Harnessing Partisan Motives to Combat Misinformation

Job Market Paper: Cameron Martel
Marketing Candidate, MIT Sloan School of Management

Cameron Martel[1*†], Jennifer Allen[1†], Gordon Pennycook[2], David G. Rand[1,3,4]

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Psychology, Cornell University, Ithaca, NY, USA
[3]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[*]Corresponding author: cmartel@mit.edu
[†]Shared first-authorship

Abstract

Partisan motives are often conceptualized as fundamentally in opposition to accuracy-directed motives. Rather than being opposed, however, it may be that partisan and accuracy motivations simply operate independently – in which case political motives may not necessarily interfere with truth discernment. Here, we test this hypothesis in the context of crowd evaluations of (mis)information. We predict that in the presence of accuracy motivations, stronger partisan motivations can actually lead to better outcomes – an increased quantity of flags, coupled with as good or better *truth discernment* – by motivating people to preferentially flag news that is both false and politically discordant. To empirically assess this prediction, we carried out a survey study and analyzed field data from X's (Twitter's) crowdsourced fact-checking platform Community Notes. These data show that more politically motivated individuals are more active community fact-checking participants, helping sustain overall contribution levels. Furthermore, our results show that more politically motivated participants engage in more politically biased flagging yet exhibit the same or better flagging discernment as compared to less politically motivated participants. Together, our results challenge the notion that partisan motives inherently undermine the ability and willingness to evaluate truth. Rather, political motivation may actually be the key to the provisioning of high quantity and quality crowdsourced fact-checks.

Keywords: misinformation, crowdsourcing, fact-checking, social media, political motivation, partisanship

Harnessing Partisan Motives to Combat Misinformation

Many scholars conceptualize directional partisan motives as fundamentally in tension with accuracy- or truth-seeking motives (Druckman, 2012; Druckman & McGrath, 2019; Kahan et al., 2017; Kunda, 1990; Taber & Lodge, 2006). This lens is commonly applied to the topic of misinformation, where it is argued that increasing polarization has created a "post-truth" world where people actively disregard accuracy in pursuit of partisan goals (Marie & Petersen, 2023; Osmundsen et al., 2021; Peterson & Iyengar, 2021; Rathje et al., 2023; Van Bavel & Pereira, 2018).

Concerns about partisan motives are particularly salient when considering "wisdom of crowds" approaches for identifying misinformation – a burgeoning strategy for keeping up with the rapid pace of false news online. Although crowds of laypeople have been shown to effectively identify misinformation at scale with high accuracy (Allen et al., 2021a; Arechar et al., 2023; Godel et al., 2021; Martel, Allen, et al., 2024; Resnick et al., 2023), partisan motives may undermine such efforts – particularly in a voluntary system in which users can choose what to fact-check. Directionally motivated partisans may exploit misinformation-reporting mechanisms to target content that is simply counter-partisan, rather than inaccurate. Recent evidence indeed shows that participants in community fact-checking systems are more likely to flag counter-partisan posts and evaluate co-partisan fact-checks as high quality (Allen et al., 2022; Yasseri & Menczer, 2023). Other research has posited that these political biases might increase flagging of reliable and politically neutral sources (Coscia & Rossi, 2020) and could decrease reporting of false news that users agree with politically.

This is not the only perspective, however, on the relationship between partisan motives and accuracy motives. Rather than being diametrically opposed, it may be that these motives simply operate independently. This alternative perspective resonates with recent findings that engaging in reasoning neither amplifies nor reduces partisan bias – but instead increases truth discernment for both politically concordant and discordant information (Bago et al., 2020; Pennycook & Rand, 2021). If partisan motives and accuracy motives are independent, then – rather than partisan motives necessarily reducing truth discernment – these two motivations may interact in more complicated ways (Duncan, 2022; Shi et al., 2019). Here we ask whether partisan motives undermine or work in tandem with accuracy motives in the domain of crowdsourced identification of misinformation.

Theoretical Background

Crowdsourced Fact-checking as a Scalable Misinformation Intervention

Online misinformation poses myriad threats to various affected parties. For individuals, encountering false and misleading claims about important societal issues can have harmful consequences. For example, increased exposure to misinformation websites before and after the 2020 U.S. Presidential Election is associated with greater belief in false claims about election fraud (Dahlke & Hancock, 2022), and false and misleading content about the COVID-19 vaccine has been demonstrated to substantially increase vaccine hesitancy (Allen et al., 2024; Loomba et al., 2021). False claims can also negatively affect individuals as consumers. For instance, misinformation can reduce consumers' willingness-to-pay for online products (Fong et al., 2023). And relatedly, fake news advertising – ads that deceptively imitate legitimate news articles – have been shown to increase traffic and interest in falsely advertised products (Rao, 2022).

Misinformation can also have detrimental effects on firms and advertisers. For example, recent work finds that companies that advertise on misinformation websites can face strong backlash from consumers, even if such advertising is inadvertent (Ahmad et al., 2024). The spread of online falsehoods is also damaging to social media companies, who have been perceived as allowing falsehoods to proliferate on their platforms (Donovan, 2020; Guess et al., 2018). Such criticisms in turn risk the loss of users and advertisers (Nakano/Bloomberg, 2023), and also increase public and regulatory pressure on social media companies.

In turn, researchers and practitioners have developed a wide variety of interventions designed to target misinformation and harmfully misleading online content (Kozyreva et al., 2024). At present, the most common approach of major tech companies for handling on-platform misinformation involves partnership with professional third-party fact-checking organizations. This approach generally involves platforms sending potentially false content to professional fact-checkers for their evaluation. Posts determined to be false may then be labeled, downranked, or removed depending on each platform's specific policies. Facebook and Instagram partner with professional fact-checkers to help label and demote false posts (Instagram, n.d.; Meta, n.d.-a). TikTok similarly partners with fact-checkers, and purports to remove some instances of false content (TikTok, 2023). Promisingly, interventions relying on fact-checker evaluations such as expert warning labels (Martel & Rand, 2023), downranking identified false content (Jia et al., 2024), and deplatforming frequent sharers of identified misinformation (McCabe et al., 2024) have been shown to have beneficial effects on reducing belief in, and spread of, misinformation and other types of harmful content. However, relying solely on fact-checkers for content moderation of misinformation has important limitation. Simply put, there are not enough fact-checkers to keep up with the amount of potentially false content produced online – and fact-checking institutional growth has even slowed in recent years (Luther, 2023). This substantially limits the ability of fact-checkers and platforms to keep up with the amount of misinformation produced on-platform. Indeed, most claims are never fact-checked, and even for those which are fact-checked, the process is often too slow to stop misinformation's viral spread (Allen et al., 2024; Meta, n.d.-b). In addition to this scalability problem, the current professional fact-checking approach also has limitations as to what content is acted upon. For example, while labeling and demotion decreased the effect of blatantly false content about vaccines on Facebook, vaccine skeptical content that was misleading – yet not outright false and thus not labeled such by fact-checkers – was left unacted upon and greatly contributed to increased vaccine hesitancy among Facebook users (Allen et al., 2024). Furthermore, U.S.-based social media platforms hire more English-language based experts, resulting in inequitable policy enforcement of professional evaluations for non-English misinformation (Avaaz, n.d.).

One proposed solution to these limitations of expert dependent approaches is to enable social media users *themselves* to identify misleading content they encounter online. This approach of crowdsourced fact-checking allows for lay users to respond to false content potentially faster than external professionals (Zhao & Naaman, 2023) and also may enable users to flag and identify content that is not only outright false, but also that which needs additional context or requires different forms of moderation. Encouragingly, recent work shows strong empirical evidence that crowds can accurately identify misinformation when presented with series of posts or articles to evaluate (Martel, Allen, et al., 2024), in line with years of research on the "wisdom of crowds" from aggregate layperson judgments (Galton, 1949; Surowiecki, 2005). And currently, Facebook deploys a similar approach of community content review, whereby non-experts are recruited and paid to fact-check-content as a way of augmenting professional fact-checkers and speeding up the

review process (Silverman, n.d.). However, despite its empirical promise, this paid non-expert reviewer approach still does not fully address the initial scalability issue. Social media platforms would still need to identify potentially misleading content, recruit and pay crowd workers, and wait for these paid workers to evaluate the content they were sent – much like the existing expert evaluation approach.

