global factors, behavioral scientists and policymakers could incorporate the lessons of attitude change occurring in the wild into educational interventions developed in the lab.

Naturalistic Study: Evidence revises beliefs in online discussions

We began to study this rich source of data by investigating how attitude change varies between posts focused on "sociomoral" and "non-sociomoral" topics. Sociomoral posts, as we are defining them, relate to social and moral issues; the most common sociomoral posts in this forum concern politics, questions of gender identity, and current events like recent elections. In contrast, posts that are not sociomoral are a grab bag of other topics including humor and debates about movies and fiction. We sought to address three questions about sociomoral attitude change, which using this sort of naturalistic dataset can uniquely address: First, as we would intuitively expect, do people change their minds less often about sociomoral issues, in general, compared to non-sociomoral issues? Second, how do the contents of arguments differ, or do they differ at all, for these two types of discussions? Finally, regardless of domain, do facts, evidence, and data promote attitude change in online forums? By answering these questions, we sought to understand the overarching factors that promote attitude change in real life settings about a variety of topics in a way that cannot be easily studied in the lab.

Methods

Our procedure had five steps: 1) Preregistration, 2) Collecting users' submissions on the website Change My View, 3) classifying posts as sociomoral or not 4) categorizing arguments that successfully change someone's mind, and 5) quantifying how much evidence was provided in a given discussion thread.

Preregistration

We preregistered this project's procedure and our hypothesis concerning attitude change in sociomoral posts on Open Science Framework. The registration for this study can be found at the following link: osf.io/jdxa8.

Submission Collection

We developed a Python script that collected 500 top Change My View posts using Version 5.3.0 of the Python Reddit API Wrapper (PRAW) (2017). Top posts are rich and mature discussions with many reply threads and participating users. Analyzing top posts allows us to consider well-developed discussions in their entirety as opposed to "young" discussions that have few comments.

Submission Classification

We coded a post as sociomoral if it concerned political, moral, or social issues. Two posts coded as sociomoral in our dataset were "U.S. military spending is unnecessarily large" and "Donald Trump has drastically changed the political landscape". Alternatively, posts that were coded as non-sociomoral sometimes involved fictional components or intended to be humorous. Two examples of posts coded as non-sociomoral in our dataset were "Thank You Cards are a waste of time and money" and "Luigi is the superior Mario Brother". All 500 posts were coded by J. Priniski, and then a second hypothesis blind coder recoded 25% of the posts (N = 125), agreeing on 88.8% of the original codings.

Measuring Belief Change

We sought to examine the factors that promoted attitude change on Change My View. To this end, we developed a way to flag posts that changed people's minds. On Change My View, there is a protocol—namely, delta awarding— which serves as a proxy of attitude change. Both the original poster and others can award comments a delta if they even partially change their mind about an issue. Delta awarding occurs when a user signifies that an argument has changed their mind. A delta can be awarded by replying to a comment with one of the following delta strings: "Δ" and "!delta".

To find Delta Awarded Comments (DACs), we traversed the discussion tree returned by the Reddit API using breadth-first search and string matching each comment for a delta signification. When a delta string is encountered, we moved upwards through the node's ancestors until the root of the thread is found. This allowed us to distill the thread of conversation that lead someone to change their mind. In some cases, there is a back-and-forth between, for instance, two users until one user is finally convinced of the argument. In these cases, the thread is multiple replies in length. In most cases, however, the thread is only a single reply deep. All delta threads gathered from the 500 top posts and the code that collected them can be found at osf.io/yvunj.

Measuring Evidence Use

In addition to collecting DACs, we also examined how Change My View users incorporate evidence in their replies. Since we are using naturalistic data, we were required to some extent to make some inferential leaps regarding just what constitutes evidence in Reddit forums. We calculated evidence use by considering two measures: (1) the number of hyperlinks that cite external websites and documents and (2) a discussant's use of "statistical language." To collect and count hyperlinks, we searched the markup text returned by the Reddit API for words containing typical website and document identifiers, such as: 'http://', 'www.', '.pdf',

'.com', etc. The complete list of identifiers can be found on the project's Github: github.com/jpriniski/CMV. We calculated statistical language by string matching words in discussion threads with statistical terms and symbols, such as: 'data', '%', 'stats', and so on. The code that completes this task can be found on the Github linked above.

Results

Analytic Strategy

Rather than performing null hypothesis significance testing, we performed Bayesian modeling using the programming language Stan in the R package brms. We specified priors to guide estimation of the data but these priors did not predetermine the results of any analysis. All the analyses reported herein are robust to different prior choices.

Our first question concerned the ways in which discussions of sociomoral issues differ compared to discussions of non-sociomoral issues. We examined how the rate of participation differs in discussions that concerned sociomoral and non-sociomoral topics. Sociomoral issues, by definition, are related to issues relevant to society at large and thus are more likely to be of interest to many people. We measured interest and participation by predicting the total number of comments in a discussion thread on the basis of discussion topic (i.e., sociomoral or non-sociomoral). As one might expect, we found that there was considerably more interest in sociomoral discussions compared to non-sociomoral discussions (see Table 4.2).

Table 4.2: Poisson regression predicting the amount of comments in a discussion based on topic type. Confidence interval represent 95% Bayesian Credible Interval. Non-sociomoral posts were the reference group.

Effect	Estimate	Lower	Upper
Intercept	5.59	5.58	5.60
Sociomoral	**0.26**	**0.25**	**0.28**

We then examined whether sociomoral posts prompt people to cite more evidence to support their beliefs than non-sociomoral posts. We calculated the number of comments that contained links, the total amount of links in a discussion, and the total amount of "statistically-oriented language" used in the discussion. These analyses indicated that evidence is more frequently provided in people's debates about sociomoral topics than non-sociomoral topics, see Figure 4.1 and Table 4.3.

However, even though users cited considerably more evidence to advance their arguments, attitude change in the sociomoral domain is as common it is in the nonsociomoral domain – that is, more evidence yielded equivalent amounts of attitude change. As Figure 4.1 shows, we found that total delta awarding and Delta

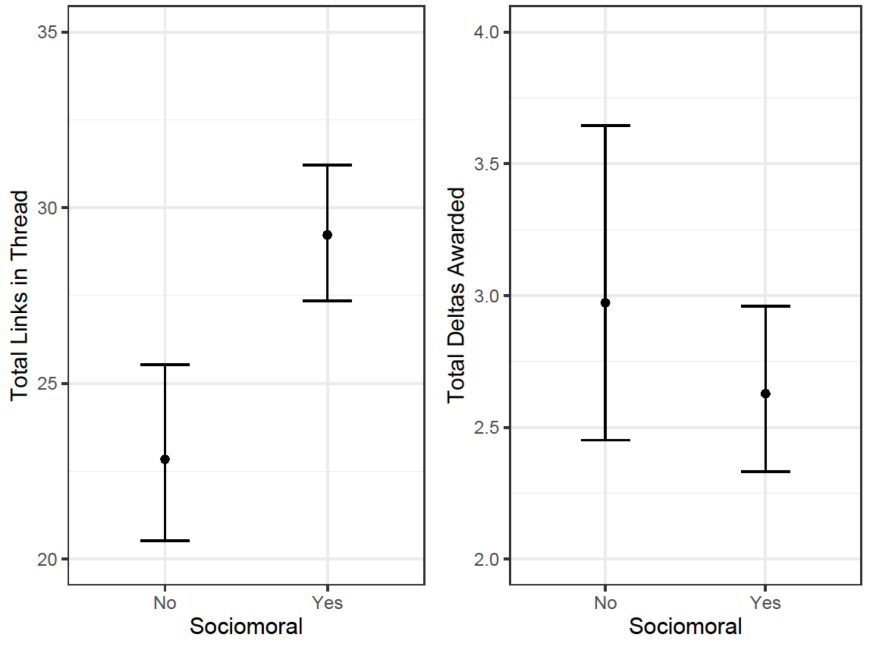

Figure 4.1: **More "evidence" offered in sociomoral discussions but no more deltas.**

Table 4.3: A multivariate negative binomial model predicting the amount of comments with links, the total amount of links in a discussion, and the amount of statistical language based on whether the thread concerned a sociomoral issue or not. Estimates are 95 % Bayesian credible intervals.

Effect	Estimate	Lower	Upper
Intercept	2.76	2.66	2.85
Comments using links	**0.19**	**0.08**	**0.30**
Intercept	3.13	3.02	3.24
Total links	**0.25**	**0.11**	**0.38**
Intercept	4.35	4.29	4.41
Statistical Language	**0.03**	**-0.04**	**0.11**

Awarded Comments (i.e., our measures of attitude change) occur at similar rates in sociomoral and nonsociomoral threads (see Table 3).

Table 4.4: A multivariate Poisson regression predicting the number of Delta Awarded Comments (DACs), the number of DACs with links, the number of DACs that include statistical language, and the total amount of deltas on the basis of topic type.

Effect	Estimate	Lower	Upper
Intercept	0.65	0.51	0.77
DACs	**-0.13**	**-0.28**	**0.02**
Intercept	-1.29	-1.63	-0.96
DACs including links	**0.00**	**-0.39**	**0.41**
Intercept	0.00	-0.18	0.17
DACs including stat. lang.	**-0.22**	**-0.43**	**-0.01**
Intercept	1.09	0.89	1.30
Deltas	**-0.13**	**-0.37**	**0.11**

Note. Non-sociomoral posts were the reference group.

These results suggest that even though commenters on Change My View participate in a forum dedicated to challenging one's own views, the members of this community are not particularly likely to change their minds. In this way, users on Change My View are perhaps more representative of how people "in the wild" change their minds than we might initially expect.

It may appear that our results are compelling evidence for the pessimistic view that people cannot agree on what facts, if any, are even relevant to debates about, for example, politics, morality, or gender. We found that, even though substantially more evidence is cited in sociomoral discussions, attitude change is no more common in these discussions than non-sociomoral discussions.

However, our data provides some reason for optimism: Consistent with prior behavioral research (e.g., Baesler & Burgoon, 1994), we observed in a highly naturalistic dataset and across several measures of "evidence", citing sources and referencing data was positively related to attitude change (see Figure 4.2). When a thread contained, for instance, more citations, links to external sources, or statistical language, it positively predicted attitude change. Furthermore, we found that this effect did not depend on the discussion being sociomoral in nature (see Table 4.4 below; more details can be found at osf.io/s3rny). This result provides "real world" evidence that when people are motivated to attend to information relevant to their beliefs, citing sources, providing data, and so forth can be an efficacious tactic for changing people's attitudes (see Petty & Cacioppo, 1986, for prior laboratory-based studies suggesting this same conclusion).

Discussion

Here, we examined the factors that promote attitude change in hotly debated topics, using a naturalistic dataset by studying attitude change in over $100,000$ comments in 500 discussion threads on Reddit's Change

Table 4.5: Belief change predicted by the amount of evidence cited in a discussion.

Effect	Estimate	Lower	Upper
Effects of comments with links and total links on number of deltas			
Intercept	0.55	0.31	0.79
Comments with links	**0.67**	**0.48**	**0.87**
Sociomoral	-0.23	-0.45	0.00
Comments	-0.12	-0.24	0.01
Intercept	0.82	0.59	1.05
Total links	**0.39**	**0.22**	**0.57**
Sociomoral	-0.18	-0.41	0.05
Comments	-0.01	-0.13	0.10
Effects of comments with links and total links on number of DACs			
Intercept	0.38	0.21	0.55
Comments with links	**0.35**	**0.21**	**0.48**
Sociomoral	-0.18	-0.33	-0.03
Comments	-0.12	-0.20	-0.03
Intercept	0.56	0.41	0.75
Total links	**0.13**	**0.03**	**0.23**
Sociomoral	-0.15	-0.30	0.01
Comments	-0.04	-0.12	0.04
Effects of statistical language on number of deltas			
Intercept	1.09	0.89	1.29
Stat. Lang.	**0.20**	**0.05**	**0.34**
Sociomoral	-0.13	-0.37	0.10
Comments	-0.02	-0.17	-0.13
Effects of statistical language on number of DACs			
Intercept	0.64	0.51	0.77
Stat. Lang.	**0.06**	**-0.03**	**0.16**
Sociomoral	-0.13	-0.28	0.03
Comments	-0.04	-0.14	0.06

Note. We distinguish between Delta Awarded Comments (DACs) and deltas because they could be distinct measures of attitude change. For example, very few comments could be awarded several deltas (one comment could receive 10 deltas). Alternatively, several comments could each be awarded just a few deltas (three comments could each receive two deltas). For this reason, we thought it was important to distinguish these indicators of belief change.

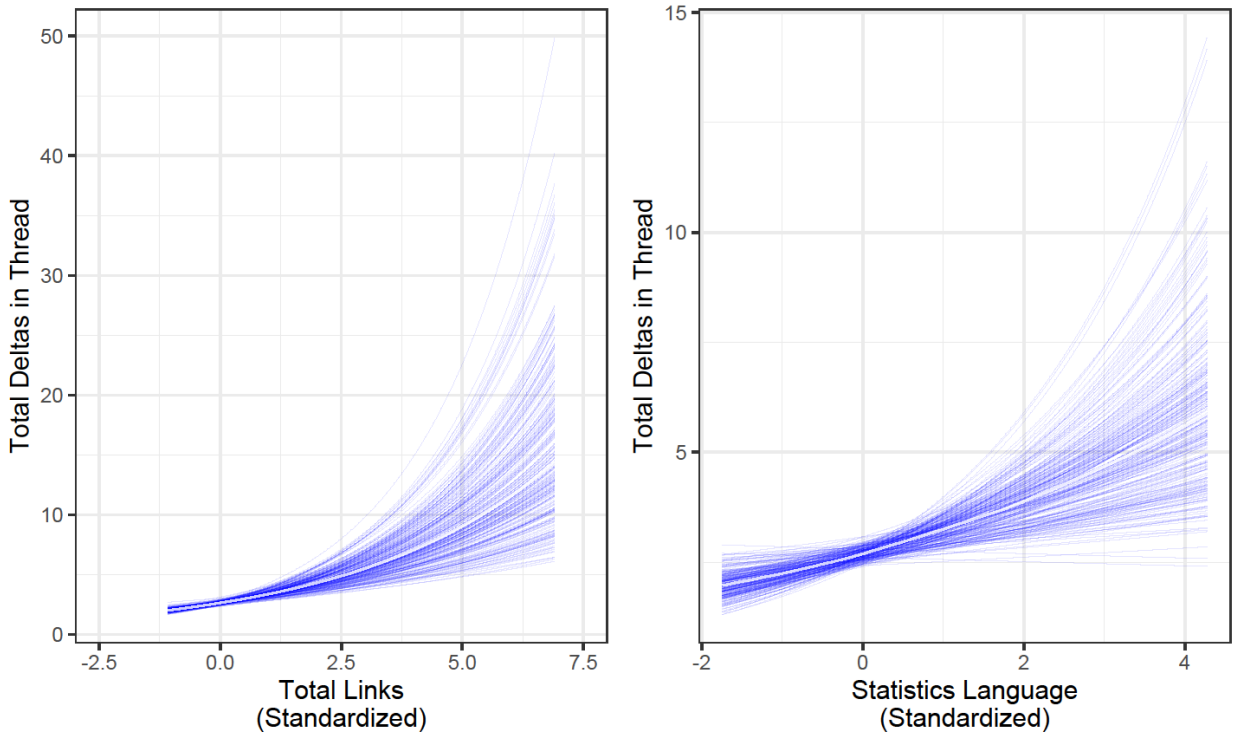

Figure 4.2: Marginal effects spaghetti plot predicting the total number of deltas awarded in a discussion thread based on the total number of links to external sources of information (left) and amount of statistical language (right). Individual blue lines represent draws from the posterior interval, and the region they produce represents 95% Credibility Intervals.

My View. This study revealed that even though users cite considerably more evidence while discussing sociomoral topics, they appear equally likely to revise their beliefs compared to topics that do not concern sociomoral issues. However, we also found that regardless of discussion type, providing evidence for a claim (for example, in the form of links to external articles) positively impacted people's willingness to change their minds. Thus, our work may suggest that while attitude change is hard-won, providing facts, statistics, and citations for one's arguments can convince people to change their minds.

Limitations and Future Directions

One concern with using Reddit's Change My View to understand how attitudes change "in the wild" is that the people who participate in this Reddit forum may be particularly willing to change their minds and to consider statistical evidence for an argument. This may not be true in the case of the general population (e.g., Nickerson, 1998), which may limit the generalizability of our findings. Several facts speak against this concern, however. First, we observed that, in a given thread, approximately 400 comments would yield only two deltas – only two people change their mind in a thread containing numerous arguments, citations, and statistics. This is the kind of proportion we would expect to observe in the general population where everyday experience tells us that attitude change rarely occurs, if ever. Second, we also found that, as we would expect about the general population, it is harder to change people's minds over sociomoral issues compared to non-sociomoral issues. So, in contrast to this initial hunch about the users of Change My View, we suspect that our results here would be more representative of how attitude change occurs beyond the artificial conditions imposed by the laboratory.

However, there are unquestionably several limitations of this naturalistic dataset that must be acknowledge. First, it may be that Redditors are unwilling to award a delta even when they have experienced attitude change, a limitation that future research may be able to address by surveying Redditors. Second, it must be noted that members of this community are motivated to deliberate on things discussed in the threads on Change My View. This quality of the forum users makes it an ideal population to study central rather than peripheral routes to persuasion (Petty & Cacioppo, 1986) but may be unrepresentative of the general population. For this reason, more research is necessary to understand the extent to which the persuasive tactics deployed by users on Change My View would generalize to populations who are not so motivated to consider the facts relevant to their beliefs.

While many researchers have examined the factors that predict belief change among Change My View users, it is unknown whether effective arguments taken from this forum would be equally effective in more controlled contexts or among a population not seeking arguments opposing their beliefs. In fact, there are several reasons why belief revision may look different on Change My View than it does in the lab. These

reasons pose concern for the generalizability of effective arguments found on Change My View and need to be experimentally addressed before Change My View can be recommended as a crowdsourcing platform for effective educational interventions.

For one, people who discuss certain topics——and particularly users on Reddit's Change My View——may be more willing to change their minds and consider evidence for an opposing argument. This may not be true for the public at large, limiting the generalizability of these prior findings. Second, people engaged in a debate on a particular topic may be more motivated to deliberate on the topics they're discussing. This fact may make online communities such as Change My View an ideal population to study central rather than peripheral routes to persuasion Petty & Cacioppo (1986). However, it may also make online communities unrepresentative of the general population who may not be so ready to entertain evidence that is contrary to their beliefs.

Altogether, controlled laboratory research is necessary to understand if the persuasive tactics deployed online can generalize to other populations and, in turn, serve as a starting place for developing educational interventions.

Behavioral experiments with crowdsourced interventions

In the present experiments, we identified successful arguments on Change My View and performed a head-to-head comparison to interventions reported in academic psychology, public policy, political science, communications, and behavioral economics articles—adopting a methodological approach most analogous to a strategy relied on in clinical trials (e.g., Leuch, et al., 2013). Namely, we compared crowdsourced arguments to academic arguments that have been shown to be somewhat effective at changing people's beliefs (or at a minimum, exert the same task demands on participants). Performing this comparison allowed us to predict whether effective educational interventions can be culled from online communities and used as effective interventions in controlled laboratory settings.

It is worth highlighting how this experimental strategy diverges from comparing the performance of an intervention to an inactive control condition. As opposed to controlling for features of naturalistic interventions to uncover what makes them effective, the paradigm we are proposing first identifies the interventions that yield desirable consequences (e.g., a reduction in misconceptions surrounding structural racism), at which point we can subsequently uncover the mechanisms that realize these positive effects. As a consequence, academic and crowdsourced interventions will differ along many unknown dimensions (including length, the task performed, the information presented, and so on). However, we do have prior evidence (either from empirical studies or from data mined from discussion forums) that signal the efficacy

of each of the interventions being compared. Ultimately, researchers aim to develop interventions that can effectively educate the public, making this dimension——efficacy——the most central on which to assess an educational intervention.

