Political Argumentation and Attitude Change in Online Interactions

Isaac D. Mehlhaff*

March 28, 2025

Abstract

Prevailing theories of public opinion and political psychology hold that human reasoning is biased and lazy, which suggests it is ill-suited to help ordinary citizens engage meaningfully with politics. In contrast, I contend that citizens' reasoning capabilities are primarily activated by high-quality argumentative exchanges, and that such exchanges motivate citizens to think through political issues more carefully and update their attitudes in response. To test this hypothesis, I train a deep neural network to simultaneously classify textual inputs on several characteristics of discussion and argumentation. I use these classifiers to annotate over one million comments from the Reddit social media platform and show that attitude change is substantially more likely to result from argumentative exchanges rather than more casual ones. Even in the often-hostile context of online political discussion, citizens can be quite skilled political reasoners.

*The University of Chicago; imehlhaff@tamu.edu; word count: 11,521. For helpful conversations and suggestions, I thank Lisa Argyle, John Brehm, Jaime Settle, Ethan Busby, Josh Gubler, Dominik Stecuła, Dave Redlawsk, Basak Taraktas, Marco Morucci, Marc Hetherington, Tim Ryan, Mary Kroeger, Rahsaan Maxwell, seminar participants at UChicago and Texas A&M, and participants at the 2022 APSA Annual Meeting.

Citizens' ability to make reasoned choices about political issues and candidates is central to the function of democracy (Dahl 1998; Madison 1787; Madison 1788). This requisite of proper democratic practice has spawned extensive work on the question of whether citizens know enough to meaningfully participate in politics. Predominant theories of public opinion generally avoid attributing high-level reasoning skills to citizens (Farrell, Mercier, and Schwartzberg 2023), instead emphasizing attitude sources ranging from media and political elites (Achen and Bartels 2016; Zaller 1992) to heuristics such as party identification (Berelson, Lazarsfeld, and McPhee 1954; Lupia and McCubbins 1998). These perspectives suggest that citizens are unskilled reasoners; they are unable to logically think through political issues, so they rely on heuristics, stereotypes, and biases to make minimally informed decisions, often through a process of motivated reasoning.

With respect to the mechanics of reasoning, the conventional wisdom is almost certainly correct: Citizens are biased and lazy political reasoners. I push back, however, on the notion that partiality and sloth make them *unskilled* reasoners. Drawing on theory from cognitive psychology, I argue that these characteristics can actually make citizens capable political reasoners, but only under certain conditions. Specifically, I contend that citizens' political reasoning skills are best utilized not when they attempt to reason on their own or even when they engage in casual political discussion. Instead, citizens are most likely to change their opinions and reduce their reliance on partisan stereotypes and heuristics when they engage in debate, exchanging and evaluating a series of arguments and counterarguments with an interlocutor.

I evaluate this theory in the context of online political discussion. Approximately 70 percent of Americans use at least one social media platform (Pew Research Center 2021), and as many as 94 percent of users attest to seeing at least some political content on their platform of choice (Pew Research Center 2016), making these types of online spaces one of the most common arenas in which citizens might encounter political talk. But despite—or perhaps because of—their near-ubiquity, scholars often portray online environments as a place where attitude change is not likely to happen. Hostility and harassment are commonplace (Matias 2019), anonymity and textual communication can undermine civil discussions among otherwise agreeable people (Brady, Crockett, and Bavel 2020), and increased visibility can amplify the voices of those who display psychological traits leading them to be hostile both online and offline (Bor and Petersen 2022). Bail et al. (2018) show that exposure to counterattitudinal opinions on Twitter can increase political polarization, and Baek, Wojcieszak, and Delli Carpini (2012) find that online political discussions generate more negative emotions than do face-to-face interactions.

At the same time, however, online interactions may not be so fundamentally different from in-person ones as to inhibit attitude change or productive discussion. There is little evidence for a psychological mismatch between online and offline environments (Bor and Petersen 2022), online discussions attract more ideological moderates and exhibit greater political and racial diversity than offline discussions (Baek, Wojcieszak, and Delli Carpini 2012), and the basic structure and content of political discussion is similar regardless of whether it is in-person or computer-mediated (Settle 2018).

To assess the potential for argumentative exchanges to change attitudes in online political discussions, I retrieve over one million comments posted on the Reddit social media platform over two years and use computational models to annotate them for characteristics of argumentation. Results suggest consistent support for an argumentative theory of political reasoning: Attitude change is most likely to occur when, for example, interlocutors address each other directly and formulate high-quality counterarguments.

I make three main contributions in this paper. First, I develop a theory of the conditions under which we should expect reasoning to lead to attitude change on political issues. In doing so, I aim to turn the prevailing notion of political reasoning on its head and outline an approach to understanding attitude change that takes the biased and lazy nature of human reasoning as an asset to leverage rather than an obstacle to overcome. Second, I demonstrate the inferential advantages of conducting research on Reddit, an underutilized data source in political science and psychology. Finally, I show how methodological advances in deep learning can assist scholars in understanding complex concepts in political speech.

Political Reasoning and its Motivated Nature

The idealized model of reasoning most common in political science is a *homo politicus* one (Crawford 2009). It suggests that citizens should be able to speak and think about political topics in a level-headed manner, are well-informed such that they can provide and assess evidence-based claims for or against a policy proposal, think and reason rationally, and are capable of arriving at the most logical and objectively optimal conclusion via this process of level-headed discussion and information consideration. As the amount and quality of information available to citizens increases, so the theory goes, so should the quality of conclusions citizens can reach (Bohman 1996; Cohen 1997; Gutmann and Thompson 1996; Somin 2010).

The empirical record, however, is mixed. Some scholars find that political discussions adhering to this idealized model have positive impacts, increasing attitude strength, promoting consensus, and decreasing ideological polarization in the short-term (Becker, Porter, and Centola 2019; Druckman and Nelson 2003; Esterling, Fung, and Lee 2021). Others argue that they can actually exacerbate polarization or otherwise fail to promote more educated, well-reasoned opinions (Jackman and Sniderman 2006; Levendusky, Druckman, and McLain 2016; Mendelberg and Oleske 2000).[1]

Indeed, this is the conclusion advanced by Taber and colleagues. They argue that the processing of political arguments is motivated, and that both confirmation (seeking out confirmatory evidence) and disconfirmation (being skeptical of counterattitudinal arguments while uncritically accepting confirmatory arguments) biases lead to attitude polarization (Taber, Cann, and Kucsova 2009; Taber and Lodge 2006). Results from studies investigating the effect of the media—the source of most political information in the present day—paint a picture of a citizenry that seeks out information reinforcing their existing beliefs if they are not pressured to do otherwise and rationalizes what little countervailing information they do consume to further support those beliefs (Hopkins 2014; Iyengar and Hahn 2009; Prior 2013).

[1]On group polarization, see Arima (2012), Myers and Bishop (1970), and Sunstein (2002), among others.

Dual-process theories—a key framework for understanding the psychology of reasoning—suggest a deeper explanation for these findings. These theories separate cognition into Type 1 (quick, subconscious, and intuitive) and Type 2 (slow, conscious, and rational) and are often concerned with the ways in which preconscious processing can shape and distort conscious reasoning and decision-making (Evans and Stanovich 2013; Evans 2003; Kahneman 2011). Scholars find that the reasoning enabled by Type 2 cognition—supposedly conscious and rational—is instead prone to motivated reasoning (Ditto and Lopez 1992; Kunda 1990), frequently resorts to fast-and-frugal heuristics instead of slow-and-deliberate ratiocination (Chaiken 1980; Kahneman, Slovic, and Tversky 1982), and is barely more skilled at prediction and forecasting than the proverbial dart-throwing chimp (Mellers et al. 2015; Tetlock and Gardner 2015). These findings suggest that humans possess limited capacity for reasoning and the reasoning we do perform is unreliable.

Scholars studying political reasoning, specifically, give a similarly grim prognosis. Framing and priming effects abound (Bizer and Petty 2005; Chong and Druckman 2007; Tesler 2015), and individuals' political attitudes seem so sensitive to these effects that an influential strand of public opinion theory argues that most citizens lack coherent ideological principles in general (Kinder and Kalmoe 2017; Zaller and Feldman 1992; Zaller 1992). More perniciously, partisan and racial stereotypes are common (Ahler and Sood 2018; Gaertner and McLaughlin 1983), and they encourage citizens to reason in a politically motivated manner (Erisen, Lodge, and Taber 2014; Jost et al. 2003; Munro, Weih, and Tsai 2010). In a vein of the dual-process tradition, Lodge and Taber (2000) develop a model of political reasoning that incorporates numerous heuristics and biases and asserts the central importance of hot cognition: the theory that reasoning is colored by an individual's affective state.[2]

In sum, many scholars argue that politics enjoys limited immunity to the "flaws" in human reasoning posited in the psychology literature (Bolsen, Druckman, and Cook 2014; Gaines et al. 2007; Sniderman, Brody, and Tetlock 1991). Working in the predominant paradigm of political reasoning, then, one might conclude that citizens are poorly equipped to produce coherent reason-

[2]See also Taber, Cann, and Kucsova (2009) and Taber and Lodge (2016).

ing on political issues, resistant to changing or even forming sensible opinions, and unlikely to engage in nuanced consideration of diverse ideas.

An Argumentative Theory of Political Reasoning

In contrast, I present a second framework, which I call "argumentative" political reasoning. This theory departs from existing theories in a major way: I consider it an empirical fact that human reasoning is biased and lazy. In the idealized model, these are *flaws* of reasoning because they complicate individuals' ability to cogitate rationally and objectively. In the argumentative framework, however, they are *features* of reasoning because they perform important functions.[3] This theory therefore takes as its point of departure the empirical literature that makes human reasoning look flawed and incorporates those findings as its core assumptions.

I define "argumentation" as giving, receiving, and responding to reasons in an iterative exchange. This type of interactive reason-giving likely approximates the environment in which human reasoning developed (Cosmides 1989), and recent theoretical advances in cognitive psychology have pinpointed these interactions as critical for forming and updating attitudes (Mercier and Sperber 2011; Mercier and Sperber 2017). I contend that argumentation, an active manifestation of this ability to evaluate competing streams of information, is more effective at changing political attitudes than solitary consideration or even casual discussion. This is chiefly because argumentation pushes individuals to recognize the potential pitfalls of their own argument as well as the merits of different ideas. Evidence from psychology supports this assertion: Stronger, less fallacious arguments are more likely to lead to attitude change, even if they contradict prior beliefs (Hahn and Oaksford 2007; Koenig 2012; Trouche, Sander, and Mercier 2014).

The mechanisms linking argumentation to attitude change are the familiar forms of bias and laziness so thoroughly documented in the political psychology literature. Humans are cognitive misers by default; the brain is a hungry organ and will avoid expending energy unless pushed to do so (e.g. De Neys, Rossi, and Houdé 2013; Donald 1991). We are thus *lazy* in our genera-

[3] See Haselton, Nettle, and Andrews (2005).

tion of reasons to justify our opinions and behavior (Mercier and Landemore 2012). The optimal argument is the one that minimizes energy expenditure while maximizing the probability of our idea being accepted by others.[4] The ideal strategy to produce these arguments is therefore to offer the lowest-quality arguments first. If they are accepted, there is no need to spend time and energy coming up with more sophisticated ones. If they are rejected, we should gradually offer more and more complex arguments, addressing more and more of our interlocutor's rebuttals until one of those arguments is finally accepted. For similar reasons, we should expend relatively little energy on anticipating counterarguments, primarily because one argument can elicit a large number of potential counterarguments. Instead of anticipating and responding to all these counterarguments—especially when some might be unconvincing or things our interlocutor never would have mentioned in the first place—it is more efficient to instead use the iterative nature of argumentation to discern which types of arguments are effective, which are not, and which sorts of counterarguments must be addressed.