As a result, several other tech platforms have initiated volunteer community fact-checking systems. Most notably, X (formerly Twitter) operates Community Notes – a feature in which users can flag misleading content, write free-response notes to add additional context on posts, and where other users can then vote on whether these free-response notes are helpful and should be appended on posts for all users to see (Allen et al., 2022; Wojcik et al., 2022). This type of crowdsourced fact-checking system – whereby selection of whether and what to evaluate is endogenous to reviewers – better addresses professional fact-checking's scalability problem by allowing all users to potentially contribute to identifying misleading content. However, endogenous selection of what to evaluate also raises a critical, and hereto untested, concern – users may choose to flag content that they disagree with politically, rather than content that is actually false or misleading. Highly partisan individuals may hijack such a system to advance their political goals rather than help improve online information quality, thus undermining the utility of a voluntary community fact-checking initiative.

Partisan Directed Motives versus Accuracy Directed Motives

Concern about extreme political motives undermining crowdsourced fact-checking systems is in line with an abundance of research positing that identity-based partisan motivation is a key driver of online misinformation sharing to begin with. For instance, Osmundsen and colleagues argue that partisan polarization is the primary motivation driving political false news sharing on Twitter (Osmundsen et al., 2021). By this account, partisan motives may similarly drive incorrect flagging of content that is true but politically disagreeable. Further exacerbating this worry, scholars have also framed partisan motivation as directly in conflict with accuracy-directed goals. In their "identity-based model of belief," for example, Van Bavel and colleagues model information belief through the process by which "accuracy goals compete with social identity goals to determine the value of belief" (p. 2, (Van Bavel et al., 2024)). And empirically, recent work also shows that financially incentivizing the accuracy of headline veracity judgments can increase correct evaluations while decreasing partisan bias in judgments, again suggesting that accuracy and partisan goals may be in direct tension (Rathje et al., 2023).

However, a growing body of research suggests that partisan motives may not be a fundamental driver of misinformation sharing. In contrast with a directional motivated reasoning account, this work has found that increased reasoning does not increase nor decrease partisan bias – but rather, increased reasoning and relevant knowledge increases accuracy judgments of true and false content (Bago et al., 2020; Pennycook & Rand, 2019, 2021). These findings suggest that partisan and accuracy goals may be independent, or at least not necessarily in fundamental conflict with one another. Rather, such motivations could jointly play a role in individuals' decisions about sharing - or flagging - potentially false online content.

Harnessing Motivations to Promote Community Fact-checking

If people are motivated by partisanship instead of accuracy, then extreme partisans in a voluntary crowdsourced fact-checking system would be predicted to flag any content they disagreed with politically – regardless of its truth. However, if partisanship and accuracy motivations operate separately, then more extreme partisans may be particularly motivated to flag content that is not only disagreeable, but also false. Furthermore, the contributions of such highly motivated individuals could also alleviate a second concern pertaining to voluntary crowdsourcing systems – low participation rates. Contributing reviews of online content – whether it be for information veracity, product quality, or commercial transaction satisfaction – has been conceptualized as a type of public good (Resnick & Zeckhauser, 2002). Flagging a post or writing a fact-check can provide helpful benefits to others but takes time and effort to do. Given these costs, a volunteer fact-checking system runs the risk of not enough users flagging any content at all. Similar online volunteer systems face similar challenges – for instance, only a small proportion of customers write reviews of products purchased online (E. T. Anderson & Simester, 2014) and far more people report relying on online reviews than writing them (A. S. and M. Anderson, 2016). Relatedly, online knowledge exchange sites (i.e., question-and-answer forums) have been documented to suffer from an undersupply of user generated contributions (Goes et al., 2016). Thus, if both partisan and accuracy goals can motivate flagging, then highly motivated individuals (including potentially extreme partisans) may actually be necessary community fact-checkers for sustaining a high enough number of fact-checks to help inform other social media users about misleading content online.

In our current work, we empirically assess whether individuals with strong political motivations help or hinder community fact-checking efforts. Across a survey-based misinformation flagging study (Study 1) and field data from X's (Twitter's) Community Notes crowdsourced fact-checking platform (Study 2), we investigate the extent to which political motivation predicts flagging bias, quantity, and quality. Our results advance theoretical understanding of the interplay between partisan and accuracy motivations, as we shed light on whether these motivations are inherently opposed or rather may operate separately and perhaps jointly work to enable community fact-checking. Practically, our findings also inform practitioners and policymakers as to how to design and implement crowdsourced misinformation identification mechanisms – and the role that individuals with extreme political motivations may play within these systems.

Study 1: Misinformation Flagging in an Online Survey Environment

We predict that accuracy motives and partisan motives separably motivate flagging of online content. Specifically, we posit that accuracy motives should drive people to flag inaccurate content (and not flag accurate content) regardless of political alignment, whereas partisan motives should drive people to flag counter-partisan content (and not flag co-partisan content) regardless of accuracy. This theoretical framework predicts that content that is both false *and* counter-partisan will be flagged the most, as both accuracy and partisan motivation predicts flagging this category of online post. This framework also suggests that under conditions of low accuracy motivation, partisan motivation may increase overall flagging quantity and quality (for a simple formal model of misinformation flagging, see SI Section 1).

We first test these predictions with a preregistered (https://aspredicted.org/NDY_YWL) survey study in which participants were shown a newsfeed of social media posts about a variety of topics. Participants were asked to help social media companies identify misinformation by

flagging the posts they believed to be false – if they believed a post to be false, they were instructed to flag it and write an explanation as to why. We then examine what types of posts participants were most likely to flag and whether participant's political motivation predicted the quantity of quality of flags they produced.

Methods

In our survey study, participants first completed several assessments of political motivation – partisanship, political knowledge, issue polarization, and out-party animosity. Participants were next presented with a pretested set of false, true, and non-political entertainment headlines (in 'Facebook format'; for details, see SI Section 2a) and were asked to identify which headlines were potentially false or misleading and why. Participants could 'flag' and write a free response explanation for why they flagged headlines as misleading; or they could scroll past headlines they did not believe were false or misleading. We predicted that false, politically discordant headlines would be more likely to be flagged as misleading than any other type of headline. We also predicted that more politically engaged participants would be more likely to (i) write a greater number of flags, (ii) flag a greater fraction of discordant content, (iii) flag a greater fraction of false content, and (iv) flag more false, discordant headlines than any other type of headline.

Participants

We preregistered a target sample of 2,000 complete responses from American participants recruited from Lucid, quota-matched to the national distribution on age, gender, ethnicity, and geographic region. We preregistered exclusion of participants who incorrectly answered either of two trivial attention checks (e.g., captcha) at the beginning of the survey, as well as participants who straightline responded by flagging over 50 headlines in the main task. In total, 4,404 participants began the survey. Only 3,410 participants correctly answered both preregistered trivial attention checks at the beginning of the survey. Of these, 2,402 continued the survey until at least the beginning of the main headline flagging task. 22 participants were excluded for straightline responding as preregistered. Our full sample of participants who passed attention filters and began the main task consisted of N=2,380 participants (mean age=47.8, 1,312 female participants, 1,049 male participants, 19 participants selecting another gender option; 1,825 White-only, 293 Black-only, 84 Asian-only, 178 who selected another race option). This study was conducted on 11 November – 12 December 2022.

Materials

Attention Items. We first presented participants with two trivial attention checks – a captcha and a single-select item asking participants to select the option 'Somewhat disagree' from the choice set listed. Participants were also asked four instructional attention items throughout the survey (Berinsky et al., 2014), which we include as a covariate in our individual-level models. We also filter by attention in preregistered secondary analyses (see SI Section 2c).

Demographics. Participants were asked demographic questions about age, gender, race, ethnicity, education, income, and belief in God(s).

Political Ideology & Identity. We assessed political ideology and identity via questionnaires adapted from the American National Election Studies (ANES, 2020). Participants

reported their political ideology on a 7-point scale from 'Strong Liberal' to 'Strong Conservative' and reported their political identity on a 7-point scale from 'Strong Democrat to 'Strong Republican.' Secondarily, we also assessed social and economic ideology on 5-point scales from 'Strongly Liberal' to 'Strongly Conservative.'

Out-party Animosity & In-party Favorability. Participants completed a political feeling thermometer (Weisberg & Rusk, 1970) in which they assessed their feelings toward Republican Party and Democratic Party voters on a slider scale from 0 (Very cold) to 100 (Very warm). We define out-party animosity as participants' reverse-coded feeling thermometer score for counter-partisans, and in-party favorability as participants' feeling thermometer score for co-partisans.