With this goal in mind, in Experiment 1 we compared the efficacy of crowdsourced and academic interventions at changing beliefs across four hotly-debated topics. Experiment 2 was an extension of Experiment 1, where we further examined whether crowdsourced arguments would be as effective as academic intervention across four new topics.

Experiment 1a: Comparing academic and crowdsourced interventions across four topics

Preregistration

The projected sample size, predictions, and analysis scripts were preregistered through Open Science Framework. Experimental scripts, analyses, scales, and Supplemental Materials are available at https://osf.io/v54ut/.

Participants

We recruited 916 participants through Amazon's Mechanical Turk to be 80% powered to detect a Cohen's d of .1 in a within-subjects design. Of the participants recruited, 816 passed attention checks and were included in the analysis of this study (333 men, 476 women, 4 non-binary, 4 preferred not to say; the median age of participants was 35 years old).

Interventions

Participants received four separate interventions that focused on either (a) reducing racist beliefs, (b) increasing support for vaccines, (c) increasing support for gun control, and (d) reducing xenophobic attitudes directed at immigrants. Participants received two crowdsourced interventions and two academic interventions (intervention type: within-subjects) with one intervention for each topic. Therefore, we tested the efficacy of eight interventions in total. Crowdsourced interventions were copied-and-pasted comments that were awarded a "delta" in a Change My View discussion—a signification that the argument changed the view of at least one user on the forum. We selected discussion comments from Change My View as crowdsourced interventions if they met the following three criteria. First, the comment was related to a topic that psychologists have traditionally studied in the lab (e.g., climate change, gun control, xenophobia, etc.). Second, the comment had been awarded a delta. Third, the content of the comments could be developed into an intervention with little-to-no editing, content change, or manipulation. Many comments on Change

My View satisfy these criteria and *could* have been empirically tested, but the aim of the present studies is to consider how several representative crowdsourced examples could be developed into effective educational interventions. (Detailed information about the interventions can be found at https://osf.io/v54ut/).

Pretest and Posttest Measures

We examined how participants' beliefs about four controversial topics changed as a function of exposure to one of two educational interventions (crowdsourced or academic) for a given topic. Prior to completing the main portion of the study, participants answered four questions assessing their pretest beliefs about each topic. For instance, participants rated their agreement with the assertion, "Gun control in America is ineffective at reducing overall violence and crime", which was taken from a Change My View post (in this case, a post about gun control). After responding to these four assertions, participants proceeded to the intervention and post-test portion of the experiment.

We developed four separate scales to measure people's beliefs about racism, vaccines, gun control, and xenophobia directed at immigrants. Each scale was composed of five items (with two items reverse coded). Items in a topic's posttest scale were created by rewording or expanding on a pretest assertion. For example, an item in the posttest gun control scale stated, "Societies with strict gun control have similar crime rates as societies with little to no gun control." See the Supplemental Materials for more details on pretest and posttest measures.

Procedure

The experiment proceeded as follows: First, participants rated their agreement with items measuring their pretest beliefs towards all four topics. Next, participants were randomly assigned either an academic or a crowdsourced intervention for a given topic. After completing this intervention (e.g., after reading information about gun control), participants responded to that topic's posttest scale. After completing the posttest scale for a given topic, participants advanced to a new topic and the procedure was reiterated until they finished reading and responding to questions about all four topics. The ordering and exposure to a given intervention type was counterbalanced and randomized.

Results and Discussion

Analytic Approach

To test our hypotheses, we performed Bayesian mixed effects modeling using the R package brms Bürkner (2017). We set regularizing priors for all population- level effects in our models, which we detail below. These

priors are recommended because they provide conservative effect size estimates and reduce the likelihood of overfitting A. Gelman et al. (2015); McElreath (2016). Following the recommendations of Liddell & Kruschke (2018), Likert data were modeled with a cumulative probability distribution. The cumulative distribution is recommended for Likert scale data because it assumes that ordered responses represent a continuous latent construct.

We tested our hypothesis by fitting an ordinal mixed-effects model predicting posttest beliefs based on the interaction between condition (Reference = Academic condition) and topic (Reference = Guns). This model controlled for participants' responses to the pretest statement, which we treated as a monotonic effect. This model included group-level effects of Subject and Topic and allowed for heterogeneity in the slopes of the effects of Condition and Topic on participants' responses. Our model is specified below in brms syntax Bürkner (2017): Response $\sim$ *Condition*Topic + mo(PreTest) + (1 + Topic + Condition | Subject).* Bayesian analyses formulate model parameters as probability distributions wherein the posterior distribution for a parameter θ is computed via the prior and likelihood of θ. To model the joint probability distribution of participants' responses, we specified priors over the possible effects each parameter could have on our response variable[1]

As shown in Figures 4.3 and 4.4, these analyses revealed that the crowdsourced interventions countering racist ($\beta = -.58$, 95% CI $[-.80, -.37]$) and anti-immigrant beliefs ($\beta = -.40$, 95% CI $[-.60, -.18]$) were credibly more effective than an academic intervention; interventions on vaccines and gun control were equally effective (see Figure 4.4). These results suggest that there are arguments being developed in online communities that are comparably effective to interventions behavioral scientists have developed. And considering crowdsourced arguments have the additional virtue of being shown to be effective in a naturalistic setting free from task demands, this may give additional motivation for beginning development of educational interventions on the basis of crowdsourced arguments.

However, given that the present design lacks a completely neutral control condition, it is important to be clear on what these results do not show. First, these results do not demonstrate the true magnitude of the effect of a given intervention. Second, there is a large amount of variance in intervention quality and effectiveness for any intervention type, and there is no reason to think that all crowdsourced arguments will always be as effective or more effective than academic interventions. Rather, one should interpret the results of Experiment 1a as suggesting that crowdsourced arguments can provide a starting place for developing educational interventions, and doing so has the additional virtue of giving us a priori reason to think they will generalize to comparatively more naturalistic settings.

[1] $\alpha_1 \sim N(2.19, 1)$; $\alpha_2 \sim N(2.94, 1)$; $\alpha_3 \sim N(3.17, 1)$; $\alpha_4 \sim N(3.47, 1)$; $\alpha_5 \sim N(3.89, 1)$; $\alpha_6 \sim N(4.59, 1)$; $\beta_{Condition} \sim N(0, .5)$; $\beta_{Pretest\ Beliefs} \sim N(4, 2)$; $\beta_{Topics} \sim N(0, 3)$; $\beta_{Topic\ \times\ Condition\ Interactions} \sim N(0, .5)$; $\Omega_k \sim LKJ(1)$ where Ω_k is a correlation matrix of group-level parameter; Group-level parameters $\sim N(1, 2)$.

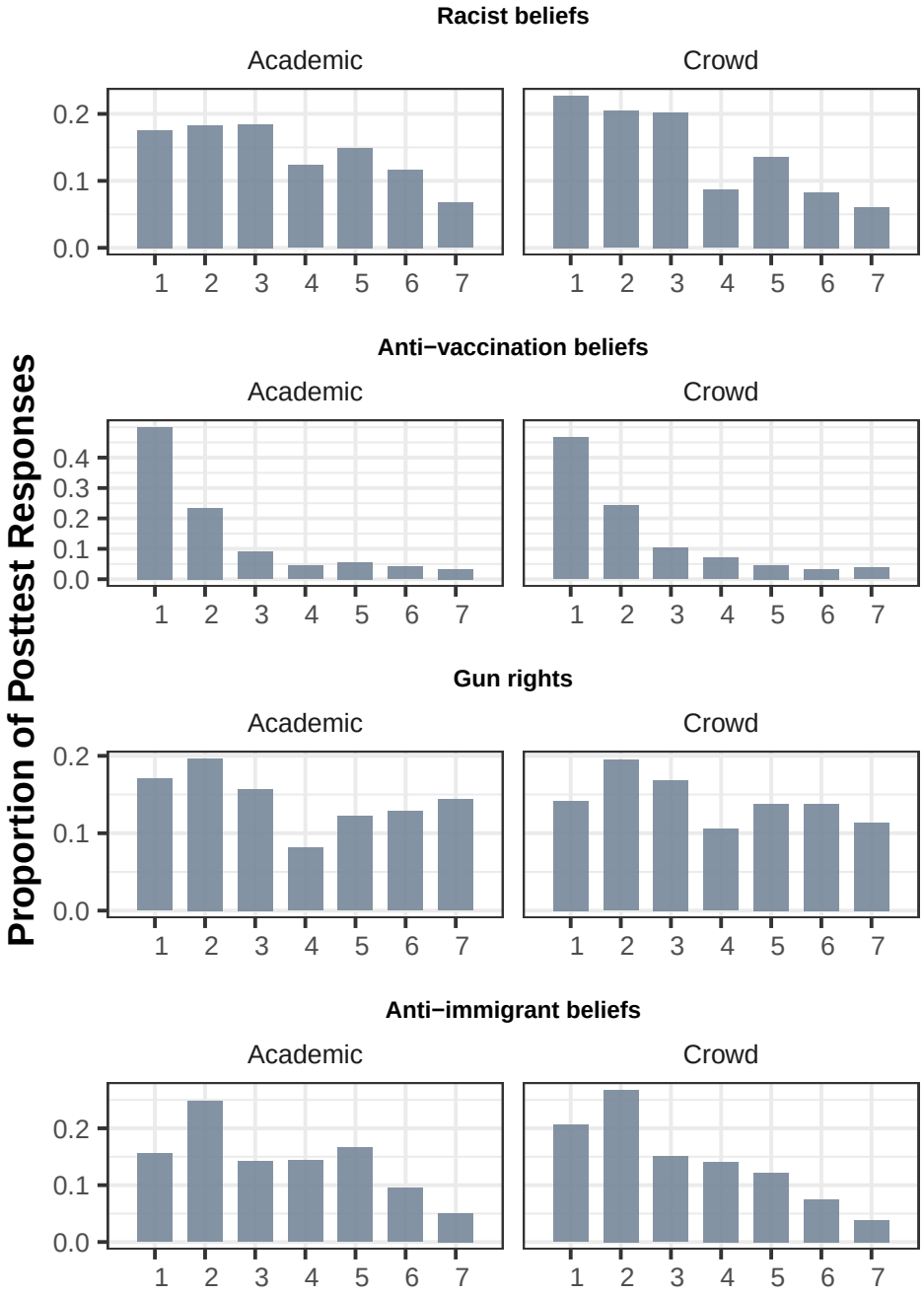

Figure 4.3: Posttest responses for each intervention tested in Experiment 1a (1 = Strongly disagree; 7 = Strongly agree). Relative effectiveness of a crowdsourced intervention can be seen by comparing the leftward shift of responses across interventions for a topic. Figures S5 through S8 in the Supplemental Materials show postintervention responses grouped by pretest response for each intervention tested in tested in Experiment 1.

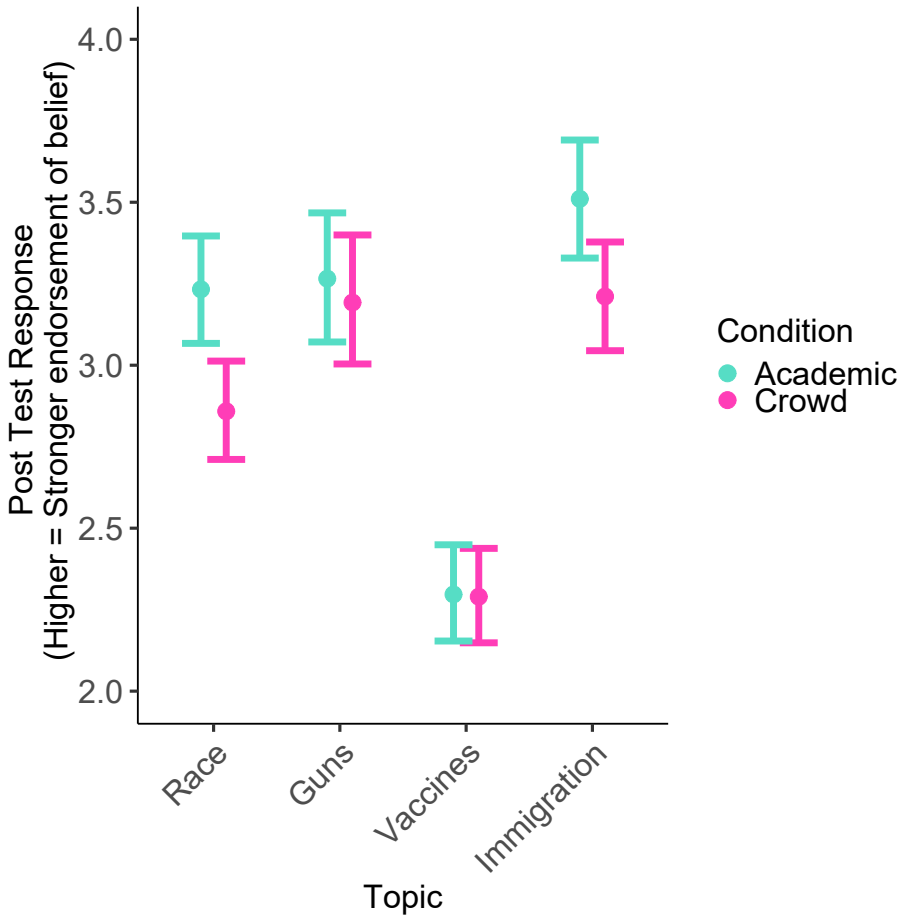

Figure 4.4: Marginal effect plot of responses for each intervention tested in Experiment 1a (1 = Strongly disagree; 7 = Strongly agree). Error bars represent 95% Bayesian credible intervals

Experiment 1b: Extending efficacy comparison with four new topics

Experiment 1b was a preregistered extension of Experiment 1a. The registration for this project can be found at https://osf.io/v54ut/. This experiment followed an identical procedure but tested the efficacy of academic and crowdsourced interventions on four new topics: (a) reducing sexist beliefs, (b) reducing transphobic beliefs, (c) reducing denial in the negative effects of climate change, and (d) reducing favor for capital punishment.

Participants

We recruited 900 participants through Amazon's Mechanical Turk to be 80% powered to detect a Cohen's d of .1 in a within-subjects design. Of the participants recruited, 745 passed attention checks and were included in the analysis of this study (325 men, 416 women, 3 non-binary, 1 preferred not to say; the median age of participants was 33 years old).

Results and Discussion

Like Experiment 1a, we predicted that crowdsourced interventions would be as effective or more effective than academic interventions for the four new topics. We fit the same ordinal regression model with the same priors as Experiment 1a.

As shown in Figures 4 and 4, in Experiment 1b, we found that crowdsourced interventions were equally effective as academic interventions across three topics; the academic intervention aimed at shifting people's beliefs about climate change was more effective than the crowdsourced intervention, $b = .24$, 95% CI [.00, .40].

Experiment 2: Embedding crowdsourced information in an interactive data narrative

However, the present experiments have some clear limitations. By design, both experiments lacked a true control condition, leaving an important question unanswered: Exactly how effective are these interventions at changing beliefs? Experiments 1a and 1b compared the *relative* effectiveness of crowdsourced interventions to academic interventions, and didn't demonstrate how effective they are with respect to a neutral control condition. In Experiment 2 we compared the most effective intervention to a true control condition in order to make explicit how effective a given intervention is at changing beliefs. We then generalized the effects by comparing to a re-representation of the core causal information in the narrative communicated by the crowdsourced intervention.

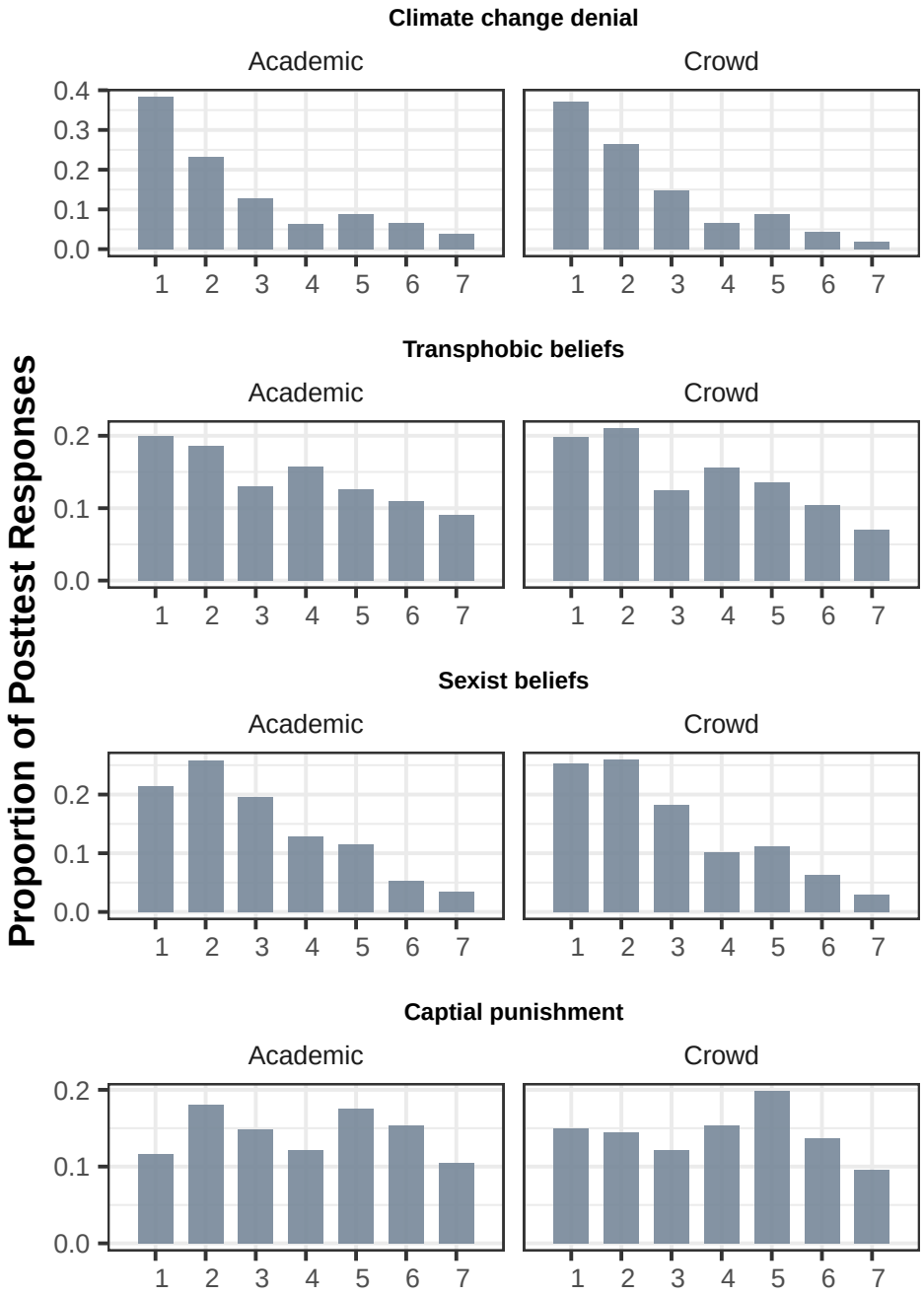

Figure 4.5: Posttest responses for each intervention tested in Experiment 1b (1 = Strongly disagree; 7 = Strongly agree). Relative effectiveness of a crowdsourced intervention can be seen by comparing the leftward shift of responses across interventions for a topic. Figures S9 through S12 in the Supplemental Materials show posttest responses grouped by pretest response for each intervention tested in tested in Experiment 1b.