We are also *biased* in our evaluation of reasons given by others (Trouche et al. 2016). On average, ideas or actions supported by only low-quality arguments are less likely to be worth serious consideration compared to ones supported by high-quality arguments. In fact, ceteris paribus, the best idea is likely to be the one for which the best arguments can be generated. This implies that knowingly accepting anything less than the highest-quality arguments possible entails a higher likelihood of reaching a suboptimal conclusion. The optimal strategy, therefore, is to only accept high-quality arguments that provide substantial reasons to believe in the veracity of our interlocutor's argument.

Evaluating high-quality reasons for an argument contrary to our predisposition while simultaneously being forced to generate high-quality reasons for our own argument are thus two components of the environment in which we should expect attitude change to occur.[5] Both dynamics must

[4]On the tradeoff between information quality and energy expenditure, see Beach and Mitchell (1978), Payne, Bettman, and Johnson (1988), and Stigler (1961).

[5]One might be concerned that exposing individuals to counterattitudinal information would cause a backlash effect, leading them to adhere more tightly to their previous beliefs (Redlawsk 2002), but a wide range of recent work suggests this effect does not occur (Guess and Coppock 2020; Porter and Wood 2021; Wood and Porter 2019).

be present: Without someone to refute your own argument, you will not engage deeply with the reasons for holding your own opinions, and without challenging your interlocutor on their ideas, the reasons they give are unlikely to be convincing.[6] This is precisely the reason why engaging in casual political talk, consuming news media, or otherwise participating in solitary consideration of political issues rarely promotes attitude change, even if multiple viewpoints are present.

Previous analyses of deliberative polling transcripts and laboratory experiments designed to evaluate these types of communication provide preliminary evidence for the persuasive power of argumentation relative to more casual political discussion. Westwood (2015) is principally concerned with determining whether opinion change in group discussion of political issues is driven by reason-based interpersonal persuasion or information-based increases and refinements in political knowledge. Using transcripts from an online deliberative poll conducted in the United States,[7] he combines content and network analysis to model argument quality and the flow of those arguments between group members.

Consistent with an argumentative theory of political reasoning, results show that arguments presented to individuals in direct debate are of higher quality than those directed at the group more generally, high-quality arguments directed at individuals in direct debate are the strongest predictor of opinion change, and this opinion change occurs in the direction of the persuader's attitude (meaning that neither group polarization nor backlash effects occur in response to counterattitudinal arguments).[8] Moreover, political knowledge appears to have no bearing on opinion change, implying that high-quality arguments presented in direct debate with an interlocutor are persuasive regardless of either participant's level of political knowledge before or after the interaction.[9]

[6] Another possible benefit to encouraging the generation of high-quality arguments is that it helps break the illusion of explanatory depth, wherein individuals believe they understand the nuances of an issue or policy but, when asked to explain it mechanistically, are forced to reckon with the fact that their knowledge is actually quite limited (Fernbach et al. 2013), cf. Crawford and Ruscio (2021).

[7] On deliberative polling, see Fishkin (2009), Luskin, Fishkin, and Jowell (2002), and Luskin et al. (2014).

[8] Gerber et al. (2014) examine a separate deliberative poll conducted in Europe and arrive at similar, though slightly more mixed, conclusions.

[9] These findings comport with the theoretical argument made by Lupia (2002), who argues that the emphasis placed on information acquisition in deliberative environments is misguided. Left to their own devices, subjects presented with new—even balanced—information will merely rationalize that information in accordance with their predispositions.

Schneiderhan and Khan (2008) contribute similar findings from a laboratory experiment. Their first treatment arm, "discussion," entails free-flowing interpersonal conversation that typically involves the exchange of information but does not necessarily involve any form of reason-giving or argumentation. Their second treatment arm, "deliberation," is more akin to argumentation, where subjects are explicitly directed to process each other's opinions, provide their own opinions with justifications, and accept conflict in the discussion. Results show a clear advantage for the argument-style treatment. Subjects in this treatment are more likely to change their opinion compared to both the discussion treatment and the control, and the discussion treatment was statistically indistinguishable from the control on the matter of opinion change. The quality of argumentation also matters, as the number of justificatory reasons provided is positively associated with opinion change. Although the discussion topic (segregated student fees) was neither explicitly political nor highly contentious, these results provide additional evidence from a controlled setting for the role of argumentation in reasoning.

Data and Methods

Reddit as a Research Tool: `r/ChangeMyView`

The choice of data source is particularly important in studies of political discussion. On one hand, while experimental and survey methods are common currency in political behavior research, they leave questions as to their external validity (Barabas and Jerit 2010). Political discussion in the modern era is a social encounter (Carlson and Settle 2022); highly controlled experiments may remove many important elements of interpersonal interaction that we could capture by observing political argumentation "in the wild." On the other hand, citizens often opt out of political talk in their real-world interactions to avoid jeopardizing their relationships with friends or family, the people with whom they most often converse (Carlson and Settle 2022). As a result, it may also be difficult to test this type of theory in organic, everyday interactions.

Instead, I turn to data from online social media platforms. These data sources have grown in popularity in recent years, as interactions are plentiful and data are frequently organized into comment-reply structures (Baughan et al. 2021; Habernal and Gurevych 2017; Settle 2018). I leverage Reddit data from a particular subreddit: `r/ChangeMyView`. This subreddit provides users a platform for engaging one another on questions both controversial and anodyne. An original poster (OP) begins a thread by stating their opinion on some topic, justifying that opinion, and inviting others to attempt to change all or part of that opinion. Other users then post comments and replies to either the original post or subsequent comments. This typically results in a deep comment forest with numerous users debating or agreeing with one another over specific points or broader arguments related to the topic.

Importantly, these users are not interacting with their friends and family. Reddit activity is anonymous, meaning users are free to express true opinions with little worry of ostracizing acquaintances more central to their social life. In a 2016 Pew Research Center survey, 84 percent of respondents agreed that social media encourages people to say things about politics they would not say in person and 74 percent said that social media helps bring new voices into the political conversation (Pew Research Center 2016). But online anonymity could be a double-edged sword; it might invite more unserious engagement or reduce the prevalence of high-quality arguments, but it also avoids the physical, tonal, and relational cues humans glean from in-person conversation that could either increase or decrease the effectiveness of argumentation. An anonymous platform is therefore a cleaner, though more simplistic and sterile, environment in which to test the effects of argumentation.

Despite their anonymity, users are not free to say whatever they please. The subreddit employs a set of specific rules and thorough moderation practices. The subreddit is maintained by a dedicated group of moderators who remove posts and comments that do not adhere to the community's rules. For an OP, these rules dictate that the initial post must thoroughly explain their opinion in no fewer than five hundred characters, they must genuinely hold that opinion (i.e. they are not playing devil's advocate) and be open to it changing, their stance may not be neutral, and they must return to the post within three hours to extensively engage with their interlocutors. For non-OP users, comments given in direct reply to the original post must challenge the OP's opinion with justification, they must refrain from making accusations of bad faith, and each comment must contribute meaningfully to the conversation (i.e. they may not submit posts with only links, jokes, or non-substantive statements of agreement). All users are additionally held to common standards of online etiquette; they may not advocate for harm or make rude or hostile comments. Posts and comments not adhering to these rules are liberally removed by the moderators.[10]

The result of these moderation criteria is an open, interactive environment that ensures the final data collected from the platform bear close resemblance to the type of data one might expect to receive from a more controlled laboratory experiment, but without the constraints and inorganic interactions imposed by such a study. In some ways, these data may even more closely approximate face-to-face interactions in more typical social settings. Social interactions typically carry a set of informal norms that govern how interlocutors should converse with each other. Failure to adhere to these norms often results in some level of social sanctioning, and this sanctioning system helps keep behavior within expected boundaries or remove individuals who cannot monitor their behavior appropriately. The `r/ChangeMyView` moderation procedures function in much the same way; moderators impose a set of norms, users adhere to the norms under the threat of sanction, and users who refuse to adhere are removed.

Finally, the key benefit to this subreddit is its method of indicating opinion change. When a user—OP or otherwise—recognizes that part or all of their opinion has been changed by a particular post, they award that post a "delta." The user awarding the delta must then explain which part of their opinion was changed by the post and why the post convinced them to change it. Moderators remove posts which award deltas sarcastically, as an expression of agreement, or for any purpose other than genuine opinion change. I use these deltas below to assess whether argumentation is more likely to result in attitude change relative to more casual modes of discussion.

[10] Matias (2019) shows that publicly displaying community norms on online discussion fora decreases harassment and increases the participation rate of new interlocutors.

I scrape all posts, comments, and associated metadata from this subreddit over two consecutive years, from July 1, 2020 to June 30, 2022. This time period encapsulates several contentious and politically relevant events from recent American history, with the 2020 presidential election cycle, the COVID-19 pandemic, and the Black Lives Matter protests among the most significant. These data collection efforts yielded 12,593 posts, each with its own comment forest and discussion. The final comment-level dataset consists of 1,045,599 unique comments.

Detecting Characteristics of Argumentation

To test the potential for argumentation to lead to attitude change in these online fora, I first need to determine whether each of these one million comments engages in debate or a more casual style of discussion. These data needs present at least three challenges. First, these types of discussion characteristics can be complex, difficult to understand, and quite subjective. Identifying high-quality arguments generally involves asking survey respondents or experiment subjects whether they perceive an argument as strong or weak (Eagly and Chaiken 1993), resulting in "an inadequate tautology" that limits scholars' ability to understand public opinion (Druckman 2022, p. 73).

Instead of relying on one "argumentative or not" indicator—the coding of which could reflect multiple distinct ways of thinking about what argumentation looks like in practice—I approach this concept from several different angles, with each indicator capturing one narrow aspect of argumentative speech as summarized in Table 1. The result is a more complete picture of argumentation in online discourse.

I begin by simply identifying whether a comment disagrees or agrees with the comment to which it is replying. Almost by definition, argumentative exchanges are more likely to involve disagreement than causal ones. Second, I capture whether the commenter is directing their comment at an individual or at a broader audience which, on this platform, could be anyone with internet access. I do so by identifying when the commenter "quotes" another individual's comment. Comments that directly engage with another individual's ideas are more likely to be persuasive than those indiscriminately broadcast into the ether.

Third, I identify whether a commenter is seeking to uncover additional information or asserting ideas of their own. Argumentative exchanges characteristically involve the assertion of ideas, so I expect those types of comments to affect attitude change more strongly than those asking questions. Fourth, I distinguish between comments that merely respond to other commenters' arguments and those that present an original argument of their own. This is one of the core characteristics of argumentative discussion that theoretically drives its persuasiveness; comments with an original argument should result in more attitude change than those that simply seek to rebut previous arguments without providing reasons to adopt an alternate viewpoint.

Fifth, I determine whether each comment employs a defeater (direct evidence that a belief is false; McCain, Stapleford, and Steup 2022) or undercutter (evidence that undermines support for a belief but stops short of providing support for an alternate belief; Kelly 2016). Connecting evidence to an alternate viewpoint should be more likely to result in attitude change than merely undercutting evidence. Finally, I distinguish between low- and high-quality arguments, under the expectation that high-quality arguments should be more persuasive than low-quality ones. Table 1 shows the indicators I use to identify argumentation alongside short descriptions of each.

The second challenge presented by the need for large-scale identification of argumentative comments is that training human annotators and paying them to read over one million documents is a process that could take months, cost thousands of dollars, and may not even result in reliable annotations despite those investments. Instead, I turn to natural language processing techniques—automated methods of extracting information from text—to accelerate comment labeling and ensure a greater degree of uniformity in the labeling process (e.g. Grimmer, Roberts, and Stewart 2021; Grimmer and Stewart 2013; Wilkerson and Casas 2017). But there are many such methods that could be suitable for this application—which one is most appropriate?