Political Knowledge. We asked participants four factual questions about political and public policy (Tappin et al., 2021) – e.g., "Whose responsibility is it to decide if a law is constitutional or not?" For each question, participants were allotted 15 seconds and asked not to look up the answers.

Issue Polarization. We assessed participants' political issue positions across 11 different items (ANES, 2020; Berinsky et al., 2021). The topics of these issues were as follows: health insurance, government job assistance, government services provisioning, government assistance for Black Americans, military spending, abortion, assault rifles, gay marriage, undocumented immigration, the Affordable Care Act, and investment in environmental protection. Issue *polarization* was then calculated as the absolute value of participants' average issue stance across items away from the scale midpoint.

Misinformation Flagging Task. In the main task, participants were presented with 80 article headlines in a random order. 20 posts were entertainment non-news items, and the remaining 60 posts were balanced on both veracity and whether they were favorable to Democrats or Republicans. Thus, participants saw five different categories of headlines – entertainment, discordant false, concordant false, discordant true, and concordant true. All 80 headlines were then presented on a single, scrollable page of the survey. The false news headlines were originally selected from the third-party fact-checking website Snopes.com. The true news headlines were accurate and selected from mainstream news outlets (Pennycook, Binnendyk, et al., 2021a). Partisan favorability of headlines was determined via a pretest (SI Section 2a).

Our full survey materials are available online (https://osf.io/3ngbt/).

Procedure

Participants first completed attention screeners, demographics, questions on political ideology and identity, and their affect towards Democratic Party and Republican Party voters. In random order, participants next completed political knowledge and political issue position questionnaires. Throughout this section, participants also completed four instructional attention items.

At the beginning of the main headline flagging task, participants were explicitly told that the survey was being conducted by MIT researchers who advise social media companies like Facebook and Twitter on how to address misinformation. Participants were then instructed that in the current study, they would be asked to help identify which content from social media is potentially false or misleading, and why – and that these decisions and reasons would be used to better help advise social media companies. Next, participants were told they would see a single, scrollable page with many news article headlines from actual social media posts. For each item, participants were informed they could either select a red flag item to indicate that post is false or

misleading, or they could continue scrolling down to the next article if they saw a headline that they did not think was false or misleading. Upon flagging an article, participants were required to write a free response explanation as to why they flagged it as false or misleading. Participants were told to write this explanation such that it could be shown to other social media users. Participants then completed a practice flagging task and were given final instruction reminders.

Finally, participants advanced to a single, scrollable page containing all 80 headlines. Each headline was presented in 'Facebook format' with a clickable red icon underneath. Posts were each separated by a divider line to clarify which flag corresponded to which article. If participants selected to flag a headline, a required free response item would pop-out, asking them to write a short justification as to why they flagged the above post as misleading.

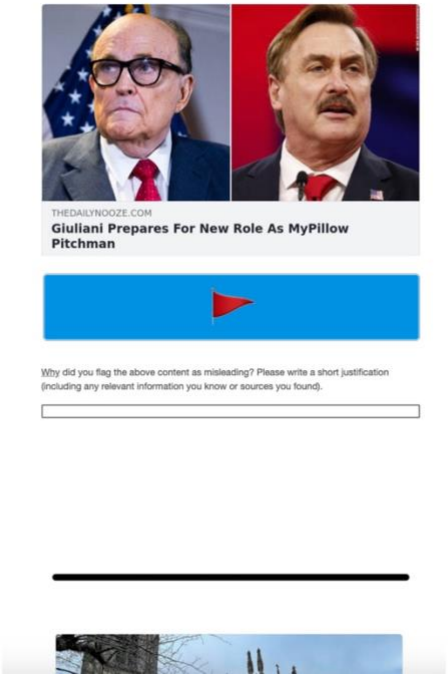

Figure 1. Survey headline flagging task. Individuals were asked to participate in a social media post flagging task to help social media companies identify misinformation. Participants saw 80 posts in total – 20 non-political (entertainment) posts, and 60 news posts balanced on veracity and partisan favorability. Participants were instructed to flag a post (i.e., click the red flag icon below the post, as shown above) if they believed the post to be false or misleading, and then write an explanation as to why. For posts participants did not believe to be false or misleading, they were instructed to keep scrolling on to further headlines.

Our study was preregistered (https://aspredicted.org/NDY_YWL) and approved as exempt by the MIT Committee on the Use of Humans as Experimental Subjects (Protocol E-4471).

Results

We first assess what type of headline participants were most likely to flag as false or misleading. We conduct a logit regression predicting the likelihood of flagging by item type category (holdout=Discordant False headline), with two-way robust standard errors clustered by participant and headline. We find that participants were more likely to flag false, politically discordant posts than any other type of headline, followed by posts that were false and politically

concordant (b=-0.234, SE=0.037, z(190,395)=-6.268, p<.001; see SI Table S2 for full logit regression table and robustness checks). Flagging likelihood was substantially lower for entertainment (b=-1.688, SE=0.125, z(190,395)=-13.553, p<.001), concordant true (b=-1.798, SE=0.118, z(190,395)=-15.182, p<.001), and, crucially, discordant true (b=-1.580, SE=0.140, z(190,395)=-11.325, p<.001) headlines relative to discordant false posts.

Next, we investigate the relationship between political motivation and flagging behavior. As pre-registered, we conducted a principal component analysis on our measures of out-party animosity, issue polarization, and political knowledge, and found that all three constructs had high weights in the first principal component (see SI Section 2b). We therefore consider this first principal component a continuous measure of aggregate political motivation across these dimensions.

We begin our analysis of political motivation by evaluating its relationship with overall flagging rates. Across all participants who started the main headline flagging task, 30.7% flagged no headlines (mean=6.54 flags, SD=8.81) – underscoring how even in a paid survey environment, motivating the provision of fact-checks is a substantial challenge. As hypothesized, we find that more politically motivated participants contributed a greater number of overall flags (ordinary least squares (OLS) regression predicting flag count by political motivation and standard controls: b=1.641, SE=0.166, z(2,351)=9.888, p<.001; see SI Table S3 for full table and robustness analyses). To illustrate this, a media split on political motivation shows that low politically motivated participants (N=1,189) wrote 5,305 flags, whereas highly politically motivated participants (N=1,191) contributed 10,259 flags. These findings are qualitatively similar when independently assessing our political motivation sub-measures – greater out-party animosity, issue polarization, and political knowledge are each associated with increased overall flagging (see SI Tables S4-S7). We also do not observe evidence of a partisan asymmetry between Democrats and Republicans in overall flagging rates (b=-0.198, SE=0.179, z(2,351)=-1.106, p=.269).

Are politically motivated participants simply flagging more counter-partisan posts, or are they providing high quality (i.e., discerning) flags? While more politically motivated participants flagged more discordant than concordant posts overall (OLS regression with analytic weighting by flag count: b=0.034, SE=0.003, z(1,628)=10.634, p<.001; SI Table S10) – that is, were more politically biased – they were also more discerning, flagging fewer true and entertainment posts relative to false posts (OLS regression with analytic weighting by flag count: b=-0.054, SE=0.005, z(1,628)=-10.673, p<.001; SI Table S15). These results are consistent when examining out-party animosity, issue polarization, and political knowledge in separate analyses (see SI Sections 2c(iii) and 2c(iv) for all regression tables and robustness checks).