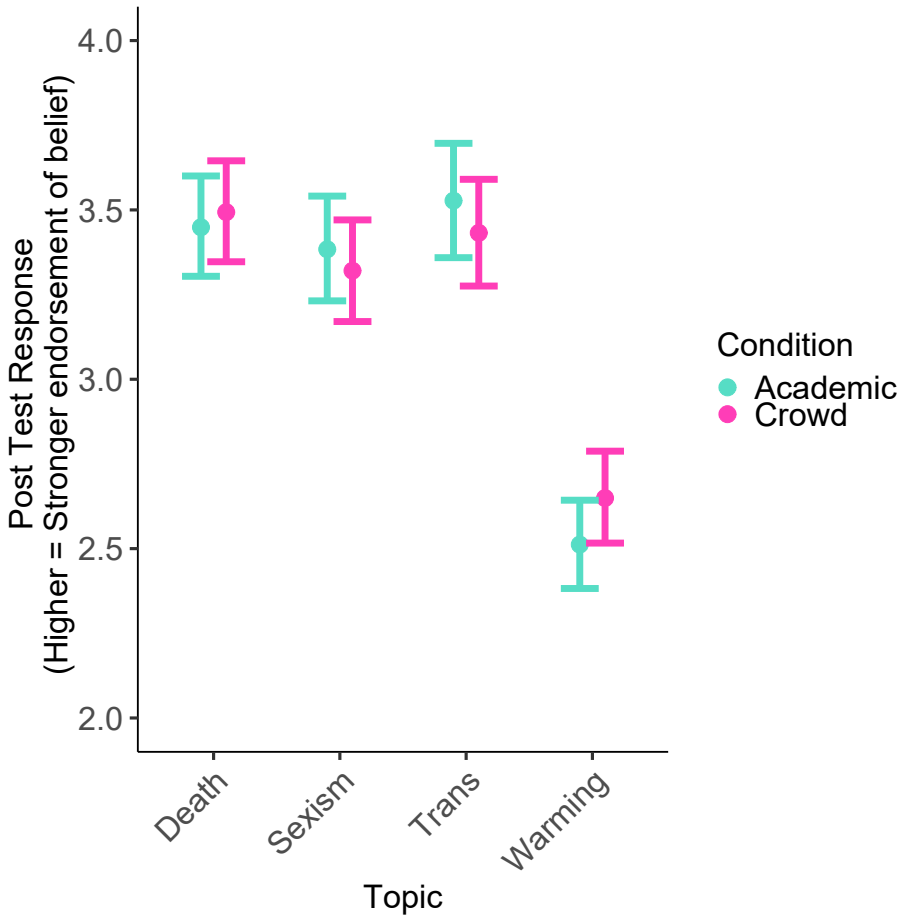

Figure 4.6: Marginal effect plot of responses for each intervention tested in Experiment 1b (1 = Strongly disagree; 7 = Strongly agree). Error bars represent 95% Bayesian credible intervals

Preregistration

We preregistered the experiment, predictions, and analysis pipeline on the Open Science Framework.

Participants

We recruited $N = 1760$ American participants using Amazon's Mechanical Turk.

Interventions

Extending the design from the previous experiments, participants were placed into one of four experimental conditions. The "academic intervention" was tested in the previous experiments, where particiapnts were asked to list out examples of racial injustice. This intervention is designed to active particaipants memory of causal mecahnisms, and is passive in that no learning occurs. In the bare control condition, participants asked to locate Waldo in an image. The crowdsourced intervention was the complete Reddit argument taken from ChangeMyView tested in Experiment 1a. We extended this information with an interactive data visualization communicating the core causal relationship in the Reddit argument. I identified this relationship by parsing the document with the causal language analysis pipeline (J. Priniski et al., 2023), which highlighted the causal link mentioned in the crowdsourced intervention between historical redlining practices and financial inequities today. Crowdviz intervention, with an added data visualization component, was used to test if highlighting salient causal information in a passage is necessary to optimize learning of causal relationships embedded in text based educational narratives.

The crowdsourced intervention communicates the effects of the racist government practice of redlining. In the Crowdviz intervention, we augmented and hoped to better demonstrate this point by collecting census and historical redlining data for neighborhoods in South Chicago and mapped it onto a digital map (see 4.7). Red zones were neighborhoods primarily composed of African Americans in the 1930s and were places where people couldn't get home loans, while purple and green zones were neighborhoods composed of White Americans in the 1930s. It was easy to get a home loan in these zones. To make the connection between 1930s redlining practices and neighborhood incomes today more salient, we highlighted two parts of South Chicago. The red box in Fig. 4.7 covers an area completely redlined, while the green box covers an area with many more positive grades. Next, we showed participants a map of present day median incomes of south Chicago residents and kept the boxes from the previous redlined map. As you can see, previously redlined neighborhoods are still poorer than the neighborhoods that were not redlined in the 1930s. When a participant clicks on a neighborhood, they see the distribution of African Americans living in that neighborhood. The red point in the scatterplot is the neighborhood they clicked on. Altogether, these

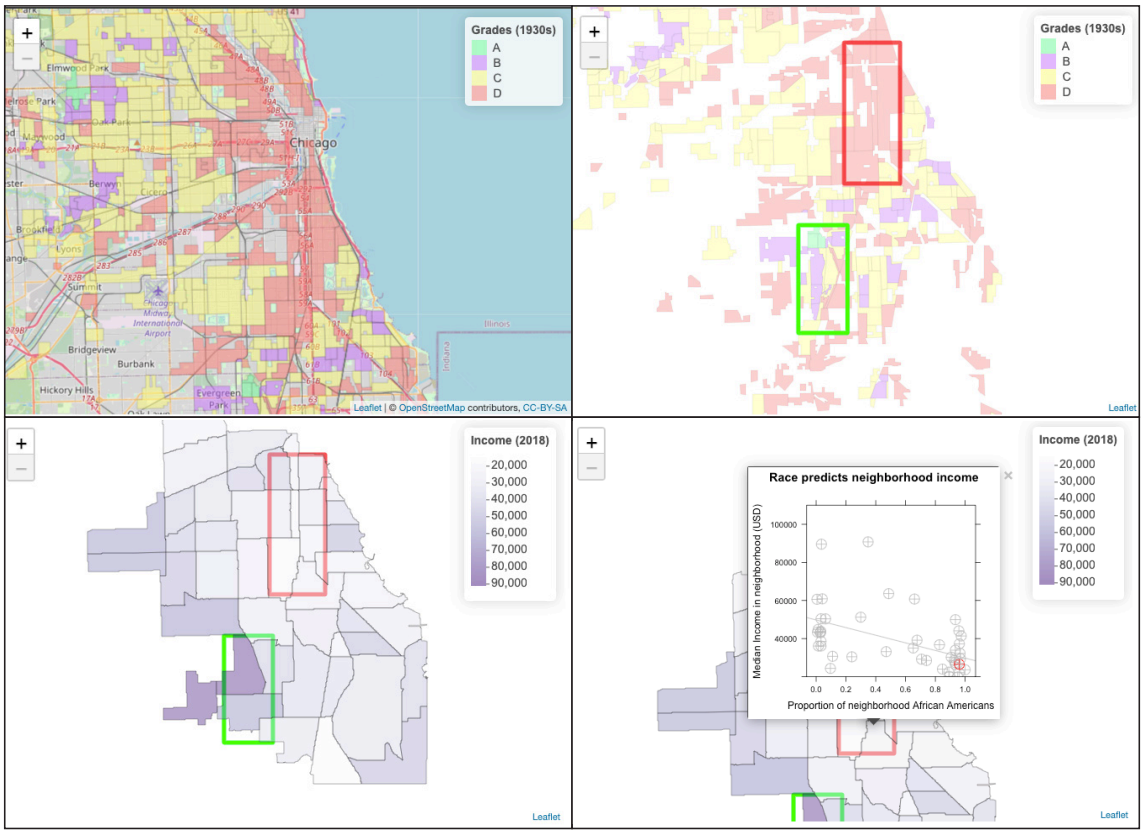

Figure 4.7: **Interactive data narrative communicating causes of racial inequity in South Chicago.** Participants were able to see how historical redlining practices in South Chicago predicted present data income disparity for African Americans. Income data sourced from Census. Redlining markers from GET.

series of maps communicate how racist practices in the 1930s structurally effect communities today. This was the main point of the crowdsourced intervention and we hoped that including it would lead to a more dramatic shift in attitudes.

Procedure

There were four conditions in a between-subjects design. Participants first responded to a pre-intervention scale, before recieving one of four educational interventions, and then post-intervention scales. We sought to replicate the initial effect of the crowdsourced argument as well as extend it by adding the Crowdviz condition.

Results

Figure 4.8 shows the distribution of postintervention responses across all four experimental conditions. The Academic and Control distribution have nearly identical distribution of responses, suggesting that having participants list of examples of racism is no more effective at shifting attitudes than the bare control. Comparing the distributoin of responses in the crowd intervention pannel to the Academic and Control pannels allows us to assess if the effect of the crwodsourced race intervention can be replicated form Experiment 1a. As was the case in the Pilot study, we see the leftward shift in the distribution of responses, suggesting the replication was successful. Finally, comparing these conditions to the extended crowdsourced condition allows us to assess if additional interactivity can make core causal redlining information more salient. When compared to the crowdsourced intervention, we don't see much of a shift. However, we still see the leftward shift in responses when compared to the Academic and Control conditions. While we had initially hoped that the visualization would make the intervention even stronger, we didn't see that. What this does demonstrate, however, is that augmenting the intervention didn't distinguish its effect. It is therefore more likely a robust intervention, that's not so fragile that tweaking its presentation and framing eradicates its underlying effect. The lack of additive effect from the visualization suggests that the causal information embedded in the language is necessary for shifting attitudes. While additional follow-up experiments are necessary, this suggests that natural language processing models that extract causal information from discourse can help produce interventions.

Discussion

People's beliefs about topics like science and morality are stubbornly resistant to new information. Developing educational interventions to correct these beliefs is a difficult task that often results in fruitless outcomes. It is also often unknown whether an intervention that manages to successfully shift beliefs in the lab will be similarly effective in a more naturalistic setting. The present studies suggest that researchers can use crowdsourced arguments to better predict and develop effective educational interventions. Furthermore, crowdsourcing effective arguments can impact the study of belief revision directly by elucidating which types of information are most effective at changing strongly held beliefs: a topic of interest to many researchers studying higher-level cognitive processes. In two experiments, we tested whether arguments crowdsourced from the Reddit forum Change My View could be used to such an end. In Experiments 1 and 2, we compared arguments crowdsourced from Change My View to interventions taken from academic research in psychology, communications, political science, behavioral economics, and public policy. In Experiment 1, we

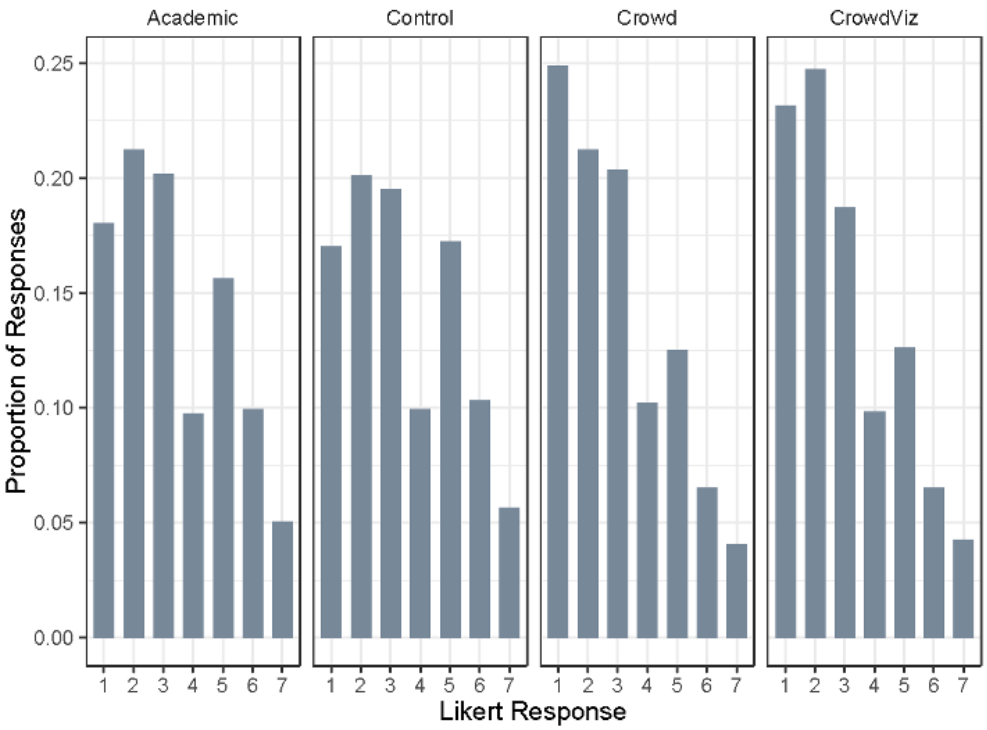

Figure 4.8: **Distribution of postintervention responses in Experiment 2**. 7 = stronger endorsement of misconception. The academic and control conditions have similar response patterns; while the crowdsourced (Crowd) and crowdsourced based interactive visualization (CrowdViz) positively shifted attitudes. However, these two conditions induce similar response patterns, which suggests that the effects of the crowdsourced intervention are robust, but that the causal information in narrative's language is enough to shift attitudes.

found that across four topics, crowdsourced arguments were as effective or more effective at changing beliefs compared to previously published or tested educational interventions developed by academics. Experiment 2 followed the same procedure, finding that crowdsourced interventions were as effective at changing beliefs in three of four topics. In only one case did an academic intervention perform better at correcting scientific misconceptions than a crowdsourced intervention.

In light of these results, we propose that arguments mined from online communities can be used to develop educational interventions. How might this process work? Consider the results in Experiment 2: We observed that an academic intervention containing an icon array (Lewandowski, et al., 2013) was more persuasive than a similar crowdsourced intervention that did not contain data visualization. This finding is consistent with a large body of research demonstrating that data visualizations can effectively communicate complex information (e.g., Fernandes, Walls, Munson, Hullman, & Kay, 2018). In future research, we propose that researchers could begin to develop an educational intervention by first turning to crowdsourced interventions that appear effective and then extending them based on well-established theoretical considerations. For instance, we found that a crowdsourced intervention about the repercussions of structural racism was much more effective than an academic intervention aimed at shifting people's implicit racial biases (Lai et al., 2014). One possibility, then, is that we could further improve the efficacy of this crowdsourced intervention by augmenting it with compelling visualizations. In this way, researchers would be able to develop interventions that have the twin virtues of demonstrating prior success in naturalistic environments and having strong empirical support from controlled laboratory studies.

Change My View is also not the only place researchers could crowdsource effective arguments; a web application could also assist in mining, for example, Facebook and Twitter for effective arguments. The tool we are proposing could take queries (e.g., topics for an intervention) and return effective arguments filtered by the searched terms. Such a system could allow researchers to not only crowdsource educational interventions more effectively, but also gain an understanding of how arguments are communicated and received among members of online communities.

A cursory look on Reddit, Twitter, and Facebook demonstrates that people naturally engage in (sometimes) persuasive argumentation. Here, we proposed that psychologists can mine this information to efficiently create educational interventions that are more likely to persuade people than the methods researchers currently use—crowdsourced interventions have the advantage of being vetted, so to speak, in naturalistic contexts. Two experiments provide support for this proposal. We observed that crowdsourced arguments were more effective or often as effective as academic interventions aimed at correcting misconceptions about several societally important topics.

Chapter 5

AI system for modeling causal beliefs from natural language

Introduction

Causal information facilitates learning (Holyoak & Cheng, 2011; Waldmann, 2007, 2017), and is crucial to how humans use and represent language (Mackie, 1980; Wolff et al., 2005; Lupyan, 2016). Causal relations are also ubiquitous in higher-level reasoning: they underlie our rich and flexible categories (S. A. Gelman & Legare, 2011), shape our explanatory preferences (Lombrozo & Vasilyeva, 2017), and structure our memories of events (Bower & Morrow, 1990).

Beliefs about causal relations can have pernicious outcomes as well. For example, beliefs that vaccines cause autism are central to antivaccination attitudes (Horne et al., 2015; D. Powell et al., 2022), and beliefs that liberal politicians have causal influence over the outcome of climate science research is a motivating consideration of climate change denialism (Cook & Lewandowsky, 2016). Because misinformation in online environments can spread rapidly to embolden these attitudes (J. H. Priniski et al., 2021; J. H. Priniski & Holyoak, 2022), new data science methods are necessary to fight these trends. However, data science algorithms generally struggle to advance a rigorous scientific understanding, as they provide correlational evidence that doesn't isolate cognitive mechanisms. Methodologists should aim to develop Natural Language Processing (NLP) algorithms that produce cognitively-plausible data representations that researchers can utilize to guide explanatory understanding and motivate future interventions.

Because causal relations are the backbone of most higher-level reasoning processes in humans, and are central to how we use language, developing systems that can isolate people's causal representations from

language data is a natural place to start. However, NLP has historically struggled to identify instances of *psychological causality* (what a speaker *thinks* causes what) (Dunietz et al., 2017). This is because the variety of ways people communicate causality is immensely vast (Talmy, 2000, 2011; Wolff, 2007), with the bulk of causal information latent in language inputs (Khoo et al., 2002; Blanco et al., 2008). Earlier methods relying on hand-labeling causal constructions to relate linguistic features to components of causal relations were extremely brittle and struggled to generalize to out-of-sample data (J. Yang et al., 2022). However, Large Language Models may help overcome this shortcoming as these models utilize rich semantic representations of language and subword tokenization that can help them identify instances of causal language not expressed in training (Devlin et al., 2018; Liu et al., 2019; Dunietz et al., 2017).

In addition to simply identifying instances of causal language, methods should account for the breadth of people's conceptual systems in which a causal claim is made. For example, causal beliefs are not held in isolation; rather, people hold rich interlocking belief systems spanning multiple topics that shape evidence integration (Quine & Ullian, 1978; J. H. Priniski & Holyoak, 2022; S. A. Gelman & Legare, 2011). Previous methods for producing representations of people's belief systems rely on experiments and are slow to develop and may not generalize outside of the lab (D. Powell, 2021; D. Powell et al., 2022). Because it is important to understand the full context of people's belief systems to reliably predict how people will interpret evidence and make decisions, tools must be designed that can identify the vast web of beliefs that people use to interpret information in the wild. NLP tools can leverage the proliferation of online social data to build these representations (Goldstone & Lupyan, 2016).

To this end, we introduce a pipeline based on the Large Language Model, RoBERTA-XL (Liu et al., 2019), that detects causal claims in text documents, and aggregates claims made across a corpus to produce a network of interlocking causal claims, or a *causal claim network* [1]. Causal claim networks can be used to approximate the breadth of content and beliefs about causal relations composing people's conceptual understanding of the entities and events discussed in a corpus. To guide future research, we host a pipeline that produces interactive visualizations of people's causal belief networks. Here, we demonstrate this software by building causal belief systems surrounding the Covid-19 vaccines from tweets.

How to build causal claim networks using our pipeline

The pipeline for extracting causal claim networks follows three main steps (see Figure 5.1). First, text documents are fed to a Large Language Model, a RoBERTa-XL transformer model (Liu et al., 2019), fine-tuned on causal language to extract causal claims made in a document (sentence to a paragraph in length).

[1] We host a no-code interface for interacting with the AI causal claim system at the following website: https://causal-claims.isi.edu/

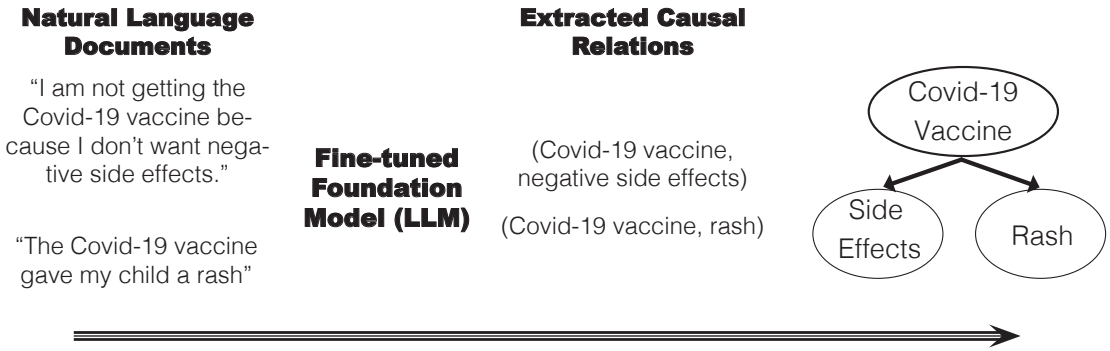

Figure 5.1: High-level schematic describing how text documents become a causal claim network. Raw text documents are first fed to a RoBERTa-XL transformer fine-tuned on causal language, which will return a list of tuples encoding expressed cause and effect relationships. These tuples will be co-referenced to produce a causal claim network.