The final challenge presented by the data provides additional guidance. I could use separate models to identify each of the indicators in Table 1, but that would require a large amount of computational resources and would ignore a great deal of information in the text. That is, the various characteristics of argumentation I extract from these comments likely share a common semantic

Table 1: Summary of Argumentation Indicators

Indicator	**Description**	**Annotation Method**	**Training Data**
Disagreement	Does the commenter express disagreement or agreement?	Deep Neural Network	Internet Argument Corpus
Object of Address	Does the commenter direct their comment at a specific individual or the broader audience?	Quote Detection	None
Question vs. Assert	Does the commenter probe for more information or assert their own ideas?	Deep Neural Network	Internet Argument Corpus
Counterargue vs. Rebut	Does the commenter present an argument of their own or merely rebut a previous argument?	Deep Neural Network	Internet Argument Corpus
Scope of Argument	Does the commenter contradict the entirety of an interlocutor's argument or just undercut one piece of evidence?	Deep Neural Network	Internet Argument Corpus
Quality of Argument	Does the comment express a clear, relevant, and well-reasoned argument?	Deep Neural Network	IBM-Rank-30k

and logical structure; a single comment could exhibit multiple characteristics, and knowing how these characteristics tend to co-occur is likely to carry implications for how characteristics are identified. To take advantage of this information, I apply a multi-task learning approach to training a deep neural network. Farzam et al. (2024) develop a multi-task neural network architecture for extracting complex concepts from unstructured text. They show this multi-task model is more efficient and more accurate than training a series of separate models, and they show evidence that different dimensions of argumentation do, indeed, share a great deal of information in common.

I train this multi-task model to classify texts on five of the six characteristics summarized in Table 1. I retrieve embeddings—a numeric representation of text that can be used in modeling—using bidirectional encoder representations from transformers (BERT), a large language model used in a wide variety of high-profile products such as Google Search (Devlin et al. 2019). BERT is pre-trained on English Wikipedia and the BooksCorpus (Zhu et al. 2015), which collectively provide a training corpus of over 3.3 billion words.

For four of these tasks, I draw training data from the Internet Argument Corpus, a collection of approximately 28,000 posts extracted from several online debate and discussion fora very similar to `r/ChangeMyView` (Abbott et al. 2016; Walker et al. 2012). The discussions in this corpus cover a variety of controversial topics relevant to politics and social life in the United States, such as same-sex marriage, gun control, and the existence of God. This diversity of issues is especially useful for training domain-general classifiers, as it prevents the model from over-fitting on words or phrases relevant to specific topics. For the fifth task, argument quality, I draw training data from IBM-Rank-30k, a corpus of approximately 30,000 crowd-sourced arguments across a similarly diverse set of 71 common topics (Gretz et al. 2020). Both sets of training data are coded on each characteristic by 5-10 human annotators. The Supplementary Information provides additional details on data, model design, and training.

Finally, one indicator—object of address—does not draw its annotations from the neural network. This indicator captures whether a commenter is responding to or engaging with an interlocutor directly, as opposed to making comments for a broader audience. The structure of the `r/ChangeMyView` text data allows me to detect this directly. Commenters can respond directly to previous comments by "quoting" them. These quotes are copied into the commenter's new post in a text box set off from the rest of the post. If a commenter quotes a previous post, I code that comment as addressing an individual. If not, I code it as being meant for the audience as a whole.

Drawing Inferences from Machine-Proxied Data

I noted above the benefits conferred by using machine learning methods to annotate text data. However, these methods also impose an additional challenge. The annotations they produce are only *proxies*, denoted here by $\hat{X}$; the true values of these indicators X are unobserved and, in many cases, unobservable. Using $\hat{X}$ to proxy for X in inferential statistical analyses requires special care, as it causes two problems.

First, using learned proxies as explanatory variables leads to attenuation bias, pushing coefficient estimates toward zero (Wooldridge 2015). However, unbiased estimates of the effect of X on Y can still be drawn under three assumptions (VanderWeele and Hernán 2012):[11]

1. **Positive average monotonicity**: When X is larger, $\hat{X}$ is also larger, on average. This is an especially weak assumption when using supervised learning methods, as the relationship between $\hat{X}$ and X can be empirically assessed. Performance metrics in the Supplementary Information suggest this assumption holds.

2. **Perfect observation of outcome and covariates**: The outcome Y and covariates W are perfectly observed. Assuming covariates are the result of sincere interactions with the platform (i.e. users did not upvote a comment on accident), this assumption is likely to hold; Y and W are not proxied and instead represent directly observable quantities available in Reddit metadata.

3. **Specification**: The covariates W are correctly specified. This is the strongest assumption in this context, but the anonymity of Reddit ameliorates many specification concerns; although it is impossible to adjust for variables that often affect political discussion—such as race, gender, or partisanship—discussion participants also lack this information about each other, decreasing the likelihood that they will exert an effect on either $\hat{X}$, X, or Y.

Second, using learned proxies will result in downward-biased standard errors, as results from the generalized linear models that I present below incorporate two sources of sampling variability. The first source of sampling variability comes from drawing one of many possible datasets for training the neural network to estimate $\hat{X}$. The second comes from drawing a sample of observations from a larger population of possible observations to estimate the effect of $\hat{X}$ on Y.[12] That is,

[11]Even under these assumptions, it is not necessarily true that a failure to find an effect suggests the lack of such an effect, owing to the familiar issues of statistical power and the nature of null-hypothesis significance testing (Knox, Lucas, and Cho 2022). With such a large dataset, however, concerns about power are likely unfounded in this case.

[12]The dataset includes all comments posted to `r/ChangeMyView` in the specified time period so, in some sense, the full dataset does not represent a sample, but rather the population. However, having well over one million observations in a linear model may lead to concerns that any statistically significant findings are merely the result of an extremely

uncertainty enters the empirical analysis in both the measurement model and the linear model, and both sources must be accounted for.

To do so, I follow Knox, Lucas, and Cho (2022), who recommend a bootstrap approach to estimating coefficient estimates with appropriate standard errors when the key explanatory variable is a learned proxy.[13] This bootstrap has two stages. In the first stage, I randomly resample the training set with replacement 500 times, re-train the neural network with each resampled training set, and estimate $\hat{X}$. In the second step, I randomly resample the final Reddit dataset with replacement 500 times, re-fitting the substantive models below with each resampled dataset. This procedure results in 250,000 total coefficient estimates, from which accurate standard errors can be calculated.[14]

Drawing Inferences from Rare-Events Data

One final methodological consideration is in order before moving to substantive results. The dependent variable, attitude change, is represented here as binary—either a comment succeeds in changing someone's opinion or it does not. I thus rely on binomial logits to estimate the effect of argumentative characteristics on attitude change. Attitude change, however, is a rare event; descriptive analyses presented below show that only 1.7 percent of all `r/ChangeMyView` comments are awarded a delta, and 36.2 percent of all comment forests award no deltas at all. Logistic regression on such unbalanced data leads to downward-biased estimates of event probabilities (King and Zeng 2001). I therefore adopt a penalized maximum-likelihood approach to correct for this bias. Firth (1993) proposed using the Jeffreys prior—the square root of the determinant of the Fisher information matrix—to penalize the log-likelihood function. This penalty has been shown to reduce bias in logistic coefficient estimates in the cases of small samples, rare events, and complete

large sample size, so I sample only 10 percent of the total number of observations. This practice also carries the practical benefit of dramatically decreasing the computational cost of fitting the model. The Supplementary Information displays results for the full dataset as well as results from post-level sampling instead of comment-level sampling.

[13]Though not directly applicable, see also Fong and Grimmer (2023) on estimating causal effects with latent treatments.

[14]To reiterate, in the second step, I randomly sample 10 percent of the total comments before fitting each iteration of the model.

or quasi-complete separation (Heinze and Schemper 2002; Puhr et al. 2017). All results derived from binomial logits below use this penalized maximum-likelihood framework.

How Do People Discuss Politics Online?

This section presents a brief descriptive analysis of the `r/ChangeMyView` data and what the neural network classifiers reveal about how citizens discuss contentious issues in online spaces. Figure 1 shows how an original post appears to users on `r/ChangeMyView`. The OP provides a post title stating their opinion, followed by a lengthy explanation of why they hold that opinion. The icon in the bottom left corner communicates that 414 comments have accumulated in the discussion tree below, and the orange badge underneath the title indicates that at least one of those comments has received a delta from the OP.

Figure 2 displays a brief excerpt of the discussion that followed the post in Figure 1. A couple features are noteworthy: First, underneath each commenter's username, the gray badge—referred to as "author flair"—shows users how many deltas that commenter has received since beginning their participation in the subreddit. Second, the discussion tree structure—denoted by the nested vertical lines on the left side of the user interface—makes it straightforward to follow the flow of the conversation: `shimmynywimminy` makes the topline comment responding directly to the OP, `Flufflebuns` (the OP) responds to `shimmynywimminy`, and `Unusual_Swordfish_40` makes a comment in response to `Flufflebuns`' rejoinder. Finally, the comment from `Unusual_Swordfish_40` demonstrates how users can take advantage of quotes to communicate directly with other users; the line of text set off in a text block is a direct quote from `Flufflebuns`' preceding comment, and `Unusual_Swordfish_40` responds directly to this quoted text. Though the discussion continues for several more exchanges, this example shows how users interact with the OP and how those interactions appear in a discussion tree. The tone and content of this exchange is generally representative of most other discussions on the forum.

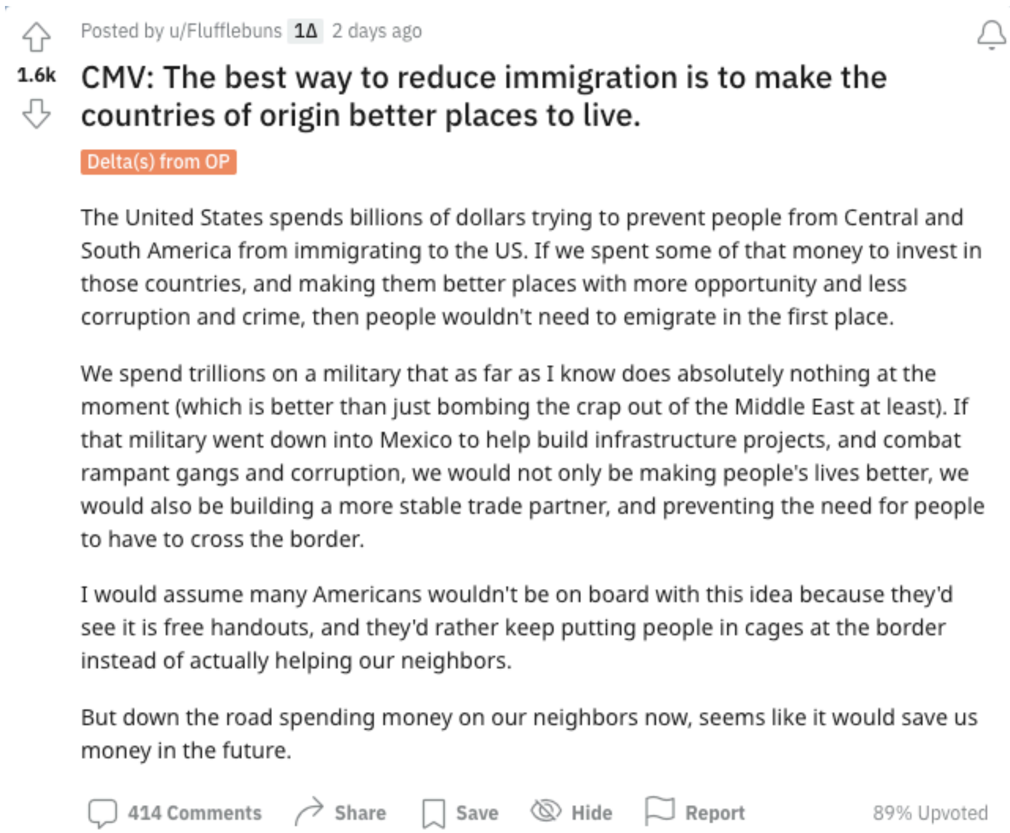

Figure 1: Example of Original Post on `r/ChangeMyView`

I turn now to summary data on how discussions unfold. Figure 3 contains four plots with information on post-level (i.e. discussion-level) characteristics. Plot A displays the number of deltas awarded in each post, plot B displays the number of unique individuals participating in the discussion, and plots C and D display the number of comments overall as well as the number of comments made by the OP. Red vertical lines locate the mean of each distribution, and labels on each plot report the mean, median, standard deviation, and range.