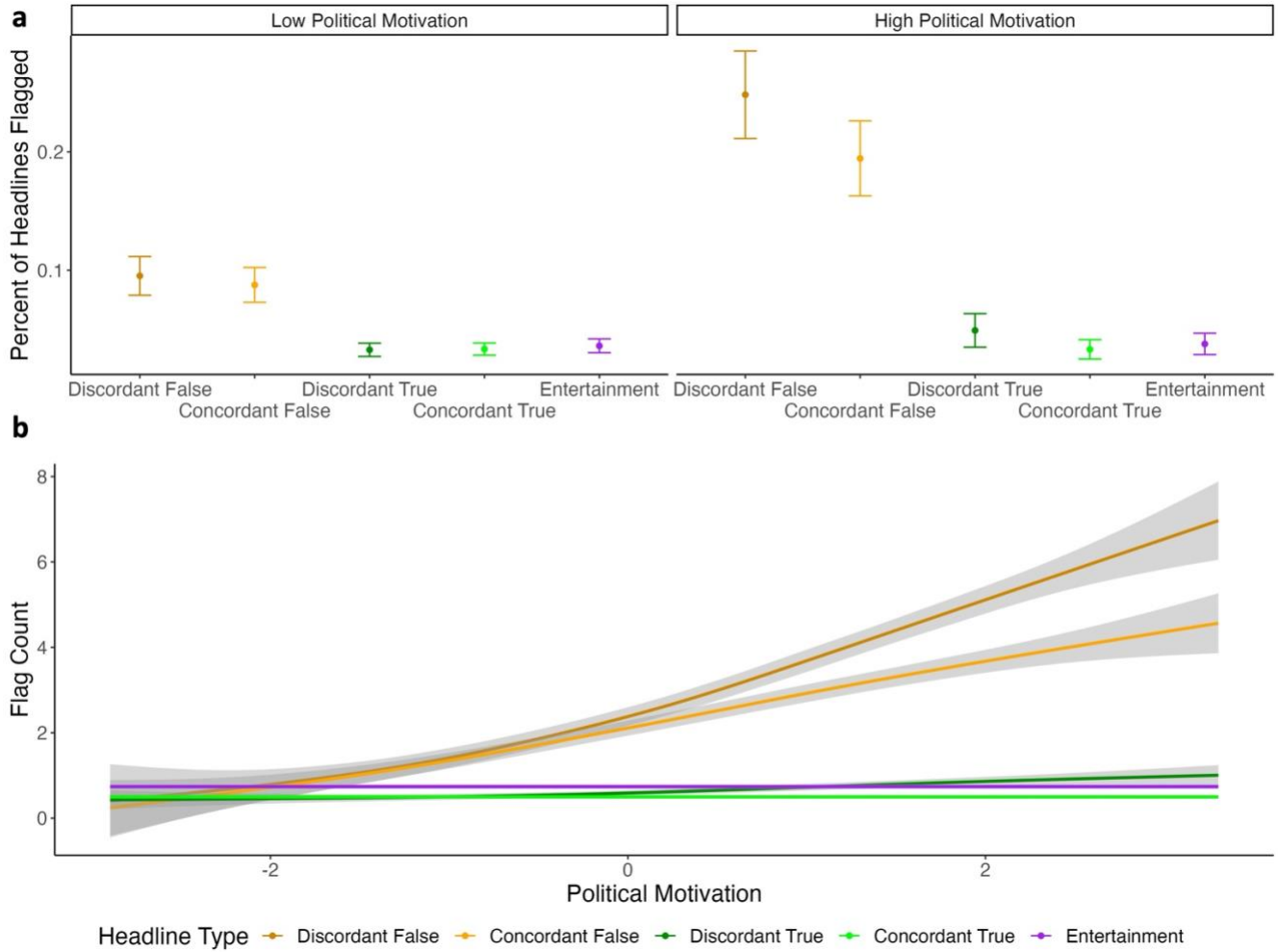

Figure 2. Politically motivated survey participants produce more flags with greater bias and discernment. More politically motivated participants (first principal component of out-party animosity, issue polarization, and political knowledge) write more flags; and specifically, are more likely to flag discordant, false headlines. (a) Shown are the proportion of headlines flagged by post type, median split by participant political motivation. Both groups flag a greater proportion of false than true headlines; crucially, highly politically motivated individuals flag a substantially greater proportion of false, discordant headlines – as well as a greater proportion of false, concordant headlines. This increase is also larger than the increase in flagging of true, discordant headlines by more politically motivated participants. Error bars reflect 95% confidence intervals. (b) Shown are the predicted number of flags by post type and political motivation. As we predicted, overall flag count increases with political motivation. Importantly, political motivation is most associated with an increase in false, discordant flagging. Excluding the most highly politically motivated participants would thus result in a decrease in both flagging quality and quantity. Error bars reflect 95% confidence intervals.

We further examine the specific types of posts that highly politically motivated participants were particularly likely to flag. We predict flag count by item type (holdout: False, Discordant), political motivation, and their interaction – and find that more politically motivated individuals flagged an even greater number of false, discordant headlines relative to all other headline types (*p*s<.001; see SI Table S20; Figure 2) including, critically, the number of flags written on true, discordant posts (*b*=-0.831, *SE*=0.066, *z*(11,755)=-12.612, *p*<.001). This result is again largely consistent when examining out-party animosity, issue polarization, and political knowledge in separate models (see SI Section 2c(v)).

We also observe partisan asymmetries in flagging bias and discernment. Republican-leaning participants write a lower proportion of discordant flags (*b*=-0.011, *SE*=0.004, *z*(1,628)=-2.661, *p*=.008); however, they also write a greater number of incorrect false-positive flags (*b*=0.024, *SE*=0.006, *z*(1,628)=4.070, *p*<.001), such that participants more affiliated with the Republican party flag more true and non-political entertainment headlines overall (see SI Figure S2).

Discussion

In sum, our survey findings indicate several key results. First, participants were much more likely to flag false headlines than true headlines. Second, participants were more likely to flag discordant than concordant headlines. And third, more politically motivated participants flagged headlines at about twice the rate of less politically motivated participants and exhibited greater political bias and veracity discernment in their flagging. Overall, less politically motivated users wrote 3,265 flags on false headlines and 2,040 flags on true and entertainment headlines, while more politically extreme users contributed 7,896 flags on false headlines and 2,363 flags on true headlines. Thus, noy only is the true positive rate greater for more politically motivated participants (77% versus 61.5%) but excluding flags from more politically motivated users would decrease the number of total flags by about two-thirds. Together, our survey study provides confirmatory evidence towards our key prediction that partisan motives are not necessarily harmful, but in fact can be beneficial for flagging propensity and discernment.

Study 2: Misinformation Flagging in the Field – X's (Twitter's) Community Notes

Our survey study results demonstrate that more politically motivated individuals not only provision more flags in a fact-checking task, but also maintain high discernment by flagging more false than true posts. These results suggest that both accuracy and partisan motives may be necessary to sustain the volume and quality necessary in a successful crowdsourced misinformation identification system. But will such results generalize beyond the context of a survey study to the field? Participants in our survey may have had outsized accuracy motivations. For example, they were instructed specifically to flag only false headlines, and to do so for the benefit of helping social media companies identify false and misleading posts on their platforms. Conversely, study participants may constitute a subpopulation with weaker political and partisan motives than individuals who would volunteer for participation on an actual fact-checking platform on social media, where political motives may be further heightened.

To evaluate whether out findings generalize to an actual online news evaluation system, we assess evidence from X's (formerly Twitter's) crowdsourced fact-checking program Community Notes (formerly named Birdwatch). X users admitted to Community Notes are able

to flag tweets as 'Misinformed or potentially misleading' and write a free response explanation – called a 'note' - as to why, in order to add further important context to misleading tweets (see Figure 3). Notes rated as helpful by enough other users would then be shown on X alongside the original tweet (the helpfulness threshold for making notes visible to everyone on tweets has been updated numerous times by Community Notes – X now uses a bridging-based algorithm to determine which notes reach consensus as broadly helpful) (Ovadya & Thorburn, 2022; Wojcik et al., 2022). Here, we assessed a dataset of flags from the first six months of the Community Notes program to investigate what types of tweets are flagged and the relationship between partisan motivation and flagging in an applied field setting.