Second, entities composing causal claims are clustered based on their embeddings to proxy *causal topics* (Grootendorst, 2022). Third, claims made across the corpus are co-referenced and strung together to make a network of cause and effect claims, and displayed in an interactive visualization.

We will now describe how a user could use our pipeline to build a causal claim network to visualize the causal claims made in a corpus of text documents. As shown in Figure 5.2, this follows two steps. First, a user uploads a .csv file containing the documents they wish to analyze. Documents should be a sentence to a paragraph in length, and can range from tweets, or journal entries, or news headlines. Next, the user selects which column in the dataframe contains the texts to be analyzed. A user can also specify if they want the pipeline to pre-process the documents and cluster the entities. It is worth noting that entity clustering works best when there are an abundance of causal claims about semantically distinct entities. If a user chooses to cluster claims and the pipeline doesn't produce an output, it doesn't mean that there are no causal claims present, but rather that there are no clear semantic clusters. In these cases, the users should deselected 'Cluster entities' and rerun the pipeline.

As seen in 5.2, we analyze a dataset of tweets about the Covid-19 vaccine with the file name *covid_tweets.csv*, and the column containing the tweet texts is titled *tweets*. We provide access to this dataset on the tool interface which can be downloaded to replicate this tutorial.

Once the document file is uploaded and the user presses submit, the job is queued and causal claims will be extracted. As shown in 5.3, a job status window will be populated and update the user on the the degree of completion. As a rough reference, extracting causal claims from about 6000 tweets takes about one minute to complete once the job begins.

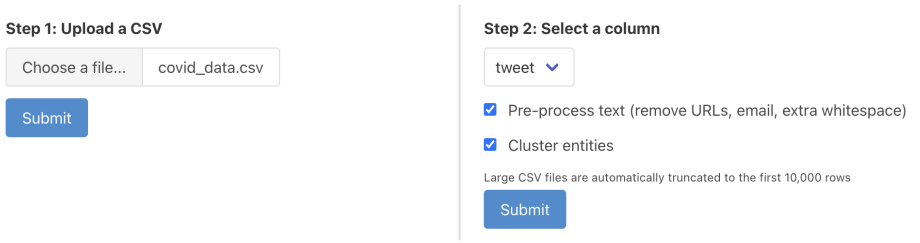

Figure 5.2: Uploading text data for causal claim analysis follows two steps. First select the .csv file you wish to analyze, then select which column in the dataframe contains the text documents to be analyzed.

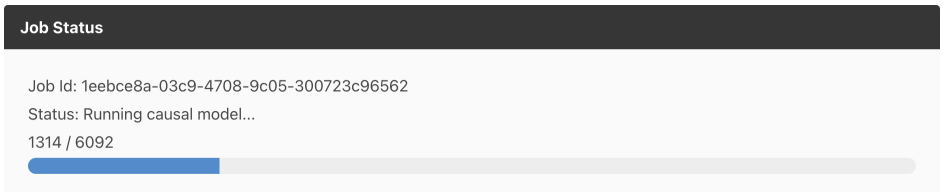

Figure 5.3: Job queue status window.

Once the job completes, the display will be populated with the causal claim network, like one in Figure 5.4. There are a few things worth highlighting here. First, each edge represents a single extracted causal claim in the corpus, and nodes are colored by their causal cluster, or topic (see Figure 5.5). Clusters proxy topics in the dataset, and can be interpreted as the central *causal topics* in the dataset. We describe how we calculate clusters in the next section.

The causal claim network produced by the pipeline is interactive. A user can click on an edge to see the document and extracted causal claim that constitutes that edge (see Figure 5.6). Furthermore, as shown in Figure 5.7, a user can simplify the network by selecting to collapse the edges between the nodes. The edge thickness is proportional to the number of documents between those two clusters. To facilitate downstream analysis of the causal claims (e.g., by analyzing sentiment or stance of causal claims), a user can download the edge list that produced the network as a .csv file. Columns in this .csv file include: cause word span, cause cluster, effect word span, effect cluster, text, and document id.

The pipeline allows users to specify different parameters for the model (see Figure 5.8). While the pipeline uncovers clusters automatically, users can specify the number of clusters to uncover in the corpus (this will equate to the number of nodes in the causal claim network). Users can also specify the N_gram range to be used during pre-processing, and can specify the number of top words used to describe each causal cluster/topic.

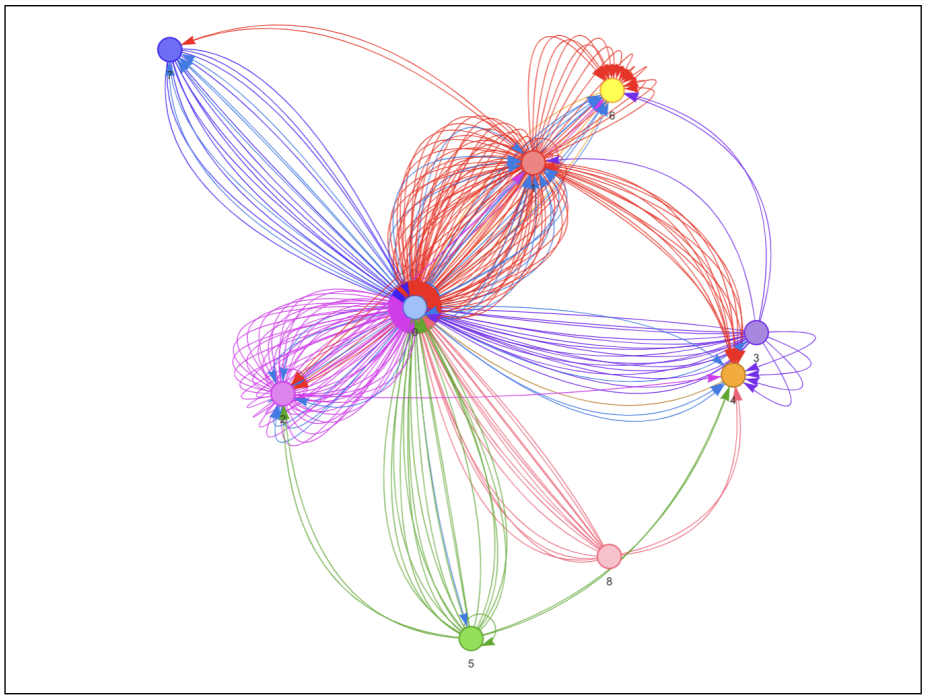

Figure 5.4: A causal claim network built from tweets about the Covid-19 vaccine. Individual nodes denote broad causal topics (i.e., clusters of cause and effect word spans based on their semantic embeddings), and edges signify a document contained a causal claim linking those two clusters.

Figure 5.5: Causal clusters, or causal topics, are shown to the right of the produced causal claim network. Each topic consists of a set of keywords that describes the cluster. Causal clusters proxy causal topics.

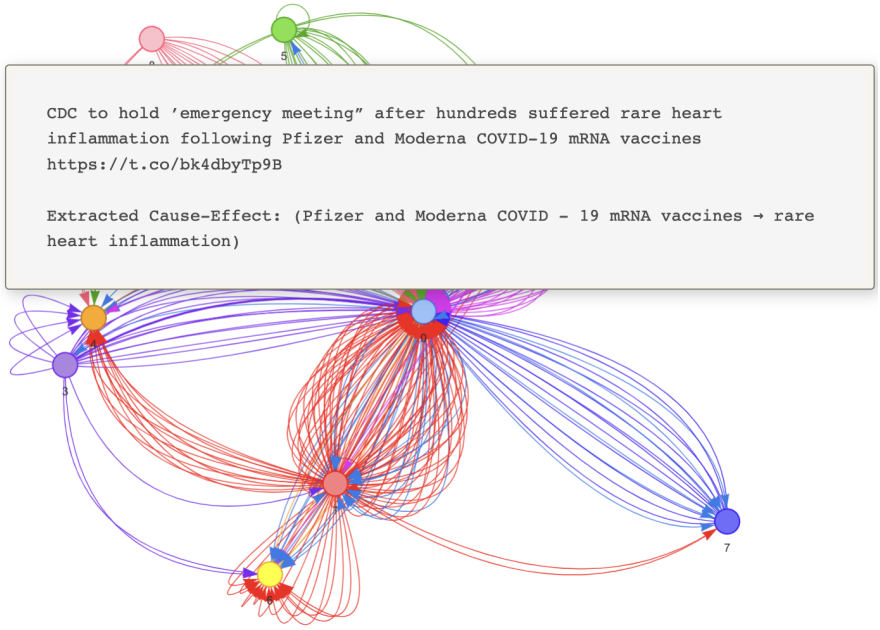

Figure 5.6: Hovering over an edge in the causal claim network displays the document and extracted causal claim that constitutes that edge. The document is shown at the top of the box, and the extracted cause claim is at the bottom.

What's happening under the hood

In this section, we describe how the pipeline build causal claim networks. This follows three steps: (1) causal claims are extracted using a RoBERTa-XL transformer model that identifies which words belong to cause and effect events (Z. Li et al., 2021), (2) claims are clustered based on their semantic topics, and (3) a causal claim network is built by stringing together the claims stated in the corpus.

Step 1: Extracting causal claims

Documents are first fed to a RoBERTa-XL transformer fine-tuned to identify cause-and-effect pairs of nominal (Hendrickx et al., 2010; Z. Li et al., 2021). The training set consists of $4,450$ sentences and contains $1,570$ causal relations, and the test set consists of 804 sentences with 296 causal relations. Following (Z. Li et al., 2021), we set up training as token classification task, where we utilize the before-inside-outside (BIO) labeling scheme to identify which words belong to a cause-span, effect-span, or embedded-causality-span (tokens belonging to a causal event in the middle of a causal chain). As seen in Table 5.1, RoBERTa-XL has a higher performance than previous state-of-the-art models on this task (Z. Li et al., 2021), and does better than the smaller transformer BERT (Devlin et al., 2018). We therefore used the RoBERTa-XL transformer in our pipeline.

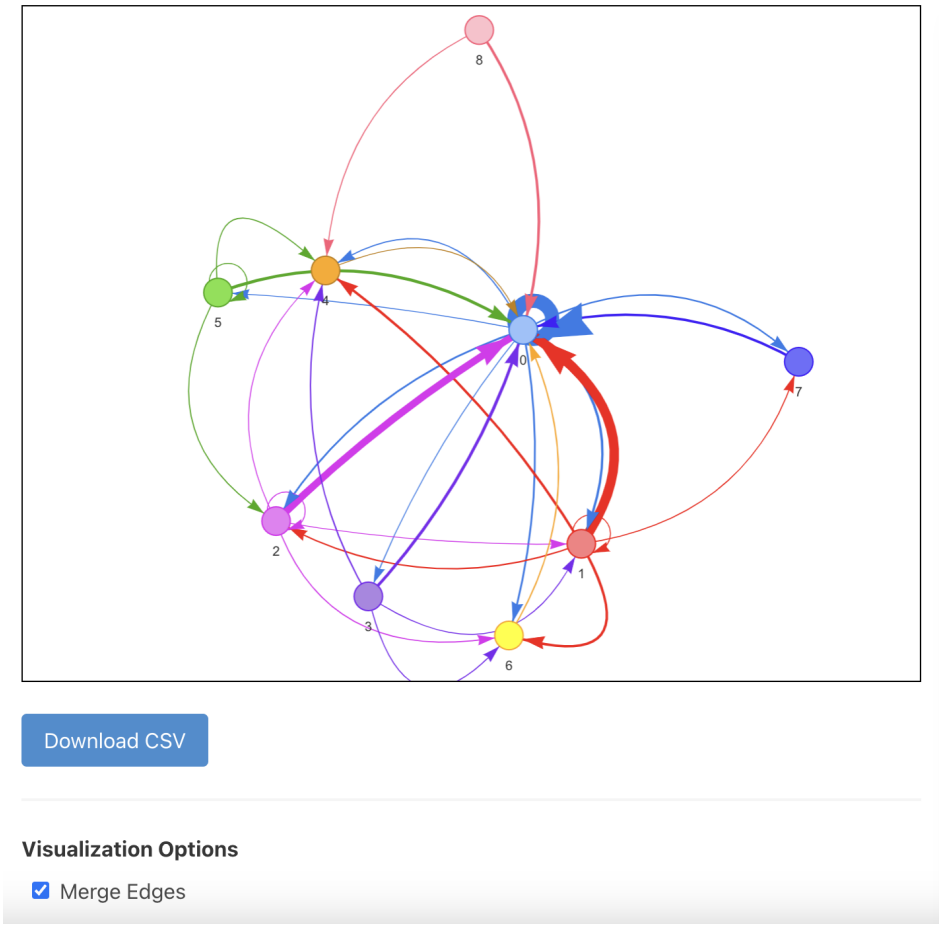

Figure 5.7: Causal claim network with merged edges, where edge weights equates to the number of documents linking two clusters. Merging edges is useful to quickly assess degree of linkage between causal clusters (nodes) in the network.

Once the sentences have been tagged using the causality tagging scheme, we run the tag2triplet algorithm as proposed Li et al., 2021 to extract the cause-effect tuples from the tagged sequence. The algorithm operates by first identifying the in-degree and out-degree of causality in the tagged sequence. Here, if the entity is labeled as a "cause" then the out-degree is incremented by 1, if the entity is labeled as an "effect" then the in-degree is incremented by 1, and if the entity is labeled as "embedded causality" then both the in-degree and out-degree are incremented by 1. The algorithm then tries to align the identified entities such that each entity that has an outgoing edge (i.e. the cause) is joined with the entity that has an incoming edge (i.e. the effect) while taking into consideration the distance between the entities in the document and whether they contain a coordinating conjunction.

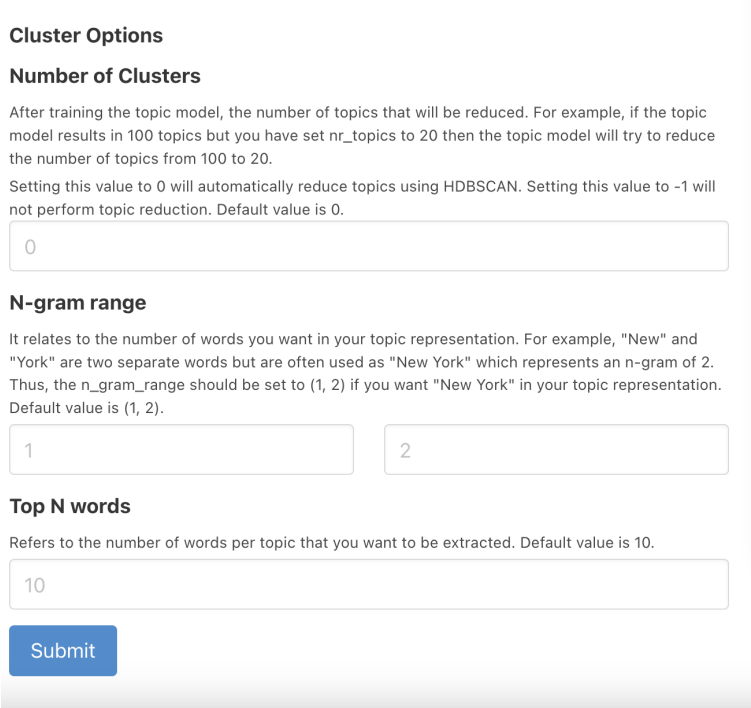

Figure 5.8: A user can specify parameters when running the pipeline to engage with exploratory data analysis. Users can pre-specify the number of clusters, the n-gram range used during processing, and set the number of words to describe each topic.

	Precision	Recall	F1
SCITE	0.833	0.858	0.845
BERT	0.824	0.858	0.841
RoBERTa-XL	**0.883**	**0.865**	**0.874**

Table 5.1: Model performance on the causal relation identification task (Hendrickx et al., 2010). The RoBERTa-XL model demonstrates increased performance over the smaller transformer BERT and previously reported state-of-the-art implementations (Z. Li et al., 2021)

Step 2: Finding causal topics via embedding clustering

Clusters of causes and effects proxy topics in the causal claim network. We cluster the embeddings of the nodes by extending tf-idf measure of embeddings (Grootendorst, 2022). This method was originally developed to clsuter BERT representations to uncover topics in a corpus, but we implemented the algorithm to cluster RoBERTa-XL embeddings. This allows users to assess latent structure in the causal claims expressed in a corpus, and simplifies the resulting causal claim graph by mapping semantically similar claims to a common node.

Constructing a causal claim network

Extracted cause-effect tuples serve as directed edges in the causal claim network, which are strung together across the corpus to form a causal claim network. Nodes are the identified cause and effect wordspans and weighted edges encode number of instances in the corpus where node i was said to cause node j. Edge direction encodes the direction of the causal relation and can be supplied with additional semantic content (e.g., relational vectors, sentiment).

Case study: Building a causal claim network about the Covid-19 vaccine from tweets

Data set of tweets

To test this pipeline, we build a causal claim network using a set of 6000 tweets about the COVID-19 vaccine (Poddar et al., 2022). The original dataset was curated by sub-setting a larger sample of tweets from before and after the release of the Covid-19 vaccines.

Pipeline results

The pipeline returns 408 extracted causal claims belonging to nine distinct clusters (see Figure 5.5). The clusters are, as one would expect, about the various Covid-19 vaccines and their anticipated consequences. By aggregating across the keywords for each cluster, we can define the set of causal topics returned by the pipeline. More specifically, cluster 0 contains keywords related to *Death*; 1: *Oxford vaccines*, 2: *Covid-19 pandemic*; 3: *Pfizer vaccine*, 4: *Side-effects*, 5: *Pfizer shot*, 6: *Immunity and antibodies*, 7: *Coronavirus*, and 8: *Covid vaccine*. As shown in Table 5.2, we see the that clusters are about a range of topics with varying semantics and valence, which suggests that the pipeline can help us understand the breadth of considerations guiding Twitter discussions about the Covid-19 vaccine.

Secondary analyses of extracted causal claims

By analyzing the causal claim .csv file that the pipeline makes available for download, we can explore how these causal clusters are linked to one another. For instance, as shown in 5.2, some of the clusters are more commonly composed of word spans denoting cause events, while others are more composed of word spans denoting effect events.

Cluster	Topic	Causes	Effects
0	Death	154	299
1	Oxford vaccine	108	15
2	Pandemic	56	16
3	Pfizer vaccine	25	1
4	Side-effects	1	25
5	Pfzier shot	22	2
6	Immunity	4	29
7	Coronavirus	13	8
8	Covid vaccine	17	0

Table 5.2: Number of identified word spans per each causal cluster. Topic label is determined by assessing the top keywords in each causal cluster. Each cluster has a different distribution of cause spans and effect spans.

Related work

While our approach is domain general, as in documents do not need to belong to a single issue or topic for the pipeline to work, we demonstrate the use of our pipeline modeling causal claims about the Covid-19 vaccine. Previous work has developed systems specifically designed to analyze claims about Covid-19. For instance, Li et al. 2022 built a system specifically designed to monitor claims made about Covid-19. This system identifies claims and arguments made in the corpus, and sources additional Wikidata information to put the claims in a richer content.

Mining causal claim networks requires isolating causal claims in text, which generally belong to arguments expressed in a text document: a reasoner often makes a causal claim when explaining a mechanism (Lombrozo & Vasilyeva, 2017) or making an argument. Claim detection is an active area of research (Palau & Moens, 2009; Goudas et al., 2014) as is detecting the components of arguments in text (Sardianos et al., 2015). Because causal reasoning is central to how people construct arguments (Abend et al., 2013), an understanding of how people posit causal claims in a corpus can shed light on the types of arguments people will endorse related to that issue.

The majority of claim detection algorithms works on the level of single documents, but approaches like Levy et al. 2014 propose corpus-wide claim detection. Our pipeline utilizes a mixing of the two: claims are detected within a document and then aggregated across the documents in the corpus to provide a corpus level representation.