These aggregate statistics suggest a couple key takeaways. First, online discussions in this context appear similar to political discussions that occur face-to-face (see, for example, Carlson and Settle 2022). Distributions in Figure 3, plots C and D are heavily right-skewed, reflecting the fact that most discussions are relatively brief. The median post contains 51 comments—12 of

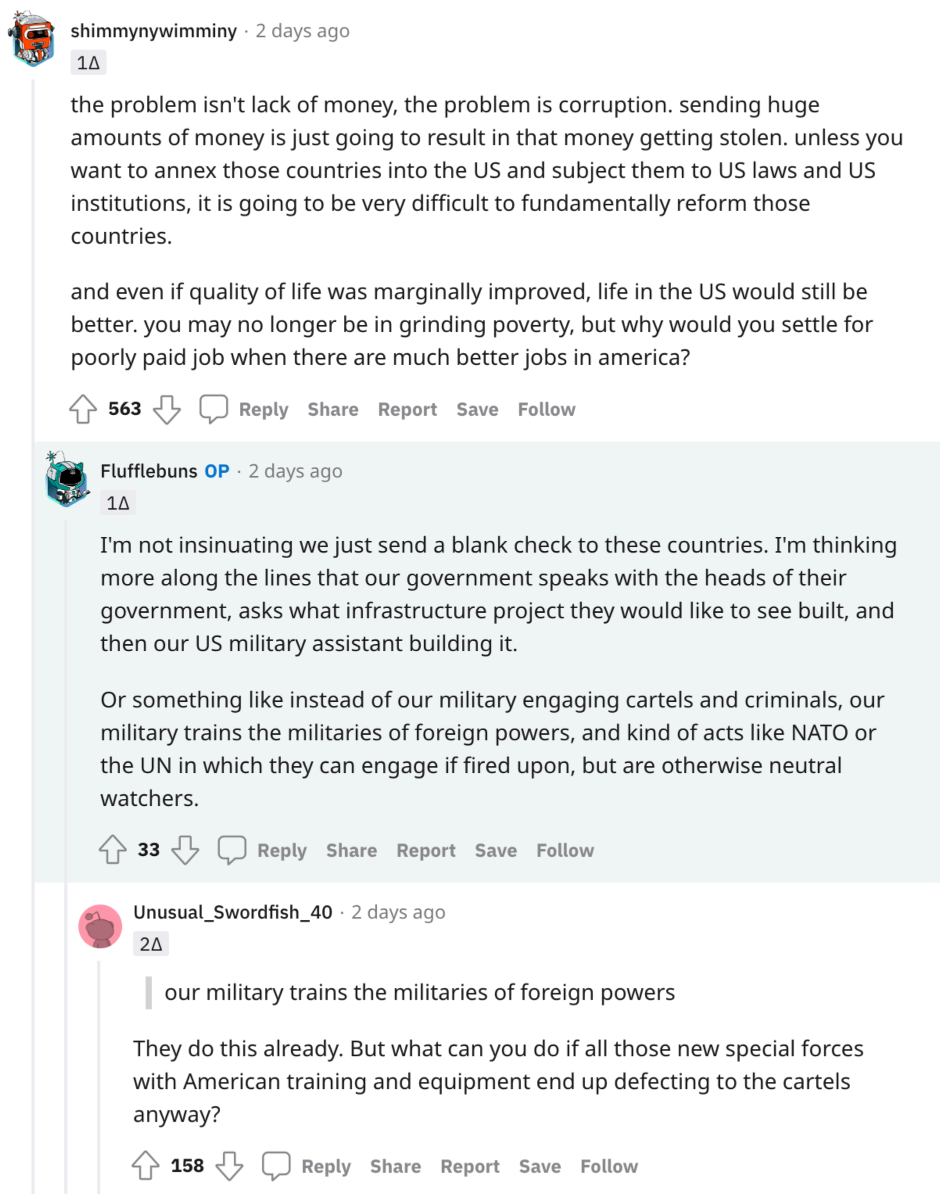

Figure 2: Excerpt of Typical `r/ChangeMyView` Discussion

which are by the OP—and these comments are typically spread among several unique threads.[15] Figure 3, plot B is also heavily right-skewed, suggesting that these discussions typically occur in small- to medium-sized groups, with a median of 20 people participating. Typically, each person in these groups contributes only a few points to the discussion before leaving, but there is a small

[15] Original posts contain 323.3 words on average, while comments average 80.3 words.

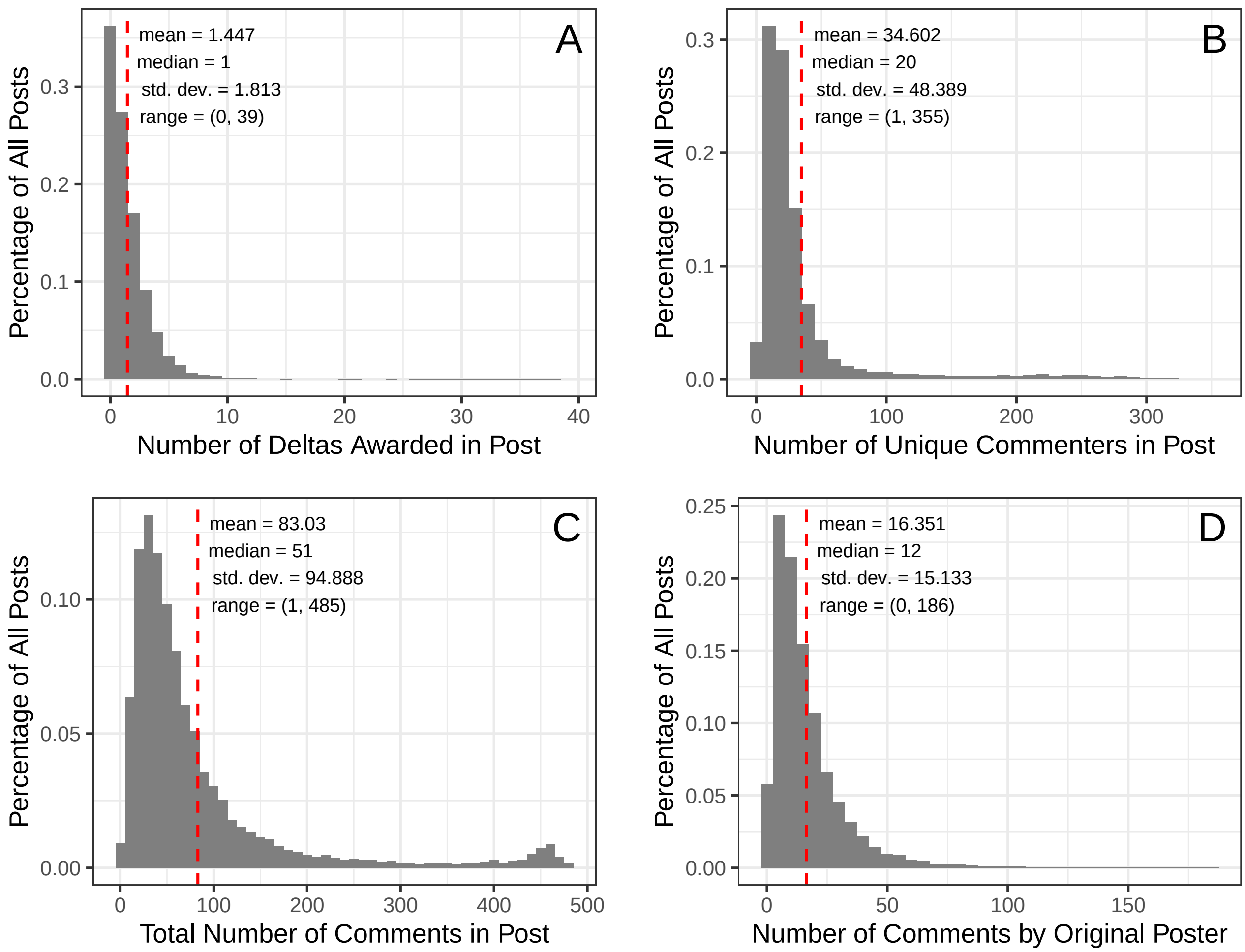

Figure 3: Descriptive Statistics on Discussion Participants and Deltas. Vertical red line shows mean of each distribution.

cadre of enthusiastic, heavily involved commenters. The mean number of comments contributed to one post by a commenter is approximately 2.4, but the observed range runs as high as 186.

Second, `r/ChangeMyView` rules stipulate that OPs must be open to their opinion changing, but Figure 3, plot A suggests that attitude change is still not a common occurrence. In fact, the modal outcome is for no deltas to be awarded anywhere in a comment forest. Narrowing the focus to only those deltas awarded by OPs, the likelihood of a delta being awarded drops further: 38.8 percent of posts never receive a delta from the OP, suggesting that nobody was able to successfully change even part of the OP's opinion. Assessed as a response to individual arguments, opinion change appears even rarer: Only 1.7 percent of comments are awarded a delta, and only 1.6 percent of comments are awarded a delta by an OP.