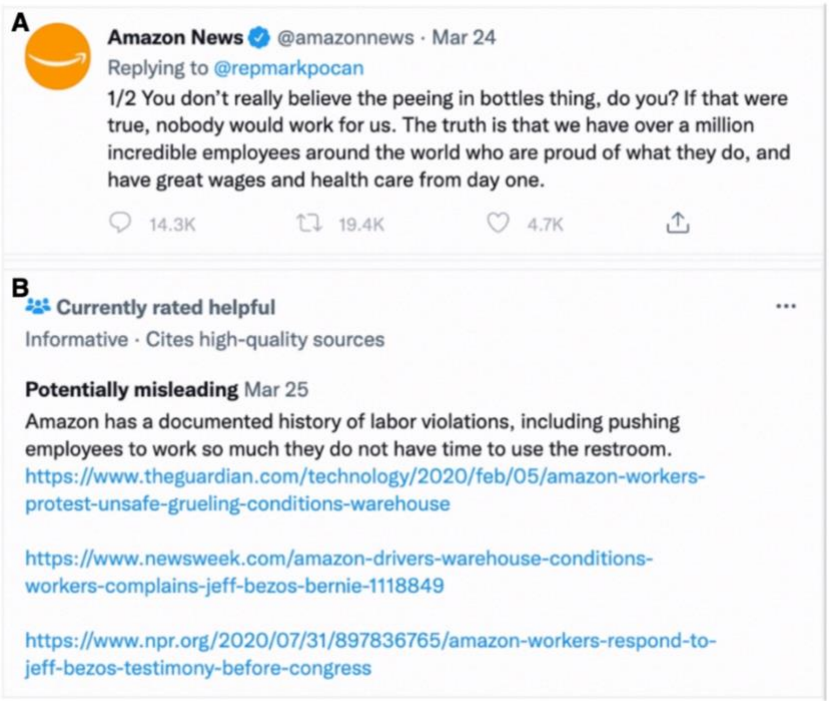

Figure 3. Example of Community Notes participation. Users in the Community Notes program have the ability flag misleading tweets and write fact-check 'notes' explaining why they flagged a tweet. (A) shows an original tweet, while (B) displays a user contributed flag and note classifying the original tweet as 'potentially misleading' and adding additional context as to why – shown as would be displayed had the note reached a threshold of helpfulness via reviews from other Community Notes users. Figure repurposed with permission (Allen et al., 2022).

Methods

Following our Study 1 survey results, we predicted that more politically extreme users contribute a greater number of overall flags, flag a greater proportion of politically discordant tweets, and flag a greater proportion of tweets actually containing false or misleading content. We also predicted in particular that more politically extreme users would flag more tweets that were both false and politically discordant.

Participants

As part of a data sharing agreement, X (Twitter at the time) provided us a dataset covering all Community Notes entries created from the program's beginning on 28 January 2021 through

29 June 2021. This dataset is similar to the publicly available Community Notes datasets (available here: https://x.com/i/communitynotes/download-data), except with additional information for internal research purposes that allowed us to link the users of Community Notes to their actual Twitter IDs. For the data analyzed here, Community Notes was in its pilot stage and participation was limited to a small subset of users who applied for participation and were accepted by Twitter. Twitter aimed to include users from a wide and balanced set of perspectives in the Community Notes pilot. In total, 4,025 unique Twitter users applied and were accepted into the Community Notes pilot. We used the M3Model (Wang et al., 2019) to estimate predicted gender and age categories from these Twitter users' available platform data. Based on these predictions, our full sample included 748 profiles classified as female, 2,934 profiles classified as not female, and 343 profiles the model could not classify. This model also predicted that our sample included 635 profiles of individuals not older than 18 years old, 1,119 individuals between 19 and 29 years old, 934 individuals between 30 and 39 years old, 994 individuals at least 40 years old, and 343 profiles the model could not classify.

Importantly, we also inferred users' partisanship from the accounts users followed, following established methodologies (Barberá et al., 2015). On a scale from -2.5 to 2.5, with more positive values indicating greater affinity towards the Republican party, the partisanship of our sample had mean -0.03 and a standard deviation of 1.39 (min=-2.42, max=2.42, median=-0.08).

Materials

Our field analyses used several datasets. First, we received from X (Twitter) an IDs dataset, containing the unique identifiers of the 4,025 users accepted to Community Notes' pilot program. Second, we received from X (Twitter) a Notes dataset, including information on all note entries available at the time. Data for each note most importantly includes a unique note ID, note writer ID, original tweet ID, original tweet writer ID, note classification ('Not misleading' or 'Misinformed or potentially misleading'), and the text of the free response note. We also used the (now discontinued) academic Twitter API to pull the full text of original tweets from the Notes dataset. Third, we further used the Twitter API to pull user characteristics of all Community Notes users, given their unique Twitter identifiers linking their Twitter behavior to their Community Notes contributions. We collected the following characteristics: inferred partisanship (Barberá et al., 2015), follower count, friends count (number of accounts the user follows), statuses count (total number of tweets and retweets posted by the user), inferred age and gender (Wang et al., 2019), toxicity score of past tweets and retweets (*Perspective API - How It Works*, n.d.), feed quality score as assessed via the aggregate domain quality ratings from each user's past tweets and retweets (Lin et al., 2022), and elite misinformation exposure score based on PolitiFact informed falsity scores of political elites users in our sample followed on Twitter (Mosleh & Rand, 2022).

We also sampled a subset of original tweets that Community Notes users had classified as misleading, up-sampling for tweets that received multiple notes (we first selected all tweets with at least three notes, then randomly sampled 300 additional tweets with at least one note). In total, we selected 461 original tweets (14.6% of total original tweets classified as misleading in our dataset). We then sent these tweets to two fact-checkers hired via Upwork for evaluation. For each tweet, both fact-checkers were asked: "Given current evidence, I believe this tweet is:" followed by choices "Not misleading" and "Misinformed or potentially misleading" (also see SI Section 3a). This is the same classification question asked of Community Notes participants. We also recruited workers from Amazon Mechanical Turk (N=355; mean age=39.9; 124 female

participants, 226 male participants, 5 participants selecting another gender identification or did not answer) to evaluate several features of these 461 original tweets. These features included whether tweets were more favorable to Democrats or Republicans, the category best describing the main topic of each tweet, and how controversial the main topic of each tweet was (see SI Section 3b).

Our full fact-checker and Mechanical Turk rating surveys are available online (https://osf.io/3ngbt/). Our Mechanical Turk rating study was approved as exempt by the MIT Committee on the Use of Humans as Experimental Subjects (Protocol E-4592). Analysis code for our field data investigation is available online. These analyses were not preregistered. Our Community Notes field data is not fully available online, given confidentiality concerns about Twitter user IDs and our data sharing agreement with X (Twitter).

Results

Overall, we examine a field dataset of 4,442 flags classifying 3,151 unique tweets as misinformed or potentially misleading. We first examine who participants in Community Notes flagging. Of 4,025 users who applied and were accepted as Community Notes users, only 1,046 contributed notes – such that 74% of users did not write a single note (mean=1.10 notes, *SD*=5.17). Given this low participation rate, we next consider the role of partisan motivation as a predictor of engaging in flagging. We assess political extremity as the absolute value of users' partisanship score, as inferred by the accounts users followed (Barberá et al., 2015). Convergent with our predictions and survey study results, we find that more politically extreme users flag more tweets (OLS model predicting flag count by political extremity: b=0.453, SE=0.129, z(3,393)=3.506, p=.001; see SI Table S25). To further illustrate this, a median split on political extremity for those with classifiable partisanship scores shows that low politically extreme users (N=1,703) flagged 1,610 tweets in total, whereas highly politically extreme users (N=1,703) flagged 2,549 tweets in total.

To evaluate the quality of these flags, we next examine the subset of 461 tweets for which we collected fact-checker veracity evaluations and crowdsourced partisan favorability ratings (see SI Section 3a,b). Strikingly, we find that 79.4% of notes agreed with the ratings of the professional fact-checkers – i.e., where both the community flagger and fact-checker classified the tweet as potentially misleading. This shows that Community Notes flagging discernment is overall quite high on average. And consistent with our predictions, we find that the substantial majority of these flags were on false discordant tweets (67.1%, omitting flags from participants for whom we could not infer partisanship; see Figure 4). A much smaller fraction of flagged tweets were on true discordant (16.2%) and false concordant (11.8%) tweets. As expected, very few flags were on true concordant tweets (4.8%).