Conclusions and future work

Interactive data visualization is an effective way for people to make sense of complex data (Janvrin et al., 2014) and can be an effective tool in guiding scientific thinking (Franconeri et al., 2021). Our pipeline is designed to help researchers explore the causal claims expressed in a corpus through interactive exploration.

There are therefore many applications of this tool to the study of human reasoning and belief change, and future work will test the efficacy of these use cases. For example, researchers from cognitive science have worked on developing methods to measure people's rich conceptual systems about vaccines (D. Powell, 2021). These methods often require developing surveys that measure people's attitudes towards a variety of related issues. Causal claim networks can give researchers a starting place to know what measures they should include in these surveys.

In future work, we will work on expanding the visualization tool to include features that allow for richer forms of interaction. For example, by allowing users to build sub-networks based on another data attribute (e.g., stance of the document, expressed sentiment), to allow for comparisons across networks. Related to this, future work will also develop quantitative measures of divergence across networks.

Chapter 6

Saving technology with cognitive and citizen science

> Defining the new environmental power as basically democratic is a feat that can obviously be credited to the Committee for National Morale of Gordon Allport, Margaret Mead, and Gregory Bateson in 1940...In the Committee for National Morale, one looked for an alternative to the authoritarian propaganda—a form of propaganda that would not just be democratic in its content, *but also in its very form.* One wondered what exercise of communication would not reproduce the unilateral submission of the receivers to the transmitter. How could one avoid one-way messages from a central post—be it at a radio mix, behind a camera, or at the editing table—which conditioned a passive, serial, robotized, fantasized subject of reception? In other words: *How does one create interactive propaganda?*
>
> Anonymous (trans Robert Hurley), *Conspiracist Manifesto*

Machines curate our news narratives, beliefs, and relationships to weaponize our minds against us. Social media and online social networks are optimized to overtake our livelihoods, because in the Attention Economy, *our* agency is *their* profit. Ubiquitous computing has toxified all corners of life and society, as influencers now infiltrate our national parks, cities, politics, and morality. Many helplessly accept this alienating and accelerating reality as the "new normal." However, technology developed with humanistic principles and cognitive science in mind can restore our vitality. Information embedded in social media, outputs from generative AI, and the internet interact with psychological mechanisms to impact society, others, and political action. Generative AI produces synthetic media that blend the real and the unreal, bringing us into a "post-truth era" (Lewandowsky et al., 2017). How can we delineate fact from fiction and rational discourse from identity politics in continually imbalanced and radically curated spaces? How do psychological and algorithmic processes interface when machines manipulate the cause-effect relationships in our narratives to systematically curate our beliefs about politics, morality, and others? What can you do to overcome these terrifying trends?

Cults, conspiracies, chaos: how narrative optimization radicalizes communities

Conspiracy theories support *subjective* epistemologies (ways of knowing the world) that allow oftentimes disenfranchised individuals to interpret complex structural mechanisms in society and interpersonal group dynamics. Conspiracies do exist, and theories about them can be correct. For example, the Tuskegee Study conducted by the U.S. Public Health Service between 1932 and 1972, tracked the progression of untreated syphilis in African American men without their informed consent. Distrust towards medical organizations stemming from this treatment exists today, which affected many Black individuals' trust in recommended public health measures. MK-Ultra is another classic example of a true conspiracy. Both serve to highlight how it is difficult to delineate reality from falsehood in many of these situations due to lack of transparency, abstract causal relations, and complex environments. Most importantly, increased sensitivity is needed on the part of behavioral scientists when understanding why individuals endorse conspiracies and related misconceptions. We must understand the *context* in which these beliefs emerge, and not characterize the adopters as lacking cognitive abilities/effort to discern fact from falsehood. Still, we take seriously when conspiracy theories hamper collective decision making and political action.

At times of political conflict, conspiracies seep in to help interpret and explain hidden causal variables generating the world's events. Conspiracy thinking is an interesting case in human reasoning because not only does it challenge theories that humans are *rational* beings, but it also make us take more seriously what constitutes belief and evidence in political decision-making. For example, does supporting the QAnon conspiracy literally imply that one believes Hillary Clinton is *actually* eating and raping babies, or does supporting the narrative signal support for a general (metaphorical) explanation of political disenfranchisement? To the supporters of QAnon and related movements, what is most important is working together to stop a small elite from overtaking the weak/many ("into the storm"), not establishing the veracity of a certain network of claims.

Conspiracy thinking can be understood as a rational process given contextual information such as a reasoner's goals and causal context. Conspiracy thinking can be *instrumental* in achieving certain objectives at the level of the individual reasoner: it may constitute practically rational decision-making in certain contexts. The second chapter of my dissertation explain how contextual information including utility information may impact people's first-order beliefs about causal and probabilistic information. Utilities can reorient people's values to cohere with alternative facts (e.g., a person's online network supports an alternate interpretation of climate science). For example, distrust in an organizing body may be warranted, and the complete causal web generating nefarious behavior may be too complex to represent. Thus, a conspiracy theory abstracts

away this information to sustain an appropriate attitude, even if the conspiracy theory may not be causally true.

However, in online environments, where group identity and evidence are radically curated, conspiracy theories embedded is political disinformation can be used to radicalize individual's belief systems, which can produce dangerous and even violent behaviors. In these cases, conspiracy narratives effectively undermine legitimacy of their targets by emboldening trust in the support targets (e.g., social entities: political candidates, scientists). This asymmetry between trust and distrust results in conspiracy thinkers interpreting socially supported evidence without skepticism, which can produce hypercohernet belief systems. However, in online environments, where group identity and information are radically curated, conspiracy theories embedded in political disinformation can *radicalize* individuals' belief systems, potentially leading to dangerous and even violent behaviors. In these cases, conspiracy narratives effectively undermine the legitimacy of their targets by emboldening trust in the supporting targets (social entities such as political candidates, scientists). This asymmetry in trust and distrust results in conspiracy thinkers interpreting evidence with local social support (local coherence) without skepticism, producing beliefs in hypercoherent belief systems; even if these beliefs are incoherent with the global set of facts and relations.

Distinguishing conspiracies from conspiracy theories

Conspiracies are networks of agents working together to uphold one's in-group over an out-group. They operate covertly and exist within social networks "out in the real world." While they can be true, many conspiracies are constructed to sow distrust in targeted out-groups, often to sustain a political ideology. Conspiracy theories are narratives that explain situations where a group is covertly acting (causal relationships must be inferred to explain data). Conspiracy narratives are given causal power in that they are used to explain information about the world. For example, climate change skeptics might reject scientific information because they hold ideological beliefs that scientists are part of a broader conspiracy working to undermine the hopes and dreams of "true America".

Both conservative and liberal attitudes are susceptible to radicalization through conspiracy theories, and our increasing ideological polarization and identity politics are being sustained by our media environments (both social and televised). Online radicalization occurs through the tuning of people's exposed information spaces, which compose causal narratives for political events. These multi-media narratives composed of hashtags, arrays of text, images, videos encode rich distributed situation models, which include key causal, agential, and relational information (Zwaan, 2022; Zwaan, Magliano, & Graesser, 1995). The mental representations encoding the situations explained by narratives can produce hypercoherent belief systems, such that all information is construed as coherent with a small set of ideological beliefs (J. H. Priniski et

al., 2021). Online media environments foster hypercoherent belief systems through reinforcing information streams (news feeds, friends' content). The reinforcement often comes from social information (e.g., retweets, what friends say about a piece of content), which can lead to echo chambers and radical beliefs, and steer how people generalize learnt information from incoming data in the future.

Generalization of previously learned information to novel situations is a hallmark of adaptive learning (Mednick & Freedman, 1960). In the context of belief formation, new beliefs also tend to generalize from, or cohere with, features of prior beliefs (Lewandowsky, Oberauer, & Gignac, 2013; Homer-Dixon et al., 2013). Indeed, a coherence mechanism has been shown to be central to various cognitive processes, from visual perception (Yuille & Grzywacz, 1988) to moral reasoning (Holyoak & Powell, 2016). According to explanatory coherence theory, beliefs are often formed on the basis of congruence with prior beliefs, insofar as the acceptance of a new belief increases the explanatory coherence of the belief network (Findlay & Thagard, 2014). Beliefs may therefore be adopted if they fit the explanatory model generated from one's prior beliefs.

But what happens when new evidence can be construed as coherent with one's prior beliefs, regardless of its veridicality? Recent work has shown that people who engage in conspiratorial thinking tend to attribute more control and structure to the world than is plausible. Conspiratorial thinkers do not typically believe in just a single conspiracy theory, but rather clusters of them (Van Harreveld et al., 2014). For example, Lewandowsky, Oberauer, & Gignac (2013) found that people's propensity to believe that NASA faked the moon-landing predicted their tendency to believe that climate change was a hoax. This association arises because conspiracy thinkers are likely to endorse completely novel conspiracies that share common conspiracy themes, thus viewing logically disjointed narratives as mutually coherent (e.g., NASA/government conspiracy $\rightarrow$ fake moon landing; climate scientists conspiracy $\rightarrow$ climate science is fake). This evidence suggests that conspiracy thinkers may readily bind new information with their conspiratorial view of the world through a maladaptive level of coherence—hypercoherence. Hypercoherence combines top-down priors based on broad core attitudes (e.g., distrust of government and scientific elites), coupled with bottom-up "data" based on the opinions of fellow believers that echo on social media. Where conspiracy thinkers go one, they go all by attributing a vast network of complex narratives to a single causal source (Saltman, 2020).

Hypercoherence may be especially easy to achieve when information is consumed within online communities where both information and social identity are radically curated. Conspiracies such as QAnon may be a natural consequence of a social media environment that: (1) prioritizes false information over verifiable information, and (2) allows for the easy and rapid formation of echochambers, or pockets of online communities that share and consume nearly identical, belief-confirming information (Sasahara et al., 2021b). Once misinformation is introduced that coheres with the narrative of a particular echochamber, it may foster the generation of additional content by simultaneously adding to the coherence of the community's narrative

while reducing its standard of plausibility. Misinformation may therefore gradually reconfigure a person's belief network toward stronger degrees of coherence, making it more capable of binding disparate and implausible beliefs. The result is belief in conspiracies that cover a wide range of narrative clusters. Providing first-order causal evidence to overcome webs of conspiratorial beliefs often doesn't work to revise attitudes because any incoming information is constured as coherent with the underlying belief in the conspiracy. For example, Wood et al. (2012) found that conspiracy theorists believed that Osama Bin Laden was both dead and a live, when that world state was evidence for Obama not having killed him and having lied about it. In the second chapter of my dissertation I discussed how social reward information to revise first-order credences about evidence. In the case of conspriacy thinking, second-order reward based interventions may be necessary to overcome these deeply-entrenched attitudes.

Testing reward-based interventions in online network experiments

Motivated reasoning, as defined in the second chapter of my dissertation, is a situation where evidence representations are biased by directional-goals. Researchers often appear to assume evidence representations must be distorted by directional-goals when participants' belief reports do not shift in ways they seemingly ought to following intervention. On its face, this suggestion is plausible because across many experiments, participants appear to hold onto their beliefs in the face of evidence contrary to what they think. However, this research often leaves many of the core representations unmeasured, and is thus unable to distinguish between situations where people may have properly integrated the evidence *before* taking the utility of holding that belief into account from situations in which motivations directly impacts how people sample from the evidence. These are distinct situations that implicate different sets of cognitive processes, which may demand altogether different kinds of interventions.

The distinction between practically rational behavior and motivated reasoning is not merely a semantic issue but impacts how we may develop interventions that correct misconceptions. For example, it may well be that someone respects the norms of practical reason and nonetheless believes that climate change is a hoax. Thus, even if a person's reasoning is rational under some set of norms, there could be a need for intervention as a matter of public policy. As a consequence, effective interventions may take very different forms when misconceptions are due to defects in one's inferential machinery (i.e., so called *first-order interventions'*) versus when misconceptions arise because of second-order influences from directional goals (i.e., so called *second-order interventions*). The goal of this section is to explore how the Bayesian decision-theoretic framework may be applied to distinguish between these two cases in order to guide intervention development.

We distinguish between two classes of interventions, first-order and second-order interventions. *First-*

order interventions aim to revise inputs to the belief-decision process, typically (but not only) focusing on doxastic factors. For example, they could take the form of a nudge to encourage participants to more accurately encode information (Pennycook & Rand, 2019; Bago et al., 2020), or by increasing trust in established bodies of knowledge (Van der Linden et al., 2017, 2019). First-order interventions operate on first-order beliefs and aim to correct the alignment between internal credences and the way the world really is. However, they need not focus exclusively on doxastic factors. For instance, interventions highlighting utilities may incentivize reasoners to use accuracy goals when interpreting evidence during reasoning (Kunda, 1990). This is a case where utility-focused information is designed to revise inputs to a doxastic representation.

Second-order interventions, on the other hand, target second-order beliefs (and consequently, belief reports), often by revising people's perceived utilities associated with holding a belief (although second-order interventions need not always operate on utilities). For example, education researchers have found that stating the real-world utility associated with learning a piece of information is a strong driver of students' eagerness to learn and remember that information (Soicher & Becker-Blease, 2020; Hulleman & Harackiewicz, 2009). Second-order, utility-focused interventions highlight the empirical consequences of a reasoner's decision to hold a belief, and thus are designed to realign second-order beliefs *after* doxastic information has been integrated. Second-order interventions aim to make salient how decisions to hold certain beliefs can facilitate achieving one's goals. First-order interventions are tested more frequently than second-order interventions. The research paradigms most typically used in social psychology partially explain this trend: Psychologists developing interventions often only measure first-order doxastic representations, focusing on how motivation biases sampling during the construction of an initial credence. However, in cases where first-order interventions fail to shift people's beliefs, it is nonetheless possible there are other means of shifting people's beliefs (J. H. Priniski & Horne, 2019). For instance, a more effective intervention might be developed by appreciating the role second-order utility plays in affecting belief reports (D. Kahan & Braman, 2006; D. Kahan et al., 2017).

To illustrate how the proposed computational framework can assist researchers in developing second-order, utility-value interventions, I will discuss two frequently-cited first-order interventions, *accuracy nudges* and *inoculation tactics*, which intervene on the separate processes of evidence representation and evidence integration, respectively. I will then discuss how they may be adapted to target second-order, directional-goals.

Example 1: Enhancing Accuracy-nudge paradigms with second-order utility information

Nudges are a popular approach in the decision sciences literature for influencing people's behavior and beliefs (Thaler & Sunstein, 2009). Recently, many researchers have been interested in applying nudges to mitigate

the spread and uptake of misinformation (e.g. Pennycook & Rand, 2019; Fazio, 2020; Pennycook et al., 2020; Pennycook & Rand, 2021; Fazio et al., 2019; Pennycook & Rand, 2020; Bronstein et al., 2019; Bago et al., 2020; Pennycook et al., 2021). So called accuracy-nudges are designed to get people to think more carefully about the accuracy of a piece of information before engaging with it (e.g., updating their beliefs, sharing the information on social media). The central idea is that a reasoner's reflective and deliberative System 2 processes are better suited at detecting a piece of information's veracity than their effortless, intuitive System 1 processes (Bago & De Neys, 2017; Pennycook & Rand, 2019; Bago & De Neys, 2019; Thompson et al., 2011). Accuracy nudges are designed to subtly nudge people to use System 2 processing when interpreting potentially false information.

There are various ways these nudges may be instantiated. For example, they could require participants to explain why a headline is true or false before gathering accuracy perceptions (Fazio, 2020). Other more implicit tacts have been taken, like varying the time participants are allotted to decide if they would share a piece of information on social media (Bago et al., 2020). Some researchers have used direct methods to nudge people towards accuracy, like reminding participants at the beginning of a misinformation detection task to consider accuracy (Pennycook et al., 2020). The authors of many of these papers report that these effects can inform us about the underlying mechanics of reasoning and belief revision (Pennycook & Rand, 2019; Pennycook et al., 2021; Brashier et al., 2021). In particular, a consistent claim across many of these studies is that because political affiliation does not correlate with participants' ability to detect or share misinformation, and accuracy nudges do not interact with political affiliation, reasoners are lazy rather than motivated when engaging with potentially false information (e.g., Pennycook & Rand, 2019; Pennycook et al., 2021, 2020).[1]

However, there are a few things worth considering when interpreting the results and assessing the viability of accuracy-nudge paradigm. The first thing to consider is that the effects of accuracy nudges (and nudges more generally) are small (e.g., correlations between responses on the Cognitive Reflection Task and ability to detect false information range from 0.1 to about 0.25; Pennycook & Rand, 2019). A recent meta-analysis found that nudge-based approaches are likely to exert extremely small effects once publication bias is accounted for (Maier et al., 2022). It is an empirical question, but the magnitude of these effect sizes may also suggest that the source of the tendency to share misinformation is not predominantly a function of a problem with people's doxastic representations. Instead, it is possible practically rational considerations are shaping people's decisions. Thus, it might be fruitful to use other equally-scalable intervention tactics that leverage the perceived utility of a belief report.

[1]This point is contentious, however (Roozenbeek et al., 2021). For instance, a motivation for accuracy is a motivation nonetheless (Kunda, 1990).

The second concern focuses on researchers' failure to measure constructs that resolve how exactly accuracy-nudges are supposed to work, and to what extent they will generalize beyond the domains they have investigated. A considerable body of research on accuracy nudges analyzes behavior involving news headlines, and consequently aims to measure representations relevant to how news headlines are interpreted and shared (e.g., Brashier et al., 2021; Pennycook & Rand, 2019, 2020; Pennycook et al., 2021, 2020). For instance, while these studies often include a measure of political ideology (which can be seen as a broad correlate of prior beliefs), they also lack precise measurement of a participant's priors for a specific belief at hand (but of course, there are exceptions; see Pennycook et al., 2020). A reasoner's prior belief about the information in a specific news headline should have considerable influence not only on their perceptions of accuracy, but also on their intention to share that information.[2] There is sure to be substantial item-level variation, because the perception of a news headline will interact with a participant's unique prior belief about that issue. This is not accounted for in many of the current designs and statistical models in the accuracy nudge literature. Further, most authors have failed to measure utility information at all. For example, group-based utility information may change how news headlines are interpreted (e.g., how do members of my group interact with this type of information; D. Kahan & Braman, 2006). After all, the very reason *sharing* is of interest to researchers is because of the assumption that information shared by one's in-group will beget further sharing. Consequently, it is possible that deploying utility-value information as an intervention could increase the efficacy of nudges that have predominantly focused on accuracy, but have not yielded particularly large behavioral effects (Roozenbeek et al., 2021; Pennycook & Rand, 2019; Bago & De Neys, 2017; Pennycook et al., 2020). To make this more explicit, I will now discuss how researchers could design an experiment that incorporates utility information into an accuracy-nudge paradigm.

Incorporating second-order information in accuracy nudge paradigms

People often read misinformation in online environments where interactions between people's social networks can serve as a rich source of second-order utility information. Consequently, making salient how a piece of information was engaged with in one's social networks could help the reader assess the social costs associated with sharing a piece of information. For example, researchers could manipulate the social consequences associated with sharing a piece of potentially false information, which could be in the form of predicted engagement with a piece of information. In this imagined experiment, a high-group support condition could inform a participant that in-group members are more likely to share and favorite a headline based on past sharing behavior of friends in their network. Researchers can then quantify these effects by varying

[2]Although perceived accuracy and intention to share may be weakly correlated in some situations, reasoners might still willingly share misinformation (e.g., X. Chen et al., 2015; Laato et al., 2020; Madrid-Morales et al., 2020; X. Chen, 2016; Altay et al., 2022)

engagement metrics to more completely define a utility space (e.g., one retweet and one "like" is a smaller group-based reward than 50 retweets, 50 likes, and five new followers). Researchers could then compare participants' intentions to share and perceptions of accuracy to participants in a low-group support condition – a condition in which participants are told that in-group members are unlikely to share and favorite a given headline, or are more likely to unfollow, mute, or block people who have shared similar content. To control for the influence of trust (a first-order, doxastic consideration), a researcher could also measure how engagement metric information shifts people's evaluation of the evidence itself, allowing them to observe the unique effects of second-order, utility-value information.