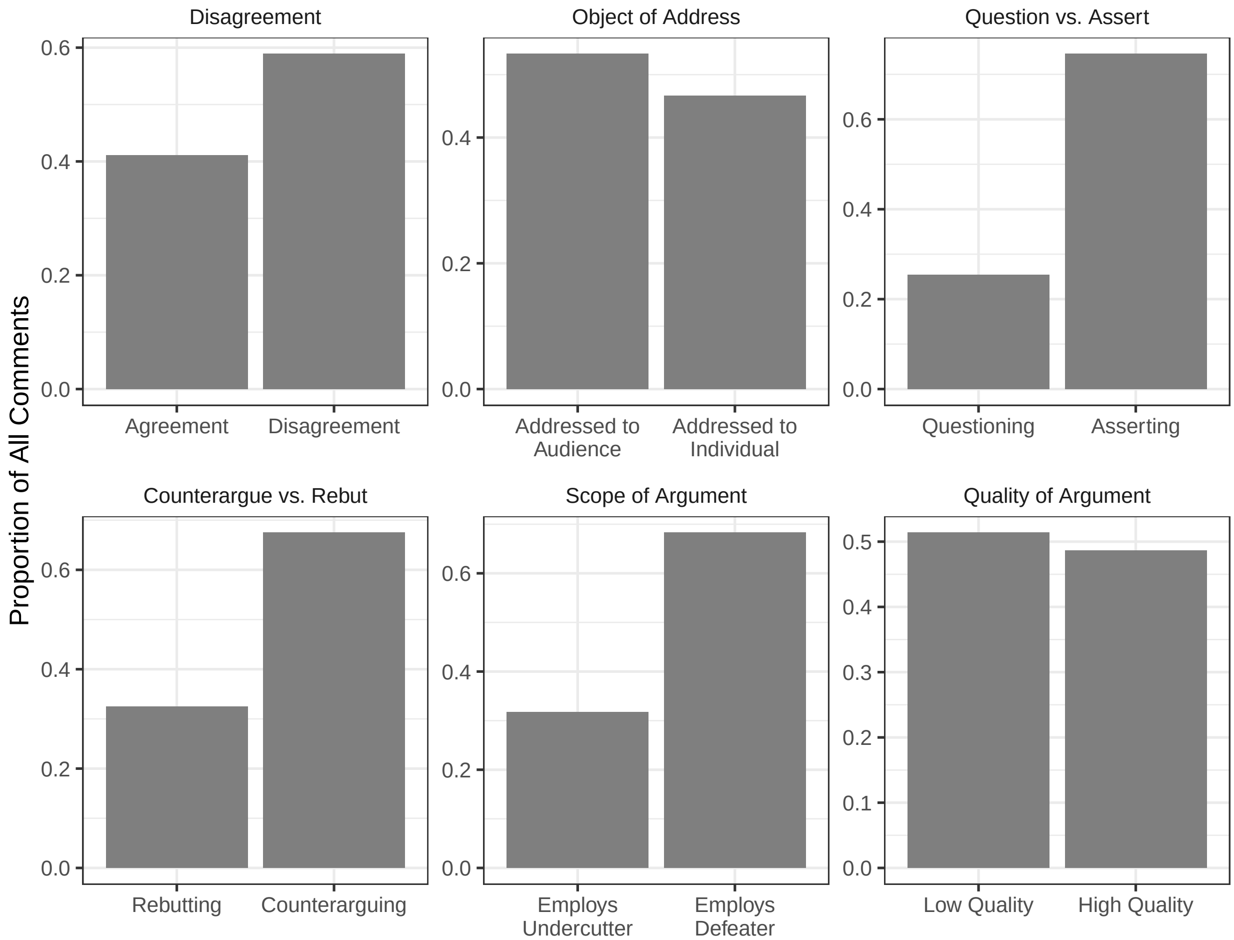

Figure 4: Class Frequencies of Argumentation Characteristics. Frequencies broken down by commenter type are presented in the Supplementary Information.

To get a sense of how prevalent argumentative characteristics are in these data, Figure 4 shows class frequencies for each of my key explanatory variables. Each exhibits substantial variation, providing reassurance that `r/ChangeMyView` is not merely hosting debates where every comment is a counterargument. Indeed, over 40 percent of comments are coded as expressing agreement with a previous comment, and approximately 25 percent of comments merely probe for more information rather than making any assertions of their own. In addition, slightly over half of all comments are *not* directed at a specific interlocutor, indicating that the majority of these interactions are not characterized by back-and-forth argumentation.

In addition to ensuring adequate variation on dependent and explanatory variables, these descriptive statistics speak to some concerns about selection bias. First, if people who are unusually open to changing their opinions self-select into this type of online forum, any apparent effects of argumentation might simply represent attitude change that would have happened no matter what. However, the rarity of attitude change depicted in Figure 3, plot A helps alleviate this concern. Although some participants may have come to these discussions more open-minded than the average politically engaged citizen, it does not appear that they were especially likely to change their opinion.

Second, this subreddit might select for users who enjoy debate and confrontation. If users almost exclusively employ argumentative tactics, a lack of variation would make it difficult to discern the effects of that argumentation. Figure 4 shows, however, that participants exhibit a fair degree of variation on each of the explanatory variables; forty percent of comments express agreement—hardly an argumentative trait—and over thirty percent refrain from advancing an original argument of their own. `r/ChangeMyView` participants may be slightly more argumentative, on average, than the general public, but they are not overwhelmingly so.

These descriptive statistics and the self-selection issues that may characterize online discussions carry implications for the interpretation of results. If participants are reluctant to change their attitudes even in this anonymous online forum where persuasion attempts are encouraged, this analysis might represent an especially hard test of the argumentative theory of political reasoning. Further, online or anonymous communication might invite a greater degree of unserious engagement than one might expect from in-person conversations. The moderation policies of this subreddit are designed to filter out much of this sort of engagement, but that which nevertheless makes its way onto the forum might dilute the meaningfulness of discussions and make it more difficult to maintain a civil, cogent conversation. These largely unavoidable features of online data likely contribute further to this representing a hard test of argumentative reasoning.

Can Argumentation Lead to Attitude Change?

I now turn to explicitly testing key theoretical implications. I focus on the six characteristics of argumentation described in Table 1 and displayed in Figure 4: disagreement, object of address, question or assertion, rebuttal or counterargument, and the scope and quality of arguments. Each of these variables taps a slightly different concept related to argumentation and, taken together, they paint a holistic portrait of the effect of argumentation on attitude change.

All models presented below use penalized maximum-likelihood to fit binomial logits, and all include three covariates. First, next to their username, `r/ChangeMyView` displays the number of deltas each commenter has been awarded over the course of their entire tenure on the subreddit. I control for this value to account for the possibility that discussion participants may view this record of deltas as a source of credibility and therefore be more likely to be swayed by commenters who have been awarded many deltas in the past. Second, the location of a comment in the comment forest may affect the likelihood that it receives a delta. In particular, the deeper in the comment forest a comment appears, the less likely that discussion participants will see that comment and award it a delta. I therefore control for the depth of each comment in the comment forest. Third, like most other subreddits, `r/ChangeMyView` allows users to up-vote or down-vote comments, and the balance of these votes is displayed next to each comment. While this voting process does not indicate attitude change, discussion participants may take the relative balance of those votes as evidence of a comment's quality (or lack thereof) and be more (or less) likely to be swayed as a result, so I include the comment's overall score as a control.

Effects of Argumentation Characteristics on Probability of Attitude Change

Figure 5 displays the predicted probability of a comment resulting in attitude change, conditional on each characteristic of argumentation, which are each presented in a separate facet.[16] To place results in context, horizontal red lines on each facet show the baseline probability that any

[16]Note that each characteristic is evaluated in a separate model.

given comment will result in attitude change, conditional on the same covariates. Error bars denote 95 percent confidence intervals, with standard errors calculated using the process outlined above.

I begin with the broad hypothesis that higher levels of disagreement should be more likely to lead to attitude change. Disagreement itself does not imply argumentation, but it is likely a necessary element of argumentation; it makes little sense to critique a series of arguments if the interlocutors hold similar attitudes. Indeed, the idea that individuals are likely to uncritically accept statements with which they agree is one of the central findings of research on motivated reasoning (e.g. Bolsen, Druckman, and Cook 2014; Lebo and Cassino 2007; Stanley et al. 2020b). It is also an important assumption for the theory I present above. This test therefore serves as both a first cut at gauging the feasibility of the theory, as well as a validation check of sorts. Consistent with previous literature, comments expressing agreement are less likely to result in attitude change compared to those expressing disagreement. Although the difference in predicted probabilities is not statistically significant at the $p < 0.05$ level, the coefficient estimate on the disagreement indicator (presented in Figure 6 below) is statistically different from zero, suggesting that individuals are more likely to change their opinions when exposed to counterattitudinal views.

Recall that a key definitional component of argumentation is an iterative exchange where interlocutors can process and respond directly to each other's arguments. Comments in which an individual responds directly to another comment and addresses their counterargument toward an individual should therefore be more likely to result in attitude change than comments that are meant for a general audience, even if those comments present high-quality arguments. The second (top middle) facet in Figure 5 suggests this is the case; comments directed at a specific discussion partner are significantly more likely to lead to attitude change, increasing the predicted probability from 2.4 to 3.8 percent. While the magnitude of these probabilities is not substantial, it bears reiterating that attitude change is a rare event, with just 1.7 percent of all comments successfully achieving persuasion. In that sense, a relative increase of 58.3 percent is an appreciable effect size.

The next three tests are concerned with how the *type* of comment affects its persuasiveness. Theoretically, I expect comments to be more likely to lead to attitude change if they assert new

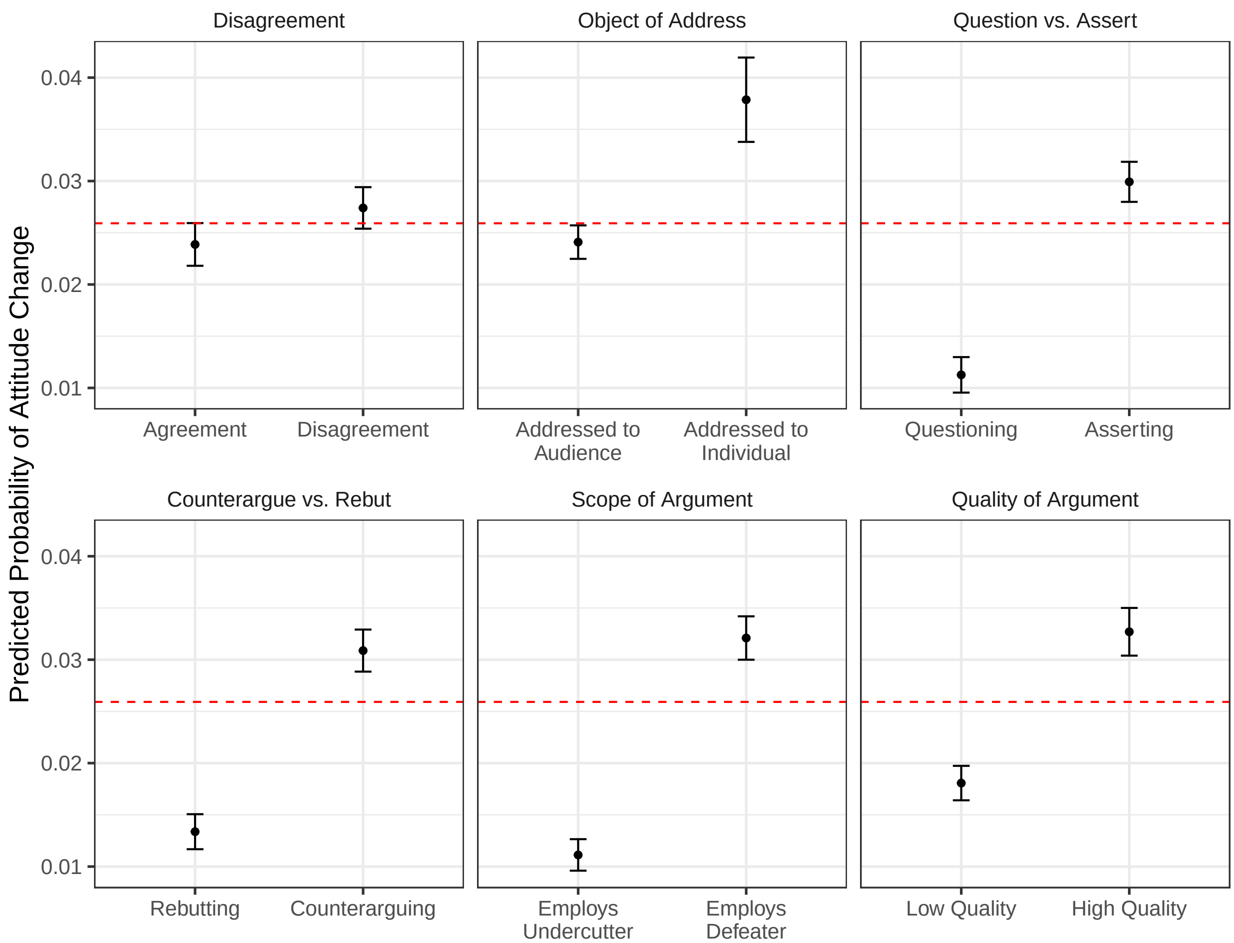

Figure 5: Effect of Argumentation Characteristics on Attitude Change. Red horizontal lines denote baseline probability of a comment resulting in attitude change. Error bars give 95% confidence intervals. Full results available in the Supplementary Information.

ideas instead of probing for information, present original counterarguments instead of merely rebutting other arguments, and use counterarguments that address the entirety of an argument (i.e. a defeater) instead of only critiquing one piece of the argument (i.e. an undercutter). Results of these indicators are presented in the top right, bottom left, and bottom middle facets, respectively, of Figure 5. Not only do results conform to expectations, but it is on these indicators that I observe the strongest effects.