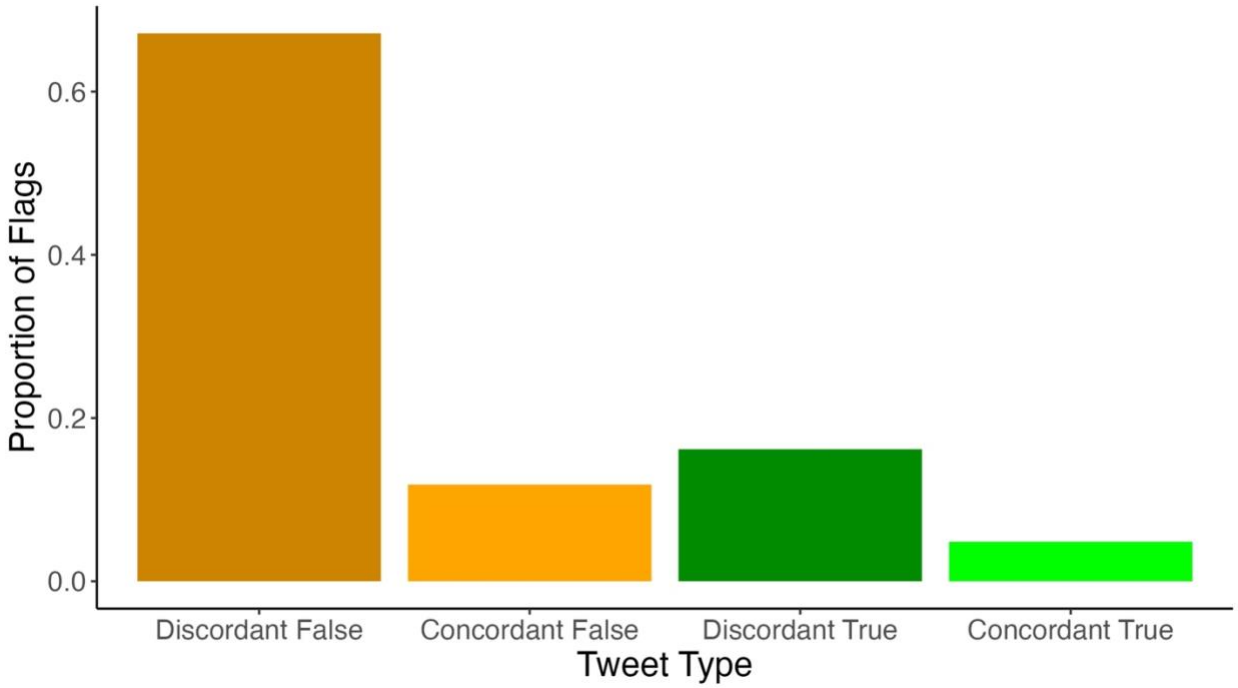

Figure 4. Distribution of Community Notes flags by tweet veracity and political concordance. Shown is a bar plot of the proportion of notes written about each tweet type for the 461 unique tweets evaluated by fact-checkers and Amazon Mechanical Turk workers on veracity and partisan favorability, respectively. The majority of notes are written on false, discordant tweets. Bar plot excludes notes from users for whom partisanship could not be inferred – e.g., if users did not follow any political accounts.

How does political extremity relate to the types of tweets that users flag? As expected, politically extreme users flag a greater proportion of politically discordant tweets (OLS regression predicting proportion of discordant notes by political extremity: b=0.051, SE=0.020, z(518)=2.564, p=.011; see SI Table S27). We do not find evidence that the proportion of false tweets flagged varies with political extremity (OLS regression predicting proportion of flags agreeing with fact-checkers: b=-0.021, SE=0.014, z(518)=-1.505, p=.133; see SI Table S28) – suggesting that in our field data, political extremity is not associated with flagging quality positively nor negatively. Putting this together with our finding that more politically extreme users contribute substantially more total flags, we also find that more politically extreme users flag a greater number of misleading tweets relative to accurate tweets (i.e., more politically extreme users have greater additive veracity discernment). Specifically, more politically extreme users flag more false discordant tweets relative to true discordant tweets (OLS regression predicting flag count by item type (holdout = False, Discordant) and political extremity; interaction term: b=-0.174, SE=0.054, z(5,772)=-3.234, p=.002; SI Table S29; Figure 5).

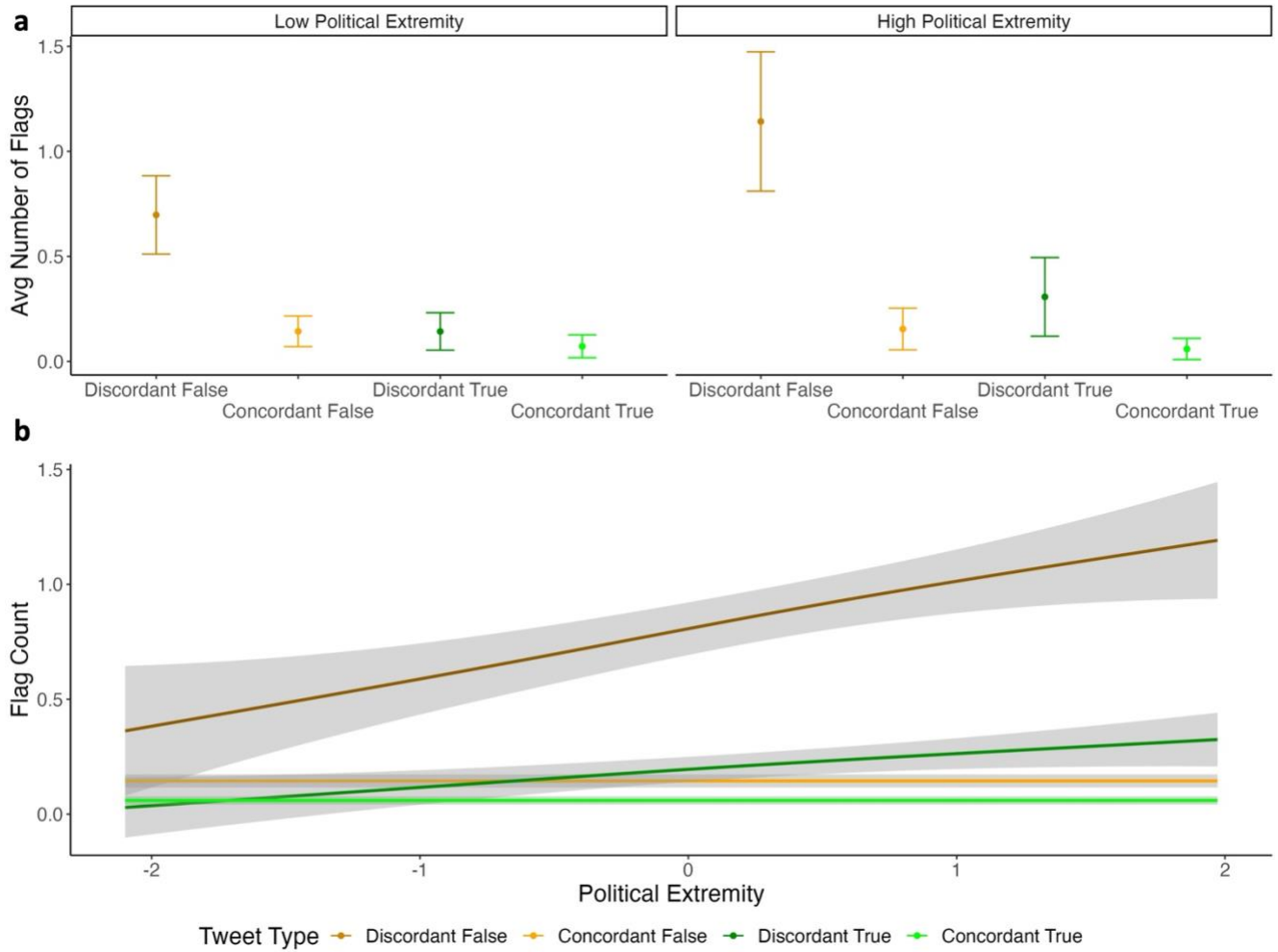

Figure 5. Politically extreme Community Notes users contribute more flags and maintain high veracity discernment. These users are more likely to flag false discordant tweets, and are also more likely to flag true, discordant tweets. However, overall flagging discernment is not worse for more politically extreme contributors. (a) Shown are the average number of tweets flagged as misleading by tweet type, median split by participant political extremity. Both subsets flag a greater overall number of false versus true headlines – and in particular flag mostly false, discordant headlines. Crucially, more politically extreme users flag a greater number of false, discordant flags (as well as true, discordant flags – but to a lesser extent). Error bars reflect 95% confidence intervals. (b) Shown is the predicted number of flags by tweet type and user political extremity. Convergent with our predictions and survey study results, overall fact-checking increases with political extremity. Furthermore, political extremity is particularly associated with flagging more false, discordant tweets. Overall flagging discernment rates (flagging false versus true tweets) remains relatively additively constant across political extremity. Error bars reflect 95% confidence intervals.

We also observe asymmetries in flagging by partisanship (as opposed to just by partisan extremity). Republican users flag a higher proportion of discordant notes overall (b=0.026, SE=0.013, z(518)=2.059, p=.040) and exhibit lower agreement with fact-checkers (b=-0.66, SE=0.010, z(518)=-6.672, p<.001; see SI Figure S3).