Misinformation may encourage misperceptions of consensus – false beliefs about what one's group uniformly agrees to believe. These misperceptions can influence people to report beliefs that are possibly out of sync with their internal credences, in turn establishing a social norm that prevents people from updating their reported beliefs about polarizing social issues. For instance, nearly three-fourths of Americans support climate change mitigation policies, but the general public believes that the proportion is closer to one-third (Sparkman et al., 2022). Many people—specifically conservatives—may therefore believe that climate change is real and would support policies designed to deal with it, but fear potential social consequences of reporting so. Assuring conservatives that there would be little-to-no social consequence in reporting support for climate change policies (because, in fact, the majority of their party also believes climate change is real and would support mitigation policies), can be one way to realign people's first-order and second-order beliefs (e.g., Lewandowsky & van der Linden, 2022; Constantino et al., 2022).

We can now see a possible benefit of developing a computational framework for distinguishing practically rational behavior from motivated reasoning: it allows us to assess what's gone unmeasured in current studies and thus suggests a way forward in the creation of new intervention strategies. The benefits of developing our framework are not unique to research on nudges. We'll now consider a second example.

Example 2: Inoculating gateway beliefs with preventative arguments

Most people are not trained to interpret scientific evidence. People's attitudes towards topics in science hinge on factors most scientists do not view as objectively relevant to how we should understand the evidential quality of a piece of scientific research. For example, it has been established that perceptions of scientific consensus shape how people evaluate a piece of science (Van der Linden et al., 2019; Cook & Lewandowsky, 2016; Roozenbeek et al., 2022; Imundo & Rapp, 2022). "Gateway beliefs" to climate attitudes and perceptions of consensus play a pivotal role in shaping how evidence for climate change is integrated into people's belief systems. Researchers have found that Bayesian networks of climate change attitudes that include a "perception of consensus" node (such that high-trust in consensus among climate scientists predicts trusting

scientific evidence for climate change) can simulate belief polarization following exposure to evidence of climate change. Because Bayesian networks are able to capture polarization effects, researchers have concluded that the cognitive processes underlying these effects are rational (in that they do not violate the axioms of probability theory, e.g., Cook & Lewandowsky, 2016). These results suggest that climate change misconceptions persist because of inputs to people's inferential machinery rather than the operations of the machinery itself. Consistent with this claim, researchers have successfully strengthened beliefs in climate attitudes by presenting participants with "consensus arguments" that change people's gateway beliefs.

Misinformation, particularly about science, often targets gateway beliefs by casting doubt on established sources of knowledge (Pierre, 2020; Stanley, 2015; Lewandowsky et al., 2019; Enders et al., 2020). The effects of misinformation which undercut established sources of knowledge can have a continued influence on people's beliefs, even after correction (Lewandowsky et al., 2012). Researchers have therefore also argued for preemptive approaches to countering misinformation by designing interventions that establish "defensive priors" by exposing people to weakened forms of misinformation – a kind of "belief vaccine" that inoculates people against misinformation (Van der Linden et al., 2017; Traberg et al., 2022; Vivion et al., 2022; Compton et al., 2021; McGuire, 1964). For example, an inoculation intervention for climate change could contain information that preemptively refutes common arguments claiming that there is no consensus among climate scientists about the realities of human-caused climate change (Traberg et al., 2022; Pilditch et al., 2022).

In the cases described above, the interventions that have been developed have focused on first-order, typically doxastic, information to correct misconceptions (or strengthen otherwise uncertain beliefs). But there is no doubt that misconceptions about climate change persist (e.g., Lewandowsky et al., 2012; Brenan & Saad, 2018), so it is worth considering how researchers can improve on existing messaging by going beyond intervening exclusively on first-order, doxastic considerations. Below, I consider how incorporating second-order, utility-value information could augment existing interventions.

Second-order inoculation to guard against conspiracy narratives

Conspiracy theories are a common type of misinformation targeting beliefs about science. They discredit scientific findings and recommendations by linking scientists to malevolent (oftentimes political) characters (Bodner et al., 2020). During the COVID-19 pandemic, for example, conspiracies quickly emerged online – generally claiming that public health experts were conspiring alongside left-wing politicians and tech elites to exert control over the public (Douglas, 2021). Conspiracy narratives are integral to climate change denialism as well. These narratives, which are often echoed on conservative news outlets, aim to construe climate science as politically motivated individuals working with left-wing politicians to limit free-market capitalism (Uscinski et al., 2017; Hornsey et al., 2018).

One way to build a second-order intervention to guard against the effects of conspiracy theories is to consider how conspiracy theories exploit group-based biases. For instance, conspiracy narratives are often structured with an in-group bias, in that they are designed to negatively construe events (e.g., public health measures during a pandemic) by positing that they are against the goals of the group (e.g., out-group totalitarian control). This is why conspiracy theories were so effective at inflaming political tensions during the COVID-19 pandemic (Douglas, 2021; Jolley et al., 2018). Conspiracies leveraging in-group bias also allow for people to rationally reinterpret contradictory evidence by dismissing the source of evidence as unreliable (Jern et al., 2014; Gershman, 2018).

To explore how we can develop a second-order *inoculation* intervention, we will focus on the case of vaccination attitudes, particularly in situations where we aim to increase vaccine uptake during disease outbreaks. Researchers have found that these situations are often accompanied by misinformation (Midgon, 2022). Misinformation inoculation follows two steps: forewarning of tactics and refutation of tactics. In the forewarning phase, researchers could provide participants with information about conspiracy theories and why they're used. For example, researchers could inform participants that conspiracy theories are designed to facilitate trust in certain people. Researchers can then provide participants with examples of how conspiracy theories have been used for harm in the past (e.g., Nazi Germany, Rwandan Genocide). Thereafter, researchers can provide participants with concrete examples of how conspiracy theories were used to undermine public health efforts during the COVID-19 pandemic (e.g., how vaccines were planted with tracking devices). Researchers can then survey people's beliefs about a novel, ongoing outbreak (e.g., Monkeypox) and measure their susceptibility to the narrative offered by emerging conspiracy theories. This will allow them to test the effects of a *norms-based message* to promote more positive vaccine attitudes (e.g., see Lewandowsky & van der Linden, 2022; Constantino et al., 2022).

Inoculation tactics usually adopt a refutation step where they logically or statistically undercut a misinformation claim researchers predict participants will have been exposed to. However, refutation interventions target first-order beliefs when second-order beliefs may play a larger role than researchers have typically assumed (Jachimowicz et al., 2018). One way researchers could take second-order beliefs into account is by providing participants with icon arrays of vaccination rates of a participant's community or a participant's political party, which would establish the norms among a participant's in-group. Indeed, it's been found that participants view anecdotal information about people as similar to themselves influences vaccine attitudes (Horne et al., 2015). Similar effects have been observed in other domains: For example, Jachimowicz and colleagues (2018) observed that second-order normative beliefs predicted participants curbing their energy use even though participants' first-order credences did not. Under the proposed computational framework, we would predict that in domains where people perceive the evidence is uncertain but the utilities favor

one theory over the other, intervening on second-order, utility oriented beliefs may be a more effective intervention tactic than interventions exclusively focused on first-order, doxastic features.

Develop more complex interventions or transform more radically?

> Communication between politically authorized contracting agencies and objectively knowledgeable and competent scientists at major research and consulting organizations marks the critical zone of the translation of practical questions into scientifically formulated questions and the translation of scientific information back into answers to practical questions. Of course this statement does not really capture the dialectic of the process. The Heidelberg Research Project in Systems Analysis has reported a revealing example. The headquarters of the U.S. Air Force using experienced contact men presents a roughly outlined problem of military technology or organization to the program department of a research and consulting organization. . . During the research itself, information is exchanged at all levels, from the president of the research organization down to the technician, with the corresponding personnel at the contracting institution. Communication may not end until the solution of the problem has basically been found, for only when the solution can be foreseen in principle is the goal of the project ultimately defined.
>
> Jurgen Habermas, *Toward a Rational Society: Student protest, science, and politics*

> One was always waiting for the man. There was always a chance. At any moment the leader might arise; the man of genius, in politics as in anything else. Probably he will be extremely disagreeable to us old fogies, thought Mr. Bankes, doing his best to make allowances, for he knew by some curious physical sensation, as of nerves erect in his spine, that he was jealous, for himself partly, partly more probably for his work, for his point of view, for his science; and therefore he was not entirely open-minded or altogether fair, for Mr. Tansley seemed to be saying. You have wasted your lives. You are all of you wrong. Poor old fogies, you're hopelessly behind the times. He seemed to be rather cocksure, this young man; and his manners were bad. But Mr. Bankes bade himself observe, he had courage; he had ability; he was extremely well up in the facts.
>
> Virginia Woolf, *To the Lighthouse*

Our rapid immersion into online life has made many ill. By producing, personalizing, and disseminating disorienting imagery, digital technologies comodify the minds and hearts of the masses with nauseating precision and scale. Generative artificial intelligence, social media, and digital news feeds establish narratives that divide and antagonize us, while those at the helm of these technologies conquer the final frontiers of our interior lives, social relations, earth, and cosmos. Generative AI will homogenize and eradicate life, not through some stupid "singularity" event, but through the production of synthetic media designed to drown us in alternative realities and images.

Antagonizing narratives communicate images (i.e., categories, causal relationships, ideologies) that reinforce a separation within our interiors; seeking them sustains a loss of equilibrium that necessitates the use of force to resolve. In the pursuit of security, we strive to balance ourselves, groups, and states by warding-off current images only to embrace the next one: a more niche aesthetic, a more radical political ideology, a more powerful weapon (whether it's a word or a gun), a higher paying job, gallery representation, a *Nature* publication, an influencer girlfriend. Images motivate us to take control of the world and others to possess the stillness of a secure interior. However, each image we desire does not settle us and only perpetuates our

separation.

When we identify with an image and want to obtain it as our own, we are insecure and desire to secure ourselves through embodying the sought after image. Obtaining an image requires calculated effort, and securing it can only happen in the future or past. Because there is a distance between now and the desired future or longed-for past, a time-interval separates oneself from all images. Thought works to collapse the interval between the present and desired image by optimizing actions (life) through efficient use of images and language (communication of images embedded in narrative structures). Images compose our memory and influence our beliefs as our memory and beliefs influence the images we pursue. Seeking images that optimize our lives are therefore bound to optimize our societies.

Images measure the world by carving it into categories and causal relationships. Categories partition the world into things. Saying something belongs to a certain category requires saying other things do not belong to that category (barring pragmatically trivial categories like 'the set of all things'). Causal relationships represent how categories should behave and what you can expect from their behavior. They also hold between categories, which define their boundaries of application. Causal relationships describe how instances of some category bring about instances of another category by describing how categories change *in time*. Categories and causal relations are two sides of the same coin; they are the media of our memory.

Grouping things according to what they are and what they are not is a necessary consequence of using categories and causal relationships to carve up the world. After all, a category would be of no use if it failed to make any discrimination between members and non-members. This discriminative capacity of categories and, by extension, the causal relationships they occupy is what gives images their *separating force*. If our image of the present moment diverges from the image we desire, we orient our attention to align our actions to embody the sought after image. Desiring to embody an image is synonymous to seeking category membership, to be one with the category. Categories provide the rigid boundaries that we hope to feel secure within, as they define for us and others who we are and how we should behave.

The separating force of images propels bodies through space and time, and situates the self over the other. Shared category membership brings things closer together; being on different sides of a boundary makes things move against one another (just think about the dynamics of nationalism and borders; the modular family structure and subdivision of space in the American suburbs). Categories structure our relationships across all dimensions: the interior and exterior world; the self and the other; the family unit; the nation-state; one's social categories (race, gender, political affiliation). Categories also impose morality on its members and non-members: the former constitute the good and the latter the bad. The morality that categories impose functions to justify the use of force to subjugate and control out-group members (at least those caste as out-group to the hegemony). Imperialism sustains itself through people's categories by

developing social networks and systems, actors, and technologies that produce images that separate people, and motivate the use of force to control those that are outside the boundaries of one's categories.

Causal relationships and, by extension, the categories that they relate to can facilitate capitalist activity by enabling effective intervention, possession, and control. We build causal models of our environment to command the world and other people to sustain our own vitality. All biological organisms represent the world to some degree to sustain their life, but the categories and causal relationships humans construct are much more flexible than those used by non-human animals. This flexibility makes the images that orient our attention and constitute our desires more open to revision, reinterpretation, and recomposition. The human mind has a certain proclivity for image-seeking, which is exploited by capitalist imagery to open our identities, relationships, and groups to the pressures of capitalist production. Crucially, the flexibility of our representations ensures that different people are oriented toward different, often conflicting goals. This is why strongly-held beliefs about conflicting images lead to depersonalization and polarization, and violence is justified by the state as a moral imperative. To escape the dissonance of conflicting images, the mind searches for a new image to act as a cure; often in the form of increased technical expertise, a more optimized police state, or a more radical political ideology. However, it is through the perpetual separation brought on by image seeking that conceptual violence manifests in the world; we will continue to fall prey to isolation as long as we depend on others to provide us cures. This is a fundamental consequence of existing in the world for the purpose to sustain one's images.

The separating force of images is most apparent in how we think about and relate to other people. We construct images of other's minds and intentions in order to predict their actions and preferences. We construct images of what they spend their attention on, in order to manipulate them to spend it on us. Images of another's attention are ultimately constrained by our beliefs about how we can better secure *their* vitality for our own gain. When understanding love on the basis of images, relationships constitute securing another's life. Sometimes one may briefly break through the possessive frame (generally on first dates, honeymoons, or while "making love"), and enjoy the pleasure of experiencing another with complete attention and silence: without their personal images of one another obscuring the lens. Those brief moments are sublime and are outside space and time. The infinite distance separating one from another in memory is completely collapsed by the stillness of love, which is experienced only through shared attention. Love is ephemeral, as attention is ephemeral, but memory works to construct an image of love to secure the cause of its stillness (i.e., the "loved-one") in the future. (Categories serve the same function when discussing and thinking about love, as evident when people talk about their "type" when dating.) When experiencing love based on our image of another, we're drawn to secure them to sustain our personal security.

However, images of love makes us believe that security is outside of ourselves, and thus strips us of our

own agency. Our life is reoriented toward securing and desiring another, rather than simply experiencing shared attention, and appreciating another's unique expression. Desiring to secure love results in pangs of fear and anxiety because the desired individual can never be secured: what we hold onto is a dying image of the referent which only partially or even erroneously captures the experience we seek. Consequently, *mistaking an image for the other degrades the self at the expense of the other.* And the only way to radically transform ourselves and our societies, is to embrace the source of our own agency in our daily lives. This is only achievable through not experiencing one's life though the maintenance of their self-image. This only follows from not positioning one love over others, and not placing oneself over others. Capitalist imagery inspires possessive love because we love with the intent to position our self-image over all else, and allow for the commodification of the most personal dimensions of our being.

Online networks disseminate imagery that expresses categories and causal relationships that embolden the self at the expense of the other, because self-oriented thinking is necessary for economic growth and empires to flourish. Senseless economic activity is based on images of oneself securing themselves over others in the future: I am not this thing now, but through effective intervention (hard work, causal manipulation, securing another), I will become some other, more desirable thing. I will obtain a new categorization for myself in my own eyes and in the eyes of others. And my life will tell a story that others will remember when they think about me. This action right here, and those I plan to execute in the future, will help bring about (cause) this dream, and I will endure forever. These self-interested causal thinkers are driven to position themselves over others. When individuals position themselves over others, they position their images and groups over others, and their state over others. Hierarchy in the world constrains and emerges from hierarchy in thought. Hierarchy in thought emerges from the separating force of images and the binary inherited by representing the world with categories and causal relationships. What is and isn't defines what is good (the self - one's interior) and bad (the other - one's exterior).

Overcoming the separating force of divisive narratives

The full power of images manifests through the use of narratives and language, which structure and relate images within and across bodies. Narratives are higher-order structures emergent in a system of images, which contextualize and enrich individual images by placing them in relation to other images in the system. Narratives are interlinking networks of images, and structure the category and causal knowledge in our belief systems. The narratives we endorse constitute the way we understand the world and are highly sensitive to the narratives endorsed by our peers.[2] They structure people into groups as a function of the content and relationships between their shared images; people are seen as coherent with one another on the basis of if their endorsed narratives are coherent with one another.

Language emerges from two or more bodies acknowledging their shared attention and desiring to communicate their personal images and narratives with each other. Language is more general than spoken and written words, language consists in any form of communication. A single gesture can communicate truths words can never capture. Language is a shared convention that bridges and aligns interior spaces.

Language aligns bodies on the basis of the content and the relations between their images and narratives. Language allows for the progression of knowledge, revising of individual and collective beliefs, societal development, and tool building. The human proclivity for using language as a tool allows for effective collective problem solving skills that rapidly accelerates technological and scientific advancements. The rapid pace of our technological development allows societies to grow, and grow, and grow, and grow... As societies grow, they increase their need for efficiency and to optimize their constituents to maintain homeostasis. Opportunists can leverage the optimizing power of language to cohere minds and groups around centralizing and polarizing narratives, which facilitate collective action and conflict.

Category boundaries and the causal relationships governing those categories arise from the parsimonious nature of images (categories and causal relationships) and language (narratives). The human representational system must be sufficiently expressive of the world, but embodied in a biophysical system fueled by consuming approximately 2,000 calories a day. If we had a maximally expressive representational system (that is, a body with infinite memory and energy consumption), all things would be a thing of its own kind and there would be no need for reducing the world with causal relationships and models. In a maximally expressive system, there is no need for representational compositionality or abstraction, and thus causality because the complete covariation of the world would be known. The parsimony required of semantic representation *is* what we understand as causality, and it results from the energy and memory constraints of our bodies. In a maximally expressive system, an infinite set of predictors could predict future images of the world, and there would be no need to abstract away the forces governing how the sequence of images change over time. Causality is the salient force(s) governing change in a sequence of images over time to the observer (reference point). Our categories are supplied with this causal information so we have sufficiently useful expectations of how things can help us sustain our lives. Sustaining life is the mandate of categories and causality, and therefore all images.

Members of competing social groups believe competing narratives. This is seen concretely in the recent anti/pro science disagreements surrounding climate change and the Covid-19 pandemic; historically through opposing religious groups and explanatory narratives; disagreements about political action and cures for social ills. We live in a society where disagreements are no longer about issues, but rather identities. Narratives categorize the other to make us believe that our peers are the causes of our problem, not those at the helms of the image generating machines. We must see the narratives we consume as constructions

designed to pull us apart. Disowning the narratives fed to us is a crucial step towards transformative change.

The narratives and beliefs that spread far and wide in online social networks are optimized to do so, because they can efficiently assimilate into the maximum number of people's belief systems. Certain narrative structures allow for homogeneity in belief systems and are thus effective tools for spreading political and religious propaganda and curating devoted followers. No wonder being an influencer is the #1 career choice among Zoomers, and a cartoonish reality T.V. star exploited social media to become Leader of the Free World. Establishing causally simplistic identity-affirming narratives can make controlling the masses easier as the group's shared beliefs form a predictable monolith that is easy to manipulate and control (i.e., 'controlling the narrative'). Disseminating disorienting narratives is the goal of propaganda, misinformation campaigns, and identity politics, which is accelerated by machine learning algorithms optimizing interactions and media content in online social networks.

Powerful elites weaponize narratives and exploit their separating force divide and conquer people. Conflicting narratives foster competing groups and hijacks the attention of those consumed with the dissonance. Psychotic opportunists disseminate narratives designed to make their supporters believe that their groups must conquer the other side, either sexually, politically, economically, or militaristically, to sustain their own life and those that they love; more precisely, members of one's groups and categories: family-unit, neighborhood, political party, race, class, religion, nationality (and their categorical intersections within-in a specific causal context).