Asserting new ideas is associated with a near-tripling of the predicted probability of attitude change compared to questioning an interlocutor. This indicator provides perhaps the closest fit with the theoretical distinction between casual discussion and argumentation, and results suggest that argumentative styles are substantially more likely to result in attitude change. Similarly, presenting an actual counterargument more than doubles the likelihood that a comment will lead to attitude change compared to merely deflecting other arguments. This is another critical test, as the counterargue/rebut indicator captures a nuanced distinction between argument types. That is, rebuttals might themselves be considered a type of argument, but they do not provide an affirmative reason to adopt a different viewpoint, only a negative reason to reject the opinion their interlocutor already holds. As I detailed in the theory section above, this approach is not likely to be successful, and results corroborate this expectation. Finally, the scope of arguments appears to have a strong association with the likelihood of a comment resulting in attitude change. Arguments aimed at disproving an entire belief (defeaters) are approximately three times as likely to result in attitude change than arguments aimed at removing evidentiary support for a belief (undercutters).

Finally, I examine the effect of argument quality in the last (bottom right) facet of Figure 5. If argumentation had no effect on attitudes, I would expect to see no difference in the persuasiveness of low- or high-quality arguments; the quality of arguments would be irrelevant if argumentation itself was ineffective. Instead, putting forth a high-quality argument is associated with a relative increase of 83.3 percent in the probability of inducing attitude change—moving from 1.5 to 3.3 percent compared to low-quality arguments. Taken together, these results provide support for the argumentative theory of political reasoning across a wide range of indicators, with both statistically and substantively significant effects.

Level of Engagement as a Moderator

Most people who view online fora are "lurkers;" they observe what others are discussing but do not (or only very rarely) contribute to the conversation themselves. Estimates of the proportion of users falling into this category range from fifty to ninety percent, depending on the social media platform (e.g. Lukin et al. 2017; McClain et al. 2021). This feature of online discussion offers an inferential advantage in the context of the argumentative theory of reasoning. OPs—those who engage most in the discussion—should be more affected by the degree of argumentation than are lurkers, who may be less likely to experience the activation of cognitive processes enabling attitude change due to their lack of engagement. Further, recall that the argumentative theory of political reasoning emphasizes the need to participate in an interative exchange. Lurkers, by definition, do not participate in such an interaction. As a consequence, they are unlikely to reap the benefits of the argumentative process.

To evaluate this possibility, I distinguish between deltas awarded by the post's OP, participants (users who comment at least one other time in addition to awarding a delta), and lurkers (users whose only contribution to the discussion is the comment in which they award a delta).[17] A delta awarded by a participant or lurker means the same as a delta awarded by an OP; they are indicating that a comment has changed all or part of their opinion, and they required by subreddit policy to explain which aspect of their opinion changed and why the comment sparked that change.

To gauge the relative size and significance of effects, Figure 6 presents the estimated coefficients on each argumentation indicator, separated by the type of user awarding the delta. Error bars show 95 percent confidence intervals, and the red horizontal line denotes zero. As expected, effects are generally stronger among OPs compared to participants and lurkers. Coefficient estimates for the former are always positive and statistically significant, while those for the latter are sometimes negative, only statistically significant in two cases, and always carry point estimates lower than those for OPs. I take these consistent results as evidence of a moderation effect. One mechanism connecting argumentative interactions to attitude change is the tailoring of counterarguments in response to an interlocutor's arguments, and the close examination of one's own beliefs that those counterarguments prompt. Individuals not participating fully in argumentation are therefore less likely to reap the benefits of them, as they are not prompted to closely engage with the reasons why they hold their own opinions, nor is the discussion likely to generate arguments that speak directly to those reasons.

[17] The Supplementary Information presents class frequencies like those presented in Figure 4 broken down by commenter type.

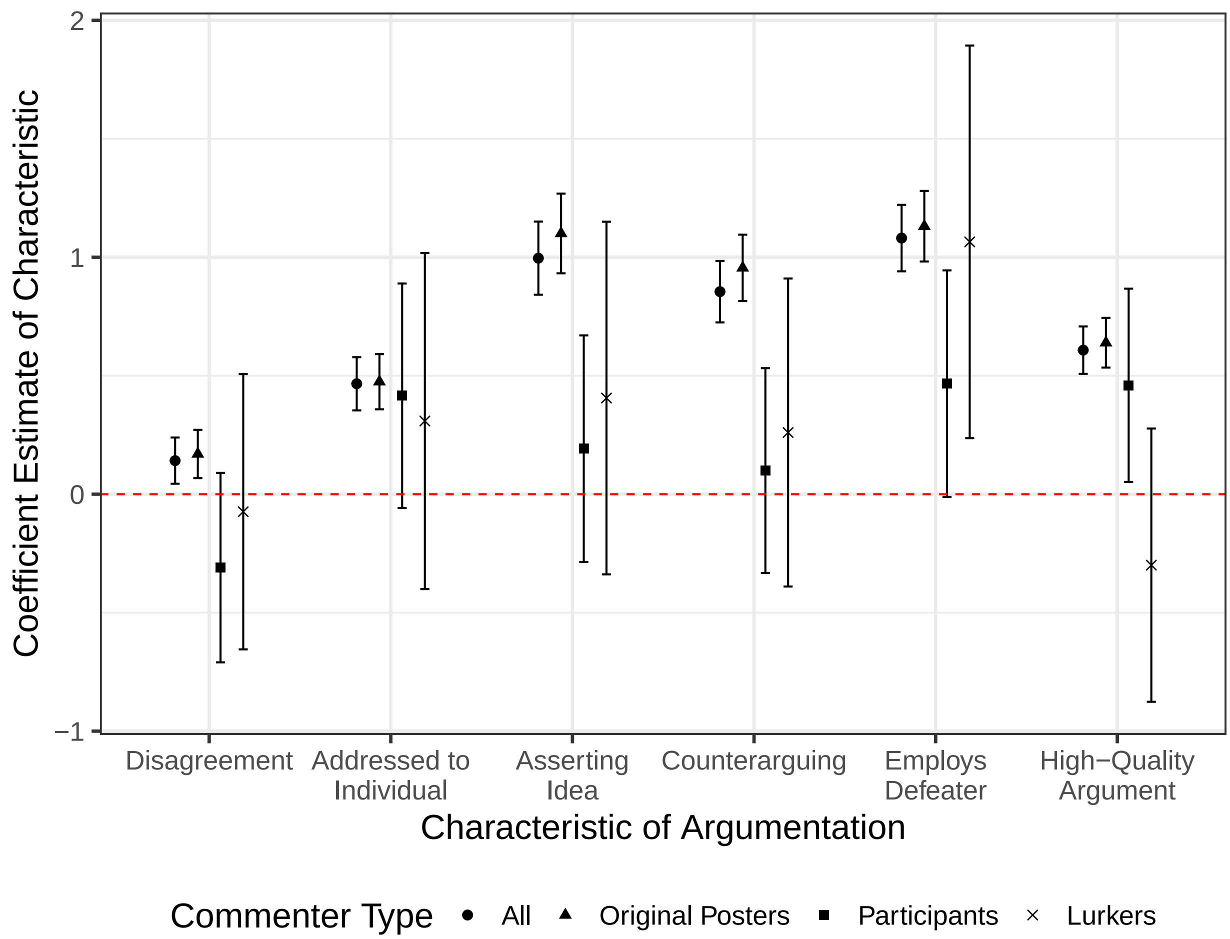

Figure 6: Coefficient Estimates of Argumentation Characteristics. Red horizontal line denotes zero. "Participants" make at least two comments. "Lurkers" only comment to award a delta. Error bars give 95% confidence intervals. Full results available in the Supplementary Information.

Discussion

I have argued that the function of reasoning is to generate reasons justifying one's behavior, evaluate reasons given by others, and determine when and to what extent attitude change is warranted based on the reasons given. The environment in which we should expect attitude change to occur, then, is one that encourages the production of high-quality reasons justifying one's attitude and high-quality reasons rebutting that attitude. Political argumentation is perhaps the only form of interpersonal interaction that naturally constructs an environment with both of these components.

Consider how this form of reasoning contrasts with the idealized style of reasoning more common throughout the political psychology literature, and just how narrow is the range of interactions encompassed by argumentative reasoning. At minimum, argumentation requires two individuals exchanging reasons in iterative fashion, directly responding to one another's arguments and counterarguments. Simply watching a debate—even if the debaters meet this requirement—or reading sets of pro and con arguments do not clear this high bar because they do not invoke direct, two-sided cognitive engagement.

Given the biased and lazy nature of human reasoning, it is not surprising that these forms of information consumption can even lead to attitude polarization, as they do in Taber, Cann, and Kucsova (2009) and Taber and Lodge (2006).[18] In these studies, subjects do not have their beliefs challenged, their own attitudes are partially validated by being presented with reasons favoring those attitudes, and there is virtually no barrier to rationalizing away the counterattitudinal information. Presenting subjects with only counterattitudinal arguments, however, can elicit some attitude change. Gibson (1998) and Sniderman and Piazza (1993) attempt to persuade survey respondents by rebutting their views on political and racial tolerance, respectively. After presenting counterattitudinal arguments, Gibson observes up to 23.5 percent of respondents becoming more politically tolerant and Sniderman and Piazza observe up to 44 percent of respondents becoming more supportive of racial policies—strong results suggesting that presenting counterattitudinal arguments in direct response to a stated attitude can indeed lead individuals to reconsider that attitude, even on hotly contested, moralized issues. The results I present above add to this empirical pattern.

Another reason that non-argumentative studies may lead to attitude polarization is that the quality of arguments in these settings is likely quite low, because presenting subjects with a predetermined set of reasons makes it impossible to address counterarguments in a flexible manner. When presented with a low-quality argument, it is easier for subjects to find fatal problems in the argument such that their confidence in the veracity of their own attitude is actually bolstered.

[18] See also Ansolabehere and Iyengar (1995), Redlawsk (2002), and Stanley et al. (2020b).

This effect is even more pronounced in individuals of high political sophistication, as they have greater domain-specific knowledge and are able to generate more counterarguments. This is why presenting high-quality counterarguments is so critical. It is also why argumentation is necessary; high-quality arguments calibrated to meet the concerns of each individual are unlikely to be produced in any other environment precisely due to the lazy nature of reason-giving.