Discussion

These field data from X's (Twitter's) Community Notes demonstrate that more politically extreme users contribute more to an online community fact-checking system, in which the majority of users do not participate at all. Despite these users exhibiting greater bias in fact-checking by preferentially flagging discordant counter-partisan content, more politically motivated users maintain the same overall level of flagging discernment between accurate and misleading posts. Practically, it is useful to consider the contributions of low versus high politically extreme users (median split) on the subset of 461 tweets we had evaluated for veracity and partisan favorability. Less politically extreme users wrote 314 flags on tweets classified as false and 83 flags on tweets classified as true, while more politically extreme users contributed 632 flags on false tweets and 169 flags on true tweets. While the trust positive rate is nearly equivalent (79.1% versus 78.9%), excluding notes from more politically extreme users would result in only about one-third as many correct notes contributed. Politically motivated users are integral to the sustained output of crowdsourced fact-checking systems – and can promote a high quantity of evaluations without sacrificing quality and flagging discernment.

General Discussion

Together, these data suggest that highly politically motivated individuals may actually be a boon for crowd-based approaches for helping social media companies identify misinformation at scale. Crowdsourced fact-checks are a public goods problem: flagging false content is societally beneficial but requires individual users to invest time and effort evaluating content. Most people may not have strong enough accuracy-directed motivations to write fact-checks on just any false content they encounter – but those with additional motivations may be willing to engage in such fact-checking. Here, we observe that partisan motivations may help solve this public goods problem. Partisan motivation and extremity are associated not only with flagging a greater number of posts but are also associated with similar – and in some cases, better – levels of flagging discernment. Specifically, political motivation corresponds with increased flagging of content that is both false and politically discordant.

Our findings show that accuracy and partisan motivations may jointly help promote voluntary crowdsourced fact-checking efforts by helping sustain both the quantity and quality and contributions. Theoretically, our results suggest that political and partisan motives are not necessarily in fundamental opposition to accuracy goals – rather, individuals may be motivated to engage in flagging of content because they have both partisan and accuracy motivations. And practically, our data show that highly politically motivated community fact-checkers flag twice as many posts as less politically motivated contributors, without any negative change in the quality of flags. Thus, any applied crowdsourced fact-checking system should not try to omit or discourage these politically motivated users from contributing flags – rather, such systems may be able to

address the underlying problem of content moderation sparsity because of these dually motivated individuals.

Our results show that the inputs into voluntary online crowdsourced fact-checking systems can be high quality, even amongst contributors with strong political and partisan motivations. In our current work, we do not examine how these useful evaluations should be best aggregated to summarize ratings or inform interventions. Future work should examine the efficacy of aggregation approaches best suited for summarizing crowdsourced ratings and mitigating the risk of bad actors hijacking the flagging process. For instance, despite our current findings, motivated bad actors in future iterations could work together to undermine good faith fact-checking efforts by purposefully flagging accurate content they dislike or disagree with in a coordinated manner. This may include the most extreme partisan zealots – for whom partisan motives could so strongly outweigh accuracy motives that they are indifferent to veracity, or perhaps even prefer flagging true, discordant content. Fortunately, prior work suggests that efforts to "game" the algorithm in crowdsourced review processes may be limited and offset by intentionally maintaining politically balanced or representative crowd ratings (Epstein et al., 2020). Nonetheless, researchers and platforms should continue to investigate and try to identify coordinated efforts and individual users attempting to game or undermine moderation efforts.

Relatedly, simple aggregation of crowd judgments (e.g., majority vote) could lead to a "tyranny of the majority" in cases where the majority of users hold an incorrect or biased belief. To mitigate this risk, researchers and practitioners have developed a variety of approaches – most notably, Community Notes itself currently uses "bridging-based ranking" to determine which notes are most widely helpful given the contents of notes and whether a range of individuals with different viewpoints rate the note as helpful (Ovadya & Thorburn, 2022; Wojcik et al., 2022). One concern with this approach is that requiring cross-ideology consensus may severely restrict the number of fact-checks that become broadly viewable on the platform (Bak-Coleman, 2023) – which is particularly an issue if not enough users evaluate the helpfulness of counter-partisan fact-checks to begin with. Thus, another avenue for future research is investigating whether soliciting or inducing users with greater political motivation can be beneficial for increasing ratings of counter-partisan fact-checks. Our theory and findings predict that recruiting and mobilizing politically motivated users should be beneficial for increasing the number of surfaced helpful fact-checks in a bridging-based evaluation system.

Limitations

We note several limitations to our current work. One limitation of our analyses is that they focus on crowdsourced fact-checking among only American participants. Although recent work has demonstrated the cross-national generalizability of effective laypeople panel fact-checking (Arechar et al., 2023), such findings do not examine whether partisan motives may help or undermine such content moderation efforts in an applied, voluntary system. Future work should assess the interplay between political and accurate motivations in other national and political contexts.

Our analyses also largely focus on fact-checking of political content. Our survey study only examined flagging of political news, and 79% of Community Notes fact-checks were written about tweets raters classified as primarily being about 'politics' (74.6% of our reviewed tweets; see SI Section 3b). Partisan motivations may be less useful for promoting evaluation of posts about other

domains such as science, business, or advertising – and further research may investigate other beneficial motivators or predictors of fact-checking across different content categories.

Another caveat on our current work is that we focus on observational, descriptive data – we do not experimentally induce greater partisan or accuracy motivation in our survey or on Community Notes. Previous work has demonstrated that shifting attention to (Pennycook, Epstein, et al., 2021) or financially incentivizing (Rathje et al., 2023) accuracy can increase news discernment, even for highly partisan individuals (Martel, Rathje, et al., 2024). Future directions may extend this to examining effects of prompting accuracy on flagging – and similarly may examine whether politically motivating or (de)polarizing individuals may likewise causally affect flagging. Nonetheless, our current analyses show that individuals with high political motivations are integral for sustaining a high number of quality contributions within a voluntary fact-checking system.

Conclusion

Widely held theories posit that political motivations undermine high quality information sharing and are fundamentally in opposition to accuracy directed motivations (Osmundsen et al., 2021; Van Bavel et al., 2024; Van Bavel & Pereira, 2018). Coupled with robust evidence of partisan selective fact-checking (Allen et al., 2022; Shin & Thorson, 2017), a major concern looming over the potential of crowdsourced fact-checking is that individuals will flag too much – inappropriately flagging true but counter-partisan content they dislike or disagree with. Our current work suggests that the much bigger hurdle is insufficient flagging – most people are not motivated enough by accuracy motivations alone to fact-check anything. Rather than undermine the system, individuals with strong partisan motives may actually help address this challenge by flagging high volumes of content without declining quality. Indeed, we find that community fact-checkers – and highly partisan motivated community fact-checkers in particular – are most likely to flag content that is both misleading *and* counter-partisan. In sum, we argue that crowdsourcing can successfully work to identify misinformation at scale because of – rather than in spite of – partisan motivations.

References

Ahmad, W., Sen, A., Eesley, C., & Brynjolfsson, E. (2024). Companies inadvertently fund online misinformation despite consumer backlash. *Nature*, *630*(8015), 123–131.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021a). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393. https://doi.org/10.1126/sciadv.abf4393

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021b). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393. https://doi.org/10.1126/sciadv.abf4393

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3502040

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), eadk3451. https://doi.org/10.1126/science.adk3451

Anderson, A. S. and M. (2016, December 19). Online Shopping and E-Commerce. *Pew Research Center*. https://www.pewresearch.org/internet/2016/12/19/online-shopping-and-e-commerce/

Anderson, E. T., & Simester, D. I. (2014). Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception. *Journal of Marketing Research*, *51*(3), 249–269. https://doi.org/10.1509/jmr.13.0209

ANES. (2020). 2020 Time Series Study. *ANES | American National Election Studies*. https://electionstudies.org/data-center/2020-time-series-study/

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., & Stagnaro, M. N. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513.

Avaaz. (n.d.). *How Facebook can Flatten the Curve of the Coronavirus Infodemic*. Avaaz. Retrieved June 28, 2024, from https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/

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.

Bak-Coleman, J. (2023). *Limiting factors in the effectiveness of crowd-sourced labeling for combating misinformation*. https://osf.io/preprints/socarxiv/ahm27/

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, *9*(2), 430–437.

Coscia, M., & Rossi, L. (2020). Distortions of political bias in crowdsourced misinformation flagging. *Journal of The Royal Society Interface*, *17*(167), 20200020. https://doi.org/10.1098/rsif.2020.0020

Dahlke, R., & Hancock, J. (2022). The effect of online misinformation exposure on false election beliefs. *OSF Preprints*, *17*. https://files.osf.io/v1/resources/325tn/providers/osfstorage/6376a5514e8c3c04e83b4bd6

Donovan, J. (2020). Social-media companies must flatten the curve of misinformation. *Nature*. https://europepmc.org/article/med/32291410

Druckman, J. N. (2012). THE POLITICS OF MOTIVATION. *Critical Review*, *24*(2), 199–216. https://doi.org/10.1080/08913811.2012.711022

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

Duncan, M. (2022). Selective rating: Partisan bias in crowdsourced news rating systems. *Journal of Information Technology & Politics*, *19*(3), 360–375. https://doi.org/10.1080/19331681.2021.1997867

Epstein, Z., Lin, H., Pennycook, G., & Rand, D. (2022). *How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation,*

and unpredictability on social media (arXiv:2207.07562). arXiv. http://arxiv.org/abs/2207.07562

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the Crowd Game the Algorithm?: Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3313831.3376232

Fong, J., Guo, T., & Rao, A. (2023). Debunking Misinformation About Consumer Products: Effects on Beliefs and Purchase Behavior. *Journal of Marketing Research*, 002224372211470. https://doi.org/10.1177/00222437221147088

Galton, F. (1949). Vox Populi (1907) Nature, n. 1949, vol. 75, pp. 450-451 (traduzione di Romolo Giovanni Capuano\copyright). *Nature*, *75*, 450–451.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1). https://www.tsjournal.org/index.php/jots/article/view/15

Goes, P. B., Guo, C., & Lin, M. (2016). Do Incentive Hierarchies Induce User Effort? Evidence from an Online Knowledge Exchange. *Information Systems Research*, *27*(3), 497–516. https://doi.org/10.1287/isre.2016.0635

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council*, *9*(3), 4.

Instagram. (n.d.). *Combatting Misinformation on Instagram | Instagram*. Retrieved June 28, 2024, from https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

Jia, C., Lam, M. S., Mai, M. C., Hancock, J. T., & Bernstein, M. S. (2024). Embedding Democratic Values into Social Media AIs via Societal Objective Functions. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–36. https://doi.org/10.1145/3641002

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

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., & Basol, M. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 1–9.

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

Lin, H., Lasser, J., Lewandowsky, S., Cole, R., Gully, A., Rand, D. G., & Pennycook, G. (2023). High level of correspondence across different news domain quality rating sets. *PNAS Nexus*, *2*(9), pgad286. https://doi.org/10.1093/pnasnexus/pgad286

Lin, H., Lasser, J., Lewandowsky, S., Cole, R., Gully, A., Rand, D., & Pennycook, G. (2022). *High level of agreement across different news domain quality ratings*. https://www.researchgate.net/profile/Hause-Lin/publication/366176520_High_level_of_correspondence_across_different_news_domain_quality_rating_sets/links/6398ff7211e9f00cda418f3d/High-level-of-correspondence-across-different-news-domain-quality-rating-sets.pdf

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

Luther, M. S., Erica Ryan, Joel. (2023, June 21). Misinformation spreads, but fact-checking has leveled off. *Poynter*. https://www.poynter.org/fact-checking/2023/duke-reporters-lab-fact-checking-census-2023/

Marie, A., & Petersen, M. B. (2023). *Motivations to affiliate with audiences drive the sharing of partisan (mis) information on social media*. https://osf.io/preprints/nmg9h/

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds Can Effectively Identify Misinformation at Scale. *Perspectives on Psychological Science*, *19*(2), 477–488. https://doi.org/10.1177/17456916231190388

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

Martel, C., Rathje, S., Clark, C. J., Pennycook, G., Van Bavel, J. J., Rand, D. G., & Van Der Linden, S. (2024). On the Efficacy of Accuracy Prompts Across Partisan Lines: An Adversarial Collaboration. *Psychological Science*, *35*(4), 435–450. https://doi.org/10.1177/09567976241232905

McCabe, S. D., Ferrari, D., Green, J., Lazer, D. M., & Esterling, K. M. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, *630*(8015), 132–140.

Meta. (n.d.-a). *How Meta's third-party fact-checking program works*. How Meta's Third-Party Fact-Checking Program Works. Retrieved June 28, 2024, from https://www.facebook.com/facebookmedia

Meta. (n.d.-b). *Request review of a fact-check rating on Facebook, Instagram, and Threads*. Meta Business Help Center. Retrieved July 1, 2024, from https://www.facebook.com/business/help/997484867366026

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), 7144.

Nakano/Bloomberg, A. C. and E. (2023, July 19). *Twitter's Surge in Harmful Content Keeps Advertiser Away*. TIME. https://time.com/6295711/twitters-hate-content-advertisers/

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015.

Ovadya, A., & Thorburn, L. (2022). Bridging-based ranking. *Harvard Kennedy School Belfer Center for Science and International Affairs*. https://lukethorburn.com/files/BridgingBasedRanking-PluralitySpringSymposium.pdf

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021a). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), 25293.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021b). A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology*, *7*(1), 25293. https://doi.org/10.1525/collabra.25293

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

Perspective API - How it works. (n.d.). Retrieved June 26, 2024, from https://www.perspectiveapi.com/how-it-works/

Peterson, E., & Iyengar, S. (2021). Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading? *American Journal of Political Science*, *65*(1), 133–147. https://doi.org/10.1111/ajps.12535

Rao, A. (2022). Deceptive Claims Using Fake News Advertising: The Impact on Consumers. *Journal of Marketing Research*, *59*(3), 534–554. https://doi.org/10.1177/00222437211039804

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, *7*(6), 892–903.

Resnick, P., Alfayez, A., Im, J., & Gilbert, E. (2023). Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, *2*(2), 263391372311734. https://doi.org/10.1177/26339137231173407

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce* (pp. 127–157). Emerald Group Publishing Limited. https://www.emerald.com/insight/content/doi/10.1016/S0278-0984(02)11030-3/full/html

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), 329–336.

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

Silverman, H. (n.d.). *Helping Fact-Checkers Identify False Claims Faster | Meta*. Retrieved June 28, 2024, from https://about.fb.com/news/2019/12/helping-fact-checkers/

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor. https://books.google.com/books?hl=en&lr=&id=_t2KDQAAQBAJ&oi=fnd&pg=PR11&dq=wisdom+of+crowds+surowiecki&ots=yAwnQGF76r&sig=fJ3tl6VTnG3r8Cx8jjTZ9iBPCdc

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tappin, B. M., Pennycook, G., & Rand, D. G. (2021). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General*, *150*(6), 1095.

TikTok. (2023, January 19). *Combating harmful misinformation*. TikTok. https://www.tiktok.com/transparency/en-us/combating-misinformation/

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

Van Bavel, J. J., Rathje, S., Vlasceanu, M., & Pretus, C. (2024). Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, *56*, 101787. https://doi.org/10.1016/j.copsyc.2023.101787

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *The World Wide Web Conference*, 2056–2067. https://doi.org/10.1145/3308558.3313684

Weisberg, H. F., & Rusk, J. G. (1970). Dimensions of candidate evaluation. *American Political Science Review*, *64*(4), 1167–1185.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., & Baxter, J. (2022). *Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation* (arXiv:2210.15723). arXiv. http://arxiv.org/abs/2210.15723

Yasseri, T., & Menczer, F. (2023). Can Crowdsourcing Rescue the Social Marketplace of Ideas? *Communications of the ACM*, *66*(9), 42–45. https://doi.org/10.1145/3578645

Zhao, A., & Naaman, M. (2023). Insights from a Comparative Study on the Variety, Velocity, Veracity, and Viability of Crowdsourced and Professional Fact-Checking Services. *Journal of Online Trust and Safety*, *2*(1). https://www.tsjournal.org/index.php/jots/article/view/118

Acknowledgments

We are thankful to A. Arechar for invaluable assistance in survey data collection and in collecting fact-checker ratings for field study tweets. We gratefully acknowledge funding from the William and Flora Hewlett Foundation, the John Templeton Foundation, the TDF Foundation, and Alfred P. Sloan Foundation Grant #2021-16891. C.M. is supported by the NSF Graduate Research Fellowship (