Conflict resulting from narrative disagreements is a key source of systemic and generational violence. Religious wars, political conflicts, military conquest, over-policing of minorities, immigrants, and working classes, the dehumanization of queer people, are justified by destroyers using spiritual, scientific, and political narratives. As long as we "pay attention" to *their* images, and define ourselves through *their* narratives, we will continue acting in *their* system: replacing our idols with new ideals in accelerating succession. Seeking images produced within their system only helps sustain their conquest as they bake the earth and seas and dislocate, homogenize, and atomize whatever stands in their way. We must awaken from their sinister phantasmagoria if we want self-determination and agency over our lives. Instead of succumbing to the separating force of images, those divided by conflicting narratives must recognize what is shared between any narrative - *the humanity of the individual* - and ultimately recognize themselves in each other's narratives. While this mutual intelligibility enables narratives to conflict it also enables narratives to cohere; any resulting mutual self-recognition is a first step toward mutual self-determination.

How often does an artist, scientist, or technocrat ask themselves why they're doing what they're doing? Do we need fruit juicers on the Internet of Things? Do we need thousands of research articles on social psychology? Do we need another artist positioning three paintings on gallery *walls* while unhoused people

Figure 6.1: Elon Musk discusses generative AI, purchasing Twitter, and overcoming the woke mind virus on Tucker Carlson Tonight. Tech and political narratives are becoming increasingly intertwined around influencer-like personalities.

die on the streets outside those walls? How often are we honest with ourselves about the purpose of our work and our contributions to society? Rarely ever, because what motivates most of this labor is the pursuit of an emboldened self and the optimization of one's social networks. Selfish individuals strive to maintain their influence at the top of their highly constructed kingdom rather than address the practical needs of their community. Artists, writers, scientists, and the like now seek microcelebrity status in their highly specialized echochambers; the prominent thinkers, entrepreneurs, and creators of our generation are more akin to influencers than stewards of a more equitable and just society (e.g., see Figure 6.1). The model of social media runs deep and is degrading all corners of life. Stepping outside of this disgraceful paradigm will require that people question the drivers of their work and thinking. It will require creation and production without the desire of fame and recognition. Transformation will require labor that addresses practical questions, rather than maintaining self-serving and highly refined personal endeavors.

The more we increase our technical prowess and refine our models and understanding of human behavior, we, as behavioral scientists, are developing the tools and frameworks that support the highly extractive technology that has imbalanced ourselves and our societies in the first place. Digital advertisements, predictive policing, and generative AI systems that compose convincing narratives are examples of models trained on our language and behaviors to understand our interior motives, manipulate us, and control us. Rather than continuing to develop the tools and paradigms that brought us here in the first place, should we simply step away and become more independent in our scientific practice? Citizen scientists seek to increase pubic engagement with one's scientific mission. How might behavioral and cognitive scientists orient their work to develop paradigms and technologies that help uphold basic principles? These are questions that I am still trying to understand.

Chapter 7

Appendices and supplementary materials

Additional mathematical details

I provide a more detailed picture of the computational mechanisms of the framework, specifically how consequences of beliefs are represented and integrated into a second-order belief, and how Bayesian reasoning updates second-order beliefs. In some places we repeat text present in the main manuscript to help contextualize details about the model.

Computing the expected utility and consequences of beliefs

The framework we present describes how utility information can alter doxastic states to produce belief reports. We describe this process in more detail.

We write the expected utility of reporting h_i as:

$$EU(h_i) = \mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{1}$$

where $\mathbb{P}(h_i)$ is the probability that a given hypothesis h_i is true, and $\mathbb{U}(h_i)$ is the utility the reasoner attributes to h_i being true. As noted in main text, adopting a new belief can be sensitive to consequences (Williams, 2021), such as maintaining or severing ties to one's social group (D. Kahan, 2013).

How may we operationalize this? A reasoner may encounter a set of K outcomes (or prospects) when believing h_i, which can be expressed as $c_{h_i} \subseteq \mathcal{C}(\mathcal{H})$. We use the notation c to denote outcomes because they

can understood as real-world *consequences* the reasoner attributes to a hypothesis. These symbols say the outcome of the hypothesis h_i is within the set of all possible outcomes. Individual outcomes associated with h_i are further indexed by $c^k_{h_i} \in \mathcal{C}_{h_i}$. The notation for $c^k_{h_i}$ can be understood as follows: the subscript, h_i denotes which hypothesis a given outcome belongs to (here, h_i) and the superscript, k denotes the indexing within the set of outcomes belonging to the hypothesis. Therefore, $c^k_{h_i}$ means the k^{th} outcome given hypothesis h_i. The reasoner needs to index a hypothesis' outcomes in order to compute $\mathbb{U}(h_i)$.

Outcomes occur with some degree of uncertainty and this uncertainty can vary as a function of the credence in a hypothesis. For instance, a pro-vaccination parent and an anti-vaccination parent may attribute similar costs to potentially negative consequences of a vaccine (e.g., their child getting sick from a vaccine dose), but anti-vaccination parents may assume the negative consequence are more likely than the pro-vaccination parent (Horne et al., 2015). We need to encode this uncertainty with a probability measure. The expression below means the joint probability of h_i being true and the probability of possible consequences of h_i manifesting as a result of h_i. We can rewrite $\mathbb{P}(h_i)$ as the probability of the conjunction of h_i and its associated outcomes as c_{h_i} as the joint probability $\mathbb{P}(h_i, c_{h_i})$. We will expand this joint probability in the equation below using a conditional probability.

For generality, we can say that h_i produces a set of k outcomes with computable utilities. By extending Equation 1 above, we can express the utility of h_i as a scaled proportion of its credence and credence of its consequences. More concretely, given a set of k outcomes to h_i, $c^k_{h_i} \in c_{h_i}$, we can then write the expected utility of a hypothesis as a function of the utilities associated with consequences conditioned on the hypothesis, which are scaled by the probability that the hypothesis is true and that the consequences obtains:

$$EU(h_i) = \mathbb{P}(h_i, c_{h_i})\, \mathbb{U}(c_{h_i}) = \mathbb{P}(c_{h_i}|h_i)\mathbb{P}(h_i)\, \mathbb{U}(c_{h_i}) \tag{2}$$

where we iterate over the possible outcomes and their associated utilities:

$$\sum_{k=1}^{K} \mathbb{P}(c^k_{h_i}|h_i)\mathbb{P}(h_i)\, \mathbb{U}(c^k_{h_i}) = \mathbb{P}(h_i) \sum_{k=1}^{K} \mathbb{P}(c^k_{h_i}|h_i)\, \mathbb{U}(c^k_{h_i}) \tag{3}$$

We need to select between hypothesis based on their relative utility values. That is, values must be normed to choose between them. Let $K_i = |C_{h_i}|$ where $|C_{h_i}|$ denotes the total number of outcomes (or the size of the set of consequences) for h_i, then:

$$z(h_i) = \frac{EU(h_i)}{\sum_{h_j \in \mathcal{H}} EU(h_j)} = \frac{\mathbb{P}(h_i) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i}|h_i)\, \mathbb{U}(c^k_{h_i})}{\sum_{h_j \in \mathcal{H}} \mathbb{P}(h_j) \sum_{k' \in K_j} \mathbb{P}(c^{k'}_{h_j}|h_j)\, \mathbb{U}(c^{k'}_{h_j})} \tag{4}$$

Here, $z(h_i)$ is the relative expected utility function which will compute the utility of a hypothesis *relative* to the other hypotheses under considerations (i.e., $h_j \in \mathcal{H}$). This is why in the second part of the expression (i.e., $\frac{\mathbb{U}(h_i)}{\sum_{h_j \in \mathcal{H}} \mathbb{U}(h_j)}$) we compute the utility of h_i in the numerator (i.e., $\mathbb{U}(h_i)$) and divide it by the summation of utilities for the remainder of the hypotheses in the denominator. Quantifying a hypothesis' utility is somewhat hand-wavy as a hypothesis in the abstract doesn't have a utility per se, rather it is represented as being evidence for possible outcomes that do have a real-world, computable utilities. So in the third part of the equation, we expand out the equation to make this fact explicit. This is an application of Equation 3, where the expected utility of h_i is equal to the summation (from 1 to K outcomes) of the joint probability of a hypothesis and its outcome, multiplied by the utility of those consequences. This is just an application of Bayes' rule: the joint probability $\mathbb{P}(h, c_{h_i})$ is equal to the prior probability of h_i ($\mathbb{P}(h_i)$) multiplied by the probability of the outcome c_{h_i} conditional on the credence in hypothesis h_i. We'll assume for simplicity that people often attempt to hold hypotheses that are maximally rewarding. Stated another way, people generally aim to maximize the expected utility they perceive as being a consequence of holding a hypothesis. We are not specific when describing how to compute a utility function, because there are various ways to do so (Maher, 1993). For instance, a reasoner may consider a trade-off between a concern for accuracy (which leads to accepting hypothesis with high probabilities) versus informativeness (lower probability). A utility function capturing preference for one desideratum over the other is a *cognitive utility function*, as it ascribes utility-values to the outcomes of a hypothesis being true.

Incorporating Bayesian inference when computing second-order beliefs

By integrating Bayesian updating in our expected utility framework, we can understand belief reports as Bayesian decisions, with the the evidence, the reasoner's prior, and the utility of holding the belief impact belief reports:

$$EU(h_i|d) = \mathbb{P}(h_i|d)\,\mathbb{U}(h_i) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{5}$$

Equation 5 is an extension of Equation 2 in which $\mathbb{P}(h_i)$ is replaced with the posterior of h_i given d. Therefore, the expected utility of a hypothesis is updated via Bayes' rule to incorporate learning new data.

We can then expand Equation 5 to its fully expressed form:

$$EU(h_i|d) = \mathbb{P}(h_i|d) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i}|h_i)\,\mathbb{U}(c^k_{h_i}) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i}|h_i)\,\mathbb{U}(c^k_{h_i}) \tag{6}$$

This equation states that we will update the utility in line with Bayesian updating. As discussed in the

main text, the second-order belief with the maximum expected utility is the belief report.

Alternative motivated reasoning models

In the motivated reasoning model discussed in the main text, a reasoner's prior is conditioned on the utilities of the possible outcomes. Utilities shape how evidence is integrated into posteriors in light of a utility-biased prior (i.e., additional credence is awarded to higher utility hypotheses which influences the likelihood calculation in Bayes theorem). However, there are other approaches a researcher might take to model how goals bias evidence integration. This could take the form of introducing additional uncertainty when computing the amount of evidence for a desired hypothesis. For example, in many motivated reasoning experiments, participants are asked to reason about unclear policy proposals and statements of potential evidence for or against that policy. Each statement about a policy proposal represents a noisy quantity of evidence. A bounding box algorithm could account for this implicit uncertainty when interpreting evidence (D. Powell, 2021; Zhu et al., 2020).

A *bounding box* defines a range of values a remembered piece of evidence may take given a ground-truth amount of evidence ev_{world}. Specifically, the amount of evidence in favor of (ev_{for}) or against ($ev_{against}$) a hypothesis are sampled from ranges of values centered by the outcome of ev_{world}. Sampled evidence values $\hat{ev}_{for}$ and $\hat{ev}_{against}$ are then integrated into a posterior belief via Bayesian updating. The main motivated reasoning model we develop describes how utilities influence evidence integration by directing priors to influence how much evidence is assimilated (i.e., evidence for higher-utility outcomes receive higher weight through the directional prior). However, a motivated reasoning model model based on a bounding box implementation describes how utilities bias memory representations of evidence. This model would use normative priors, but the *amount* of evidence assimilated into posterior beliefs would be biased by the utilities.

We will introduce this concept more precisely using the discrete case, however, a continuous case can be developed as well. Consider an example where three out of five facts support a hypothesis. A reasoner will likely represent the number of facts supporting a hypothesis, ev_{for}, and the amount of facts against a hypothesis, $ev_{against}$, as *distinct* representations – as two separate random variables. With length $l = 1$, we define bounding boxes around $ev_{world} = 3$ for evidence in support of, $ev_{for} \in [2, 3, 4]$), and against the hypothesis, $ev_{against} \in [1, 2, 3]$). Note that $ev_{against}$ is centered at 2 because if 3 out of 5 facts favor the hypothesis, 2 out of 5 are against it. $\hat{ev}_{for}$ and $\hat{ev}_{against}$ are then sampled from these two ranges, respectively, and posterior beliefs result from Bayesian updating over the sampled evidence representations.

Researchers can superimpose a probability distribution over the box to introduce systematic noise in how

evidence counts are sampled from the respective ranges.[1] Let $\mathcal{F}$ indicate the sampling distribution over the bounding box with distributional parameters $\vec{\sigma}$. Given a maximum amount of evidence for a hypothesis k, the full distributional model for constructing a posterior credence from evidence sampled from a bounding box of length l is:

$$ev_{world} \sim unif[1, k]$$

$$ev_{against} = k - ev_{world}$$

$$\hat{ev}_{for} \sim \mathcal{F}([ev_{world} - l_{lower} : ev_{world} + l_{upper}]; \vec{\sigma})$$

$$\hat{ev}_{against} \sim \mathcal{F}([ev_{against} - l_{upper} : ev_{against} + l_{lower}]; \vec{\sigma})$$

$$credence \sim \beta(a + \hat{ev}_{for}, b + (k - \hat{ev}_{against}))$$

How might an experimenter model the impact of utilities on evidence representations? In a discrete case where we have pieces of evidence of equal weight, shifting the limits of the box toward the motivated hypothesis by a single unit leads a "motivated reasoner" to remember one more piece of evidence for the desired hypothesis with probability p. This p value for box index i is equal to the density on the bounding box specified by $\mathcal{F}(I = i; \vec{\sigma})$. In an experiment where we manipulate the amount of evidence for a motivated hypothesis, participants could be asked to remember the amount of evidence after each trial. We can then compute the difference between how much evidence participants remember and the true evidence value. More precisely, by estimating $\mathcal{F}(I = i, \vec{\sigma})$. To better discern how motivations are biasing people's thinking for a specific case, psychologists can compare model fits from this model against the directional priors model implemented in the main text.

[1] This noise can also be instantiated with variable-length bounding boxes, where a reasoner has an increased bounding box for evidence states that produce high credence in desired hypothesis. This hyperparamter can be defined such that it accounts for individual differences in how motivational goals bias evidence sampling.

Bibliography

Abend, G., Petre, C., & Sauder, M. (2013). Styles of causal thought: An empirical investigation. *American Journal of Sociology*, *119*(3), 602–654.

Alker, H., & Poppen, P. (1973). Personality and ideology in university students. *Journal of Personality*, *41*(4), 653–671.

Altay, S., de Araujo, E., & Mercier, H. (2022). "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, *10*(3), 373–394.

Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., & Leonardi, S. (2012). Online team formation in social networks. In *Proceedings of the 21st international conference on world wide web* (pp. 839–848).

Arceneaux, K., & Vander Wielen, R. J. (2013). The effects of need for cognition and need for affect on partisan evaluations. *Political Psychology*, *34*(1), 23–42.

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of investigative psychology and offender profiling*, *2*(1), 43–63.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526.

Avolio, M. L., Carroll, I. T., Collins, S. L., Houseman, G. R., Hallett, L. M., Isbell, F., ... Wilcox, K. R. (2019). A comprehensive approach to analyzing community dynamics using rank abundance curves. *Ecosphere*, *10*(10), e02881.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, *11*(1), 109–126.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299.

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

Bai, X., Griffiths, T., & Fiske, S. (2022). Explore-exploit tradeoffs generate cascading societal stereotypes.

Barkoczi, D., & Galesic, M. (2016, October). Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, *7*(1), 13109. Retrieved 2023-12-01, from `https://www.nature.com/articles/ncomms13109` (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/ncomms13109

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Blanco, E., Castell, N., & Moldovan, D. (2008). Causal relation extraction. In *Proceedings of the sixth international conference on language resources and evaluation (lrec'08)*.

Bodner, J., Welch, W., & Brodie, I. (2020). *Covid-19 conspiracy theories: Qanon, 5g, the new world order and other viral ideas*. McFarland.

Booten, K. (2016). Hashtag drift: Tracing the evolving uses of political hashtags over time. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2401–2405). ACM.

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3), 443–445.

Bower, G. H., & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, *247*(4938), 44–48.

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences* (pp. 1–10).

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020, July). The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspectives on Psychological Science*, *15*(4), 978–1010. Retrieved 2024-03-20, from `https://doi.org/10.1177/1745691620917336` (Publisher: SAGE Publications Inc) doi: 10.1177/1745691620917336

Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M., & Crockett, M. J. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Na-*

ture human behaviour, *7*(6), 917–927. Retrieved 2024-03-20, from https://www.nature.com/articles/s41562-023-01582-0 (Publisher: Nature Publishing Group UK London)

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017, July). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. Retrieved 2024-03-20, from https://pnas.org/doi/full/10.1073/pnas.1618923114 doi: 10.1073/pnas.1618923114

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Brenan, M., & Saad, L. (2018). *Global warming concern steady despite some partisan shifts.* Retrieved from https://news.gallup.com/poll/231530/global-warming-concern-steady-despite-partisan-shifts.aspx.

Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Brown, G. D., Lewandowsky, S., & Huang, Z. (2022). Social sampling and expressed attitudes: Authenticity preference and social extremeness aversion lead to social norm effects and polarization. *Psychological Review*, *129*(1), 18.

Burton, J. W., Cruz, N., & Hahn, U. (2021, December). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635. Retrieved 2024-03-20, from https://www.nature.com/articles/s41562-021-01133-5 (Publisher: Nature Publishing Group) doi: 10.1038/s41562-021-01133-5

Bürkner, P.-C. (2017, August). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. Retrieved 2024-01-31, from https://doi.org/10.18637/jss.v080.i01 doi: 10.18637/jss.v080.i01

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116-131.

Caddick, Z. A. (2016). Evaluating contradicting and confirming evidence: A study on beliefs and motivated reasoning. *ProQuest Dissertations and Theses*, 85. Retrieved from `http://ezproxy.msu.edu/login?url=https://www.proquest.com/dissertations-theses/evaluating-contradicting-confirming-evidence/docview/1867574098/se-2`

Caddick, Z. A., & Feist, G. J. (2021). When beliefs and evidence collide: psychological and ideological predictors of motivated reasoning about climate change. *Thinking & Reasoning*, 1–37.

Caddick, Z. A., & Rottman, B. M. (2021). Motivated reasoning in an explore-exploit task. *Cognitive Science*, *45*(8), e13018.

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of experimental psychology: Applied*, *7*(2), 91.

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*, 807–840.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994.

Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., & Hwang, A. (2020). Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*.

Chen, D. L., Schonger, M., & Wickens, C. (2016, March). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Chen, X. (2016). The influences of personality and motivation on the sharing of misinformation on social media. *IConference 2016 Proceedings*.

Chen, X., Sin, S.-C. J., Theng, Y.-L., & Lee, C. S. (2015). Why do social media users share misinformation? In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries* (pp. 111–114).

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). Elsevier.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*(2), 139–153.

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., ... Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest*, *23*(2), 50-97.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179.

Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011, June). Analyzing the Dynamic Evolution of Hashtags on Twitter: a Language-Based Approach. In M. Nagarajan & M. Gamon (Eds.), *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 58–65). Portland, Oregon: Association for Computational Linguistics.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387.

Dawson, P. (2020). Hashtag narrative: Emergent storytelling and affective publics in the digital age. , *23*(6), 968–983.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv. (arXiv:1810.04805 [cs])

Dijksterhuis, A., Van Knippenberg, A., Kruglanski, A. W., & Schaper, C. (1996). Motivated social cognition: Need for closure effects on memory and judgment. *Journal of Experimental Social Psychology*, *32*(3), 254–270.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584.

Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin*, *29*(19), 1120–1132.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, *75*(1), 53–69.

Dixon, G., Bullock, O., & Adams, D. (2019). Unintended effects of emphasizing the role of climate change in recent natural disasters. *Environmental Communication*, *13*(2), 135–143.

Douglas, K. M. (2021). Covid-19 conspiracy theories. *Group Processes & Intergroup Relations*, *24*(2), 270–275.

Druckman, J. (2015). Communicating policy-relevant science. *PS: Political Science & Politics*, *48*(S1), 58–69.