I suggest that political argumentation provides at least three additional benefits that future research should assess. The first of these benefits is cognitive and the other two stem from the social nature of political argumentation. First, conditional on the provision of high-quality arguments, participants should become less confident in the veracity of their opinion, they should hold that opinion less strongly, and they should be more willing to compromise in the future, even if their actual preferences or ideal points do not move. Second, argumentation should elicit higher-quality justifications for political attitudes by pushing participants to search for more refined reasons to support their predispositions.[19] Finally, it exposes participants to real out-group partisans presenting real out-group partisan arguments; encroaching on partisan echo chambers and isolating out-group partisan arguments from media influences may humanize out-group partisans and decrease the existential threat they are considered to pose to one's in-group. The opposing viewpoint may seem accessible rather than foreign and each side may feel like their voices were heard.[20]

This final prediction may seem a lofty goal in a polarized political environment, but two recent studies provide empirical support. Stanley et al. (2020a) found that partisans believed their counter-partisans were less likely to have good reasons for their political attitudes, and that this expected lack of high-quality reasons spilled over into doubts about counter-partisans' intellectual and moral fortitude. Exposing these partisans to counterattitudinal arguments, however, led them to produce more favorable views of the counter-partisans who produce those arguments. Moreover, this effect was entirely independent of persuasion, suggesting that even if individuals do not change their

[19] Scholars of science education have produced a substantial amount of work on epistemic quality (e.g. Erduran and Jiménez-Aleixandre 2007; Kuhn 1992; Sandoval 2003).

[20] This may partially depend on interlocutors providing each other with "high-quality listening" (Itzchakov, Kluger, and Castro 2017; Kalla and Broockman 2020) in addition to high-quality arguments.

attitudes as a result of argumentation, simply being exposed to high-quality arguments by their interlocutors may decrease the negative affect they feel toward those with whom they disagree.[21]

Dorison, Minson, and Rogers (2019) explicate the link between information search and partisan affect, showing that partisans' unwillingness to voluntarily expose themselves to counterattitudinal information is partially due to overestimating the strength of negative affect they are likely to feel in response to that information. More importantly, correcting that "affective forecasting error" led to greater voluntary exposure to counterattitudinal arguments. Combining these findings with those of Stanley et al. (2020a) suggest a positive feedback loop: Exposure to counterattitudinal arguments via argumentation leads to a decrease in negative partisan affect,[22] this decrease in negative partisan affect leads to increased willingness to engage with counterattitudinal arguments, and so on. Parsons (2010) uses observational data to test these effects and comes away with a clear result: Exposure to political disagreement depolarizes party affect.

In sum, argumentation appears to hold promise for changing political attitudes. Online discussions adhering to an array of argumentative characteristics are substantially more likely to result in attitude change compared to discussions with more casual characteristics. These findings comport with a wide range of related, though theoretically and methodologically distinct, studies showing the value of an interactionist approach to cognitive psychology and an argumentative theory of reasoning. In my view, this suggests reason for optimism. Under the appropriate conditions, humans can, in fact, be quite skilled political reasoners.

[21]It may also be the case that negative partisan affect will decrease as a downstream effect of decreased ideological distance, but Stanley et al. (2020a) do not test this directly.

[22]It should be emphasized that negative partisan affect is distinct from anxiety or negative core affect more generally. The former describes conscious dislike of opposing partisans while the latter two describe internal emotional and affective states. Negative partisan affect should therefore not be construed to regulate information search, openness to persuasion, or persuasiveness as argued above for anxiety and negative affect.

References

Abbott, Rob et al. (May 2016). "Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go With It". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 4445–4452.

Achen, Christopher H. and Larry M. Bartels (2016). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Ahler, Douglas J. and Gaurav Sood (July 2018). "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences". In: *The Journal of Politics* 80.3, pp. 964–981.

Ansolabehere, Stephen and Shanto Iyengar (1995). *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. New York: The Free Press.

Arima, Yoshiko (Apr. 2012). "Effect of Group Means on the Probability of Consensus". In: *Psychological Reports* 110.2, pp. 607–623.

Baek, Young Min, Magdalena Wojcieszak, and Michael X. Delli Carpini (May 2012). "Online Versus Face-to-Face Deliberation: Who? Why? What? With What Effects?" In: *New Media & Society* 14.3, pp. 363–383.

Bail, Christopher A. et al. (Sept. 2018). "Exposure to Opposing Views on Social Media Can Increase Political Polarization". In: *Proceedings of the National Academy of Sciences* 115.37, pp. 9216–9221.

Barabas, Jason and Jennifer Jerit (May 2010). "Are Survey Experiments Externally Valid?" In: *American Political Science Review* 104.2, pp. 226–242.

Baughan, Amanda et al. (Apr. 2021). "Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces". In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1), pp. 1–22.

Beach, Lee Roy and Terence R. Mitchell (July 1978). "A Contingency Model for the Selection of Decision Strategies". In: *The Academy of Management Review* 3.3, pp. 439–449.

Becker, Joshua, Ethan Porter, and Damon Centola (May 2019). “The Wisdom of Partisan Crowds”. In: *Proceedings of the National Academy of Sciences* 116.22, pp. 10717–10722.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee (1954). *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: The University of Chicago Press.

Bizer, George Y. and Richard E. Petty (Aug. 2005). “How We Conceptualize Our Attitudes Matters: The Effects of Valence Framing on the Resistance of Political Attitudes”. In: *Political Psychology* 26.4, pp. 553–568.

Bohman, James (1996). *Public Deliberation*. Cambridge, MA: The MIT Press.

Bolsen, Toby, James N. Druckman, and Fay Lomax Cook (June 2014). “The Influence of Partisan Motivated Reasoning on Public Opinion”. In: *Political Behavior* 36.2, pp. 235–262.

Bor, Alexander and Michael Bang Petersen (Feb. 2022). “The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis”. In: *American Political Science Review* 116.1, pp. 1–18.

Brady, William J., M. J. Crockett, and Jay J. Van Bavel (July 2020). “The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online”. In: *Perspectives on Psychological Science* 15.4, pp. 978–1010.

Carlson, Taylor N. and Jaime E. Settle (2022). *What Goes Without Saying: Navigating Political Discussion in America*. New York: Cambridge University Press.

Chaiken, Shelly (Nov. 1980). “Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion”. In: *Journal of Personality and Social Psychology* 39.5, pp. 752–766.

Chong, Dennis and James N. Druckman (Nov. 2007). “Framing Public Opinion in Competitive Democracies”. In: *American Political Science Review* 101.4, pp. 637–655.

Cohen, Joshua (1997). “Deliberation and Democratic Legitimacy”. In: *Deliberative Democracy: Essays on Reason and Politics*. Ed. by James Bohman and William Rehg. Cambridge, MA: The MIT Press, pp. 67–92.

Cosmides, Leda (Apr. 1989). "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task". In: *Cognition* 31.3, pp. 187–276.

Crawford, Jarret T. and John Ruscio (Apr. 2021). "Asking People to Explain Complex Policies Does Not Increase Political Moderation: Three Preregistered Failures to Closely Replicate Fernbach, Rogers, Fox, and Sloman's (2013) Findings". In: *Psychological Science* 32.4, pp. 611–621.

Crawford, Neta C. (Mar. 2009). "Homo Politicus and Argument (Nearly) All the Way Down: Persuasion in Politics". In: *Perspectives on Politics* 7.1, pp. 103–124.

Dahl, Robert A. (1998). *On Democracy*. 2nd ed. New Haven, CT: Yale University Press.

De Neys, Wim, Sandrine Rossi, and Olivier Houdé (Apr. 2013). "Bats, Balls, and Substitution Sensitivity: Cognitive Misers Are No Happy Fools". In: *Psychonomic Bulletin & Review* 20.2, pp. 269–273.

Devlin, Jacob et al. (May 2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". Pre-Print. Google AI Language. arXiv: `1810.04805`.

Ditto, Peter H. and David F. Lopez (Oct. 1992). "Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions". In: *Journal of Personality and Social Psychology* 63.4, pp. 568–584.

Donald, Merlin (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Dorison, Charles A., Julia A. Minson, and Todd Rogers (July 2019). "Selective Exposure Partly Relies on Faulty Affective Forecasts". In: *Cognition* 188, pp. 98–107.

Druckman, James N. (2022). "A Framework for the Study of Persuasion". In: *Annual Review of Political Science* 25.

Druckman, James N. and Kjersten R. Nelson (Oct. 2003). "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence". In: *American Journal of Political Science* 37.3, pp. 729–745.

Eagly, Alice H. and Shelly Chaiken (1993). *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Erduran, Sibel and María Pilar Jiménez-Aleixandre, eds. (2007). *Argumentation in Science Education: Perspectives from Classroom-Based Research*. Science & Technology Education Library 35. New York: Springer.

Erisen, Cengiz, Milton Lodge, and Charles S. Taber (Apr. 2014). "Affective Contagion in Effortful Political Thinking". In: *Political Psychology* 35.2, pp. 187–206.

Esterling, Kevin M., Archon Fung, and Taeku Lee (Apr. 2021). "When Deliberation Produces Persuasion Rather than Polarization: Measuring and Modeling Small Group Dynamics in a Field Experiment". In: *British Journal of Political Science* 51.2, pp. 666–684.

Evans, Jonathan St. B. T. and Keith E. Stanovich (May 2013). "Dual-Process Theories of Higher Cognition: Advancing the Debate". In: *Perspectives on Psychological Science* 8.3, pp. 223–241.

Evans, Jonathan St. B.T. (Oct. 2003). "In Two Minds: Dual-Process Accounts of Reasoning". In: *Trends in Cognitive Sciences* 7.10, pp. 454–459.

Farrell, Henry, Hugo Mercier, and Melissa Schwartzberg (May 2023). "Analytical Democratic Theory: A Microfoundational Approach". In: *American Political Science Review* 117.2, pp. 767–772.

Farzam, Amirhossein et al. (Aug. 2024). "Multi-Task Learning Improves Performance in Deep Argument Mining Models". In: *Proceedings of the 11th Workshop on Argument Mining*. Bangkok: Association for Computational Linguistics, pp. 46–58.

Fernbach, Philip M. et al. (June 2013). "Political Extremism Is Supported by an Illusion of Understanding". In: *Psychological Science* 24.6, pp. 939–946.

Firth, David (Mar. 1993). "Bias Reduction of Maximum Likelihood Estimates". In: *Biometrika* 80.1, pp. 27–38.

Fishkin, James S. (2009). *When the People Speak: Deliberative Democracy & Public Consultation*. New York: Oxford University Press.

Fong, Christian and Justin Grimmer (Apr. 2023). "Causal Inference with Latent Treatments". In: *American Journal of Political Science* 67.2, pp. 374–389.

Gaertner, Samuel L. and John P. McLaughlin (Mar. 1983). "Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics". In: *Social Psychology Quarterly* 46.1, pp. 23–30.

Gaines, Brian J. et al. (Nov. 2007). "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq". In: *The Journal of Politics* 69.4, pp. 957–974.

Gerber, Marlène et al. (Sept. 2014). "Deliberative and Non-Deliberative Persuasion: Mechanisms of Opinion Formation in EuroPolis". In: *European Union Politics* 15.3, pp. 410–429.

Gibson, James L. (July 1998). "A Sober Second Thought: An Experiment in Persuading Russians to Tolerate". In: *American Journal of Political Science* 42.3, p. 819. JSTOR: 2991731.

Gretz, Shai et al. (Apr. 2020). "A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.5, pp. 7805–7813.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2021). "Machine Learning for Social Science: An Agnostic Approach". In: *Annual Review of Political Science* 24, pp. 395–419.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, pp. 267–297.

Guess, Andrew and Alexander Coppock (Oct. 2020). "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments". In: *British Journal of Political Science* 50.4, pp. 1497–1515.

Gutmann, Amy and Dennis Thompson (1996). *Democracy and Disagreement*. Cambridge, MA: Belknap Press.

Habernal, Ivan and Iryna Gurevych (Apr. 2017). "Argumentation Mining in User-Generated Web Discourse". In: *Computational Linguistics* 43.1, pp. 125–179.

Hahn, Ulrike and Mike Oaksford (July 2007). “The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies”. In: *Psychological Review* 114.3, pp. 704–732.