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

Dunietz, J., Levin, L., & Carbonell, J. (2017). Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, *5*, 117–133.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417.

Emler, N., Renwick, S., & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology*, *45*(5), 1073–1080.

Enders, A. M., Uscinski, J. E., Klofstad, C., & Stoler, J. (2020). The different forms of covid-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review*.

Enisman, M., Shpitzer, H., & Kleiman, T. (2021). Choice changes preferences, not merely reflects them: A meta-analysis of the artifact-free free-choice paradigm. *Journal of Personality and Social Psychology*, *120*(1), 16.

Falk, A., & Zimmermann, F. (2016). *Beliefs and utility: Experimental evidence on preferences for information*. CESifo Working Paper Series.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2).

Fazio, L., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*(5), 1705–1710.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Findlay, S. D., & Thagard, P. (2014). Emotional change in international negotiation: Analyzing the camp david accords using cognitive-affective maps. *Group Decision and Negotiation*, *23*, 1281–1300.

Fiore, M. C., Novotny, T. E., Pierce, J. P., Giovino, G. A., Hatziandreu, E. J., Newcomb, P. A., ... Davis, R. M. (1990). Methods Used to Quit Smoking in the United States: Do Cessation Programs Help? *JAMA*, *263*, 2760–2765.

Fishkin, J., Keniston, K., & McKinnon, C. (1973). Moral reasoning and political ideology. *Journal of Personality and Social Psychology*, *27*(1), 109–119.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), 110–161.

Gallup. (2014). *Evolution, creationism, intelligent design.* Retrieved from `https://news.gallup.com/poll/21814/evolution-creationism-intelligent-design.aspx`

Gelman, A., Lee, D., & Guo, J. (2015, October). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. Retrieved 2024-01-31, from `https://doi.org/10.3102/1076998615606113` (Publisher: American Educational Research Association) doi: 10.3102/1076998615606113

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual review of anthropology*, *40*, 379.

Gershman, S. (2018). How to never be wrong. *Psychonomic Bulletin and Review*, *26*(1), 1–16.

Giaxoglou, K. (2018). #jesuischarlie? hashtags as narrative resources in contexts of ecstatic sharing. *Discourse, context & media*, *22*, 13–20.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, *44*(6), 1110–1126.

Goldstone, R. L., & Lupyan, G. (2016). *Discovering psychological principles by mining naturally occurring data sets* (Vol. 8) (No. 3). Wiley Online Library.

Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, *5*(3), 143–164.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149.

Goudas, T., Louizos, C., Petasis, G., & Karkaletsis, V. (2014). Argument extraction from news, blogs, and social media. In *Artificial intelligence: Methods and applications: 8th hellenic conference on ai, setn 2014, ioannina, greece, may 15-17, 2014. proceedings 8* (pp. 287–299).

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J., Friese, M., ... others (2020). The implicit association test at age 20: What is known and what is not known about implicit bias.

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

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Hahn, U. (2022). Collectives and epistemic rationality. *Topics in Cognitive Science*.

Hallett, L. M., Jones, S. K., MacDonald, A. A. M., Jones, M. B., Flynn, D. F. B., Ripplinger, J., ... Collins, S. L. (2016). codyn: An r package of community dynamics metrics. *Methods in Ecology and Evolution*, *7*(10), 1146–1151.

Harman, G. (1986). *Change in view: Principles of reasoning.* The MIT Press.

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Harrison, L., & Startin, N. (2013). *Political research: An introduction.* Routledge.

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701–723.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; A case study. *Journal of Abnormal and Social Psychology*, *49*(1), 129–134.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., ... Szpakowicz, S. (2010, July). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 33–38). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/S10-1006`

Hershey, J. C., & Schoemaker, P. J. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, *31*(10), 1213–1231.

Hickling, A., Wellman, H., & Dannemiller, J. L. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668–683.

Hidey, C., & McKeown, K. (2018). Persuasive influence detection: The role of argument sequencing. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135–163.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179–1203.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*(1), 3.

Homer-Dixon, T., Maynard, J. L., Mildenberger, M., Milkoreit, M., Mock, S. J., Quilley, S., ... Thagard, P. (2013). A complex systems approach to the study of ideology: Cognitive-affective structures and the dynamics of belief systems. *Journal of social and political psychology*, *1*(1).

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324.

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. *Nature Climate Change*, *8*(7), 614–620.

Howard, P. N., & Hussain, M. M. (2013). *Democracy's Fourth Wave?: Digital Media and the Arab Spring*. Oxford University Press.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410–1412.

Imundo, M. N., & Rapp, D. N. (2022). When fairness is flawed: Effects of false balance reporting and weight-of-evidence statements on beliefs and perceptions of climate change. *Journal of Applied Research in Memory and Cognition*, *11*(2), 258.

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences*, *107*(51), 22014–22019.

Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E., & Galinsky, A. D. (2018). The critical role of second-order normative beliefs in predicting energy conservation. *Nature Human Behaviour*, *2*(10), 757–764.

Jackson, E. G. (2019). Belief and credence: Why the attitude-type matters. *Philosophical Studies*, *176*(9), 2477–2496.

Jain, A., Marshall, J., Buikema, A., Bancroft, T., Kelly, J. P., & Newschaffer, C. J. (2015). Autism occurrence by mmr vaccine status among us children with older siblings with and without autism. *Jama*, *313*(15), 1534–1540.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763.

Janvrin, D. J., Raschke, R. L., & Dilla, W. N. (2014). Making sense of complex data using interactive data visualization. *Journal of Accounting Education*, *32*(4), 31–48.

Jern, A., Chang, K. M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Jin, Y., Jensen, G., Gottlieb, J., & Ferrera, V. (2022). Superstitious learning of abstract order from random reinforcement. *Proceedings of the National Academy of Sciences*, *119*(35), e2202789119. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.2202789119` doi: 10.1073/pnas.2202789119

Jo, Y., Poddar, S., Jeon, B., Shen, Q., Rose, C., & Neubig, G. (2018). Attentive Interaction Model: Modeling Changes in View in Argumentation. In M. Walker (Ed.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103 – 116). Stroudsburg, PA: Association for Computational Linguistics.

Jolley, D., Douglas, K. M., & Sutton, R. M. (2018). Blaming a few bad apples to save a threatened barrel: The system-justifying function of conspiracy theories. *Political Psychology*, *39*(2), 465–478.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, 1–17.

Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, *8*, 407–424.

Kahan, D., & Braman, D. (2006). Cultural cognition and public policy. *Yale Law and Policy Review*, *24*, 149-172.

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

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735. doi: 10.1038/nclimate1547

Kalis, A., Mojzisch, A., Schweizer, T. S., & Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 402–417.

Kashima, Y., Perfors, A., Ferdinand, V., & Pattenden, E. (2021). Ideology, communication and polarization. *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200133.

Khoo, C., Chan, S., & Niu, Y. (2002). The many facets of the cause-effect relation. In *The semantics of relationships* (pp. 51–70). Springer.

Killen, M., & Stangor, C. (2001). Children's social reasoning about inclusion and exclusion in gender and race peer group contexts. *Child Development*, *72*(1), 174–186.

Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two–process approach to adolescent cognition. *Child Development*, *71*(5), 1347–1366.

Klayman, J. (1995, January). Varieties of Confirmation Bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of Learning and Motivation* (Vol. 32, pp. 385–418). Academic Press. doi: 10.1016/S0079-7421(08)60315-1

Kraus, M. W., & Tan, J. J. (2015). Americans overestimate social class mobility. *Journal of Experimental Social Psychology*, *58*, 101–111.

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: need for closure and the unfolding of group-centrism. *Psychological review*, *113*(1), 84-100.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kundra, Z., & Sinclair, L. (1999). Motivated reasoning with stereotypes: Activation, application, and inhibition. *Psychological Inquiry*, *10*(1), 12–22.

Laato, S., Islam, A., Islam, M. N., & Whelan, E. (2020). Why do people share misinformation during the covid-19 pandemic? *arXiv preprint arXiv:2004.09600*.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., ... others (2014). Reducing implicit racial preferences: I. a comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785.

Lee, D., & Coricelli, G. (2020). An empirical test of the role of value certainty in decision making. *Frontiers in psychology*, *11*, 574473.

Lee, D., & Daunizeau, J. (2020). Choosing what we like vs liking what we choose: How choice-induced preference change might actually be instrumental to decision-making. *PloS one*, *15*(5), e0231081.

Lee, D. G., & Holyoak, K. J. (2021). Coherence shifts in attribute evaluations. *Decision*, *8*(4), 257.

Lee, D. G., & Holyoak, K. J. (2023). Transient value refinements during deliberation facilitate choice. *Decision*.

Levy, A. G., Thorpe, A., Scherer, L. D., Scherer, A. M., Drews, F. A., Butler, J. M., ... Fagerlin, A. (2022). Misrepresentation and nonadherence regarding covid-19 public health measures. *JAMA Network Open*, *5*(10), e2235837–e2235837.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., & Slonim, N. (2014, August). Context dependent claim detection. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1489–1500). Dublin, Ireland: Dublin City University and Association for Computational Linguistics. Retrieved from `https://aclanthology.org/C14-1141`

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, *47*(8), 1445–1456.

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., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. , *6*(4), 353–369. Retrieved 2023-12-08, from `https://www.sciencedirect.com/science/article/pii/S2211368117300700` doi: 10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ginac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*, 399 – 404.

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

Lewandowsky, S., & van der Linden, S. (2022). Interventions based on social norms could benefit from considering adversarial information environments: Comment on constantino et al. (2022). *Psychological Science in the Public Interest*, *23*(2), 43-49.

Li, M., Reddy, R. G., Wang, Z., Chiang, Y.-S., Lai, T., Yu, P., ... Ji, H. (2022). Covid-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th annual meeting of the association for computational linguistics: System demonstrations* (pp. 135–144).

Li, Z., Li, Q., Zou, X., & Ren, J. (2021). Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, *423*, 207–219.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lin, Y.-R., Margolin, D., Keegan, B., Baronchelli, A., & Lazer, D. (2013). # bigbirds never die: Understanding social dynamics of emergent hashtags. In *Proceedings of the international aaai conference on web and social media* (Vol. 7, pp. 370–379).

Little, A. T. (2021). Directional motives and different priors are observationally equivalent. *University of California-Berkeley (Unpublished Manuscript)*, 1–31.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *The Oxford Handbook of Causal Reasoning*, 415.

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), 2098–2109.

Luce, R. D. (1991). Rank-and sign-dependent linear utility models for binary gambles. *Journal of Economic Theory*, *53*(1), 75–100.

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, *66*(3), 516–553.

Mackie, J. L. (1980). *The cement of the universe: A study of causation*. Clarendon Press.

Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanku, E., Tully, M., ... Uzuegbunam, C. (2020). Motivations for sharing misinformation: a comparative study in six sub-saharan african countries. *International Journal of Communication*.

Maher, P. (1993). *Betting on theories*. Cambridge University Press.

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e2200300119.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

McGuire, W. J. (1964). Inducing resistance to persuasion. some contemporary approaches. *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing), 1981*, 192-230.

Mednick, S. A., & Freedman, J. L. (1960). Stimulus generalization. *Psychological Bulletin*, *57*(3), 169.

Midgon, B. (2022). Experts worry monkeypox disinformation will harm LGBTQ+ community. *The Hill*.

Morrow, D. G., Bower, G. H., & Greenspan, S. L. (1989). Updating situation models during narrative comprehension. *Journal of memory and language*, *28*(3), 292–312.

Moskowitz, G. B. (1993). Individual differences in social categorization: The influence of personal need for structure on spontaneous trait inferences. *Journal of Personality and Social Psychology*, *65*(1), 132-142.

Neuberg, S. L., Judice, T. N., & West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of personality and social psychology*, *72*(6), 1396.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of personality and social psychology*, *65*(1), 113-131.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220.

Nir, L. (2011). Motivated reasoning and public opinion perception. *Public Opinion Quarterly*, *75*(3), 504–532.

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

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? an experimental evaluation of the effects of corrective information. *Vaccine*, *33*, 459 – 464.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014a). Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, *133*(4), e835–e842.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014b). Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, *133*, 835 – 842.

Orne, M. T. (2017). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In *Sociological methods* (pp. 279–299). Routledge.

Palau, R. M., & Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law* (pp. 98–107).

Papacharissi, Z. (2015). *Affective publics: sentiment, technology, and politics.* New York, NY: Oxford University Press.

Papacharissi, Z. (2016). Affective publics and structures of storytelling: sentiment, events and mediality. *Information, Communication & Society*, *19*(3), 307–324.

Pasek, J. (2018). It's not my consensus: Motivated reasoning and the sources of scientific illiteracy. *Public Understanding of Science*, *27*(7), 787–806.

Pennycook, G. (2022). A framework for understanding reasoning errors: From fake news to climate change and beyond. *PsyArXiv*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). 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., & Rand, D. G. (2020). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402.

Pettigrew, R. (2019). Epistemic Utility Arguments for Probabilism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2019/entries/epistemic-utility/`.

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior*. The Guilford Press.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24). Springer.

Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, *8*(2), 617–641.

Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & van der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, *9*(8), 211953.

Poddar, S., Mondal, M., Misra, J., Ganguly, N., & Ghosh, S. (2022). Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In *Proceedings of the international aaai conference on web and social media* (Vol. 16, pp. 782–793).

Powell, D. (2021). Comparing probabilistic accounts of probability judgments.

Powell, D., Weisman, K., & Markman, E. (2022). Modeling and leveraging intuitive theories to improve vaccine attitudes.

Powell, M., Kim, A. D., & Smaldino, P. E. (2023, June). Hashtags as signals of political identity: #BlackLivesMatter and #AllLivesMatter. *PLOS ONE*, *18*(6), e0286524. Retrieved 2024-03-20, from `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0286524` (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0286524

Priniski, J., Verma, I., & Morstatter, F. (2023). Pipeline for modeling causal beliefs from natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 436–443). Toronto, Canada: Association for Computational Linguistics.

Priniski, J. H., & Holyoak, K. J. (2020). Crowdsourcing to analyze belief systems underlying social issues. In *Proceedings of the annual conference of the cognitive science society.*

Priniski, J. H., & Holyoak, K. J. (2022). A darkening spring: How preexisting distrust shaped covid-19 skepticism. *PloS one*, *17*(1), e0263191.

Priniski, J. H., & Horne, Z. (2018). Attitude change on reddit's change my view. In *40th annual meeting of the cognitive science society* (pp. 2279–2284).

Priniski, J. H., & Horne, Z. (2019). Crowdsourcing effective educational interventions. In *41st Annual Meeting of the Cognitive Science Society* (pp. 2599–2605).

Priniski, J. H., McClay, M., & Holyoak, K. J. (2021). Rise of qanon: A mental model of good and evil stews in an echochamber. *ArXiv Preprint*.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (Vol. 2). Random House: New York.

Radcliffe, E. S. (1999). Hume on the generation of motives: Why beliefs alone never motivate. *Hume Studies*, *25*(1), 101–122.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56.

Roozenbeek, J., Freeman, A. L., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al.(2020). *Psychological Science*, *32*(7), 1169–1178.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), eabo6254. Retrieved from `https://www.science.org/doi/abs/10.1126/sciadv.abo6254` doi: 10.1126/sciadv.abo6254

Saltman, K. J. (2020). Salvational super-agents and conspiratorial secret agents: Conspiracy, theory, and fantasies of control. *symploke*, *28*(1), 51–63.

Sardianos, C., Katakis, I. M., Petasis, G., & Karkaletsis, V. (2015). Argument extraction from news. In *Proceedings of the 2nd workshop on argumentation mining* (pp. 56–66).

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2021a). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, *4*(1), 381–402.

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2021b). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, *4*(1), 381–402.

Schoenfield, M. (2018). Permissivism and the value of rationality: A challenge to the uniqueness thesis. *Philosophy and Phenomenological Research*, *99*(2), 286–297.

Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, *62*(3), 291–324.

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and moral development. *The Emergence of Morality in Young Children*, 199–283.

Shweder, R. A., & Sullivan, M. A. (1993). Cultural psychology: Who needs it? *Annual Review of Psychology*, *44*(1), 497–523.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* New York: Cambridge University Press.

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

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1250–1260.

Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394.

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49*(2), 123–138.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Soicher, R. N., & Becker-Blease, K. A. (2020). Utility value interventions: Why and how instructors should use them in college psychology courses. *Scholarship of Teaching and Learning in Psychology*.

Sparkman, G., Geiger, N., & Weber, E. U. (2022). Americans experience a false social reality by underestimating popular climate policy support by nearly half. *Nature Communications*, *13*, 4779.

Stanley, J. (2015). How propaganda works. In *How propaganda works.* Princeton University Press.

Stetzka, R. M., & Winter, S. (2021). How rational is gambling? *Journal of Economic Surveys*.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159 – 171.

Taber, C., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155.

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

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 2). MIT press.

Talmy, L. (2011). Cognitive semantics: An overview. *Semantics: an international handbook of natural language meaning. Berlim*.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In J. Bourdeau, J. A. Hendler, &

R. N. Nkambou (Eds.), *Proceedings of the 25th International Conference on World Wide Web* (pp. 613 – 624). New York, NY: Association for Computing Machinery.

Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. (2020, June). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLOS ONE*, *15*(6), e0233879.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* Penguin.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

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

Turetsky, K. M., & Sanderson, C. A. (2017). Comparing educational interventions: Correcting misperceived norms improves college students' mental health attitudes. *Journal of Applied Social Psychology*, *48*, 46 – 55.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Uscinski, J. E., Douglas, K., & Lewandowsky, S. (2017). Climate change conspiracy theories. In *Oxford research encyclopedia of climate science.*

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49–58.

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 Harreveld, F., Rutjens, B. T., Schneider, I. K., Nohlen, H. U., & Keskinis, K. (2014). In doubt and disorderly: Ambivalence promotes compensatory perceptions of order. *Journal of Experimental Psychology: General*, *143*(4), 1666.

Vineberg, S. (2022). Dutch Book Arguments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2022/entries/dutch-book/`.

Vivion, M., Anassour Laouan Sidi, E., Betsch, C., Dionne, M., Dubé, E., Driedger, S. M., ... others (2022). Prebunking messaging to inoculate against covid-19 vaccine misinformation: An effective strategy for public health. *Journal of Communication in Healthcare*, 1–11.

Voigt, K., Murawski, C., Speer, S., & Bode, S. (2019). Hard decisions shape the neural coding of preferences. *Journal of Neuroscience*, *39*(4), 718–726.

Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton University Press.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive science*, *31*(2), 233–256.

Waldmann, M. R. (2017). Causal reasoning: An introduction. *The Oxford Handbook of Causal Reasoning*, 1–9.

Walker, M., & Matsa, K. E. (2021). News consumption across social media in 2021.

Wallace, R. J. (2020). Practical reason. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford University. `https://plato.stanford.edu/archives/spr2020/entries/practical-reason/`.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of personality and social psychology*, *67*(6), 1049-1062.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*(2), 357–378.

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press.

Wiggins, D. (1978). Weakness of will commensurability, and the objects of deliberation and desire. In *Proceedings of the aristotelian society* (Vol. 79, pp. 251–277).

Williams, D. (2021). Socially adaptive belief. *Mind & Language*, *36*(3), 333–354.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General*, *136*(1), 82.

Wolff, P., Klettke, B., Ventura, T., & Song, G. (2005). Expressing causation in english and other languages.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, *3*(6), 767–773.

Yang, G. (2016). Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter. *Media and Communication*, *4*(4), 13–17.

Yang, J., Han, S. C., & Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 1–26.

Yu, A. J., & Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? *Advances in neural information processing systems*, *21*.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282.

Yuille, A. L., & Grzywacz, N. M. (1988). A computational theory for the perception of coherent visual motion. *Nature*, *333*(6168), 71–74.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748.

Zimper, A., & Ludwig, A. (2009). On attitude polarization under Bayesian learning with non-additive beliefs. *Journal of Risk and Uncertainty*, *39*(2), 181–212.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*(2), 245–287.

Zwaan, R. A. (2022). Conspiracy thinking as situation model construction. , *47*, 101413.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, *6*(5), 292–297.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, *21*(2), 386.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.