Haselton, Martie G., Daniel Nettle, and Paul W. Andrews (2005). “The Evolution of Cognitive Bias”. In: *The Handbook of Evolutionary Psychology*. Ed. by David M. Buss. 1st ed. Hoboken, NJ: John Wiley & Sons, Inc., pp. 724–746.

Heinze, Georg and Michael Schemper (Aug. 2002). “A Solution to the Problem of Separation in Logistic Regression”. In: *Statistics in Medicine* 21.16, pp. 2409–2419.

Hopkins, Daniel J. (Mar. 2014). “The Consequences of Broader Media Choice: Evidence from the Expansion of Fox News”. In: *Quarterly Journal of Political Science* 9.1, pp. 115–135.

Itzchakov, Guy, Avraham N. Kluger, and Dotan R. Castro (Jan. 2017). “I Am Aware of My Inconsistencies but Can Tolerate Them: The Effect of High Quality Listening on Speakers’ Attitude Ambivalence”. In: *Personality and Social Psychology Bulletin* 43.1, pp. 105–120.

Iyengar, Shanto and Kyu S. Hahn (Mar. 2009). “Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use”. In: *Journal of Communication* 59.1, pp. 19–39.

Jackman, Simon and Paul M. Sniderman (May 2006). “The Limits of Deliberative Discussion: A Model of Everyday Political Arguments”. In: *The Journal of Politics* 68.2, pp. 272–283.

Jost, John T. et al. (May 2003). “Political Conservatism as Motivated Social Cognition”. In: *Psychological Bulletin* 129.3, pp. 339–375.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, Daniel, Paul Slovic, and Amos Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Kalla, Joshua L. and David E. Broockman (June 2020). “Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments”. In: *American Political Science Review* 114.2, pp. 410–425.

Kelly, Thomas (Win. 2016). “Evidence”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Stanford, CA.

Kinder, Donald R. and Nathan P. Kalmoe (2017). *Neither Liberal nor Conservative: Ideological Innocence in the American Public*. Chicago: The University of Chicago Press.

King, Gary and Langche Zeng (2001). "Logistic Regression in Rare Events Data". In: *Political Analysis* 9.2, pp. 137–163.

Knox, Dean, Christopher Lucas, and Wendy K. Tam Cho (2022). "Testing Causal Theories with Learned Proxies". In: *Annual Review of Political Science* 25, pp. 419–441.

Koenig, Melissa A. (May 2012). "Beyond Semantic Accuracy: Preschoolers Evaluate a Speaker's Reasons". In: *Child Development* 83.3, pp. 1051–1063.

Kuhn, Deanna (Sum. 1992). "Thinking as Argument". In: *Harvard Educational Review* 62.2, pp. 155–178.

Kunda, Ziva (Nov. 1990). "The Case for Motivated Reasoning". In: *Psychological Bulletin* 108.3, pp. 480–498.

Lebo, Matthew J. and Daniel Cassino (Dec. 2007). "The Aggregated Consequences of Motivated Reasoning and the Dynamics of Partisan Presidential Approval". In: *Political Psychology* 28.6, pp. 719–746.

Levendusky, Matthew S., James N. Druckman, and Audrey McLain (Apr. 2016). "How Group Discussions Create Strong Attitudes and Strong Partisans". In: *Research & Politics* 3.2, pp. 1–6.

Lodge, Milton and Charles Taber (2000). "Three Steps toward a Theory of Motivated Political Reasoning". In: *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*. Ed. by Arthur Lupia, Mathew D. McCubbins, and Samuel L. Popkin. New York: Cambridge University Press, pp. 183–213.

Lukin, Stephanie M. et al. (Aug. 2017). "Argument Strength Is in the Eye of the Beholder: Audience Effects in Persuasion". Pre-Print. University of California, Santa Cruz. arXiv: 1708.09085.

Lupia, Arthur (Sum. 2002). "Deliberation Disconnected: What It Takes to Improve Civic Competence". In: *Law and Contemporary Problems* 65.3, p. 133. JSTOR: 1192406.

Lupia, Arthur and Mathew D. McCubbins (1998). *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* New York: Cambridge University Press.

Luskin, Robert C., James S. Fishkin, and Roger Jowell (July 2002). “Considered Opinions: Deliberative Polling in Britain”. In: *British Journal of Political Science* 32.3, pp. 455–487.

Luskin, Robert C. et al. (Mar. 2014). “Deliberating across Deep Divides”. In: *Political Studies* 62.1, pp. 116–135.

Madison, James (Nov. 1787). “Federalist No. 10: The Same Subject Continued: The Union as a Safeguard Against Domestic Faction and Insurrection”. In: *The New York Packet*.

— (Feb. 1788). “Federalist No. 49: Method of Guarding Against the Encroachments of Any One Department of Government by Appealing to the People Through a Convention”. In: *The New York Packet*.

Matias, J. Nathan (May 2019). “Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions”. In: *Proceedings of the National Academy of Sciences* 116.20, pp. 9785–9789.

McCain, Kevin, Scott Stapleford, and Matthias Steup, eds. (2022). *Epistemic Dilemmas: New Arguments, New Angles*. New York: Routledge.

McClain, Colleen et al. (Nov. 2021). *The Behaviors and Attitudes of U.S. Adults on Twitter*. Washington, DC: Pew Research Center.

Mellers, Barbara et al. (Mar. 2015). “The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics”. In: *Journal of Experimental Psychology: Applied* 21.1, pp. 1–14.

Mendelberg, Tali and John Oleske (Apr. 2000). “Race and Public Deliberation”. In: *Political Communication* 17.2, pp. 169–191.

Mercier, Hugo and Hélène Landemore (Apr. 2012). “Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation”. In: *Political Psychology* 33.2, pp. 243–258.

Mercier, Hugo and Dan Sperber (Apr. 2011). “Why Do Humans Reason? Arguments for an Argumentative Theory”. In: *Behavioral and Brain Sciences* 34.2, pp. 57–74.

Mercier, Hugo and Dan Sperber (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Munro, Geoffrey D., Carrie Weih, and Jeffrey Tsai (May 2010). "Motivated Suspicion: Asymmetrical Attributions of the Behavior of Political Ingroup and Outgroup Members". In: *Basic and Applied Social Psychology* 32.2, pp. 173–184.

Myers, David G. and George D. Bishop (Aug. 1970). "Discussion Effects on Racial Attitudes". In: *Science* 169.3947, pp. 778–779.

Parsons, Bryan M. (June 2010). "Social Networks and the Affective Impact of Political Disagreement". In: *Political Behavior* 32.2, pp. 181–204.

Payne, John W., James R. Bettman, and Eric J. Johnson (July 1988). "Adaptive Strategy Selection in Decision-Making". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.3, pp. 534–552.

Pew Research Center (Oct. 25, 2016). *The Political Environment on Social Media*. URL: https://www.pewresearch.org/internet/2016/10/25/the-political-environment-on-social-media/.

— (Apr. 7, 2021). *Social Media Fact Sheet*. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=3814afe3-3f3c-4623-910b-8a6a37885ab8.

Porter, Ethan and Thomas J. Wood (Sept. 2021). "The Global Effectiveness of Fact-Checking: Evidence from Simultaneous Experiments in Argentina, Nigeria, South Africa, and the United Kingdom". In: *Proceedings of the National Academy of Sciences* 118.37.

Prior, Markus (May 2013). "Media and Political Polarization". In: *Annual Review of Political Science* 16.1, pp. 101–127.

Puhr, Rainer et al. (2017). "Firth's Logistic Regression with Rare Events: Accurate Effect Estimates and Predictions?" In: *Statistics in Medicine* 36, pp. 2302–2317.

Redlawsk, David P. (Nov. 2002). "Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making". In: *The Journal of Politics* 64.4, pp. 1021–1044.

Sandoval, William A. (Jan. 2003). "Conceptual and Epistemic Aspects of Students' Scientific Explanations". In: *Journal of the Learning Sciences* 12.1, pp. 5–51.

Schneiderhan, Erik and Shamus Khan (Mar. 2008). "Reasons and Inclusion: The Foundation of Deliberation". In: *Sociological Theory* 26.1, pp. 1–24.

Settle, Jaime E. (2018). *Frenemies: How Social Media Polarizes America*. New York: Cambridge University Press.

Sniderman, Paul M., Richard A. Brody, and Philip E. Tetlock (1991). *Reasoning and Choice: Explorations in Political Psychology*. New York: Cambridge University Press.

Sniderman, Paul M. and Thomas Piazza (1993). *The Scar of Race*. Cambridge, MA: Belknap Press.

Somin, Ilya (May 2010). "Deliberative Democracy and Political Ignorance". In: *Critical Review* 22.2, pp. 253–279.

Stanley, Matthew L. et al. (Nov. 2020a). "Exposure to Opposing Reasons Reduces Negative Impressions of Ideological Opponents". In: *Journal of Experimental Social Psychology* 91.

Stanley, Matthew L. et al. (Sept. 2020b). "Resistance to Position Change, Motivated Reasoning, and Polarization". In: *Political Behavior* 42.3, pp. 891–913.

Stigler, George J. (June 1961). "The Economics of Information". In: *Journal of Political Economy* 69.3, pp. 213–225.

Sunstein, Cass R. (June 2002). "The Law of Group Polarization". In: *The Journal of Political Philosophy* 10.2, pp. 175–195.

Taber, Charles S., Damon Cann, and Simona Kucsova (June 2009). "The Motivated Processing of Political Arguments". In: *Political Behavior* 31.2, pp. 137–155.

Taber, Charles S. and Milton Lodge (July 2006). "Motivated Skepticism in the Evaluation of Political Beliefs". In: *American Journal of Political Science* 50.3, pp. 755–769.

— (Feb. 2016). "The Illusion of Choice in Democratic Politics: The Unconscious Impact of Motivated Political Reasoning". In: *Political Psychology* 37 (Suppl. 1), pp. 61–85.

Tesler, Michael (Oct. 2015). “Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed”. In: *American Journal of Political Science* 59.4, pp. 806–824.

Tetlock, Philip E. and Dan Gardner (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown Publishers.

Trouche, Emmanuel, Emmanuel Sander, and Hugo Mercier (Oct. 2014). “Arguments, More than Confidence, Explain the Good Performance of Reasoning Groups”. In: *Journal of Experimental Psychology: General* 143.5, pp. 1958–1971.

Trouche, Emmanuel et al. (Nov. 2016). “The Selective Laziness of Reasoning”. In: *Cognitive Science* 40.8, pp. 2122–2136.

VanderWeele, Tyler J. and Miguel A. Hernán (June 2012). “Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs”. In: *American Journal of Epidemiology* 175.12, pp. 1303–1310.

Walker, Marilyn A. et al. (May 2012). “A Corpus for Research on Deliberation and Debate”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Istanbul, pp. 812–817.

Westwood, Sean J. (Oct. 2015). “The Role of Persuasion in Deliberative Opinion Change”. In: *Political Communication* 32.4, pp. 509–528.

Wilkerson, John and Andreu Casas (2017). “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges”. In: *Annual Review of Political Science* 20, pp. 529–544.

Wood, Thomas and Ethan Porter (Mar. 2019). “The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence”. In: *Political Behavior* 41.1, pp. 135–163.

Wooldridge, Jeffrey M. (2015). *Introductory Econometrics: A Modern Approach*. Mason, OH: Cengage Learning.

Zaller, John and Stanley Feldman (Aug. 1992). “A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences”. In: *American Journal of Political Science* 36.3, p. 579.

Zaller, John R. (1992). *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zhu, Yukun et al. (Dec. 2015). “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books”. In: *2015 IEEE International Conference on Computer Vision*. Santiago, Chile: Institute of Electrical and Electronics Engineers, pp. 19–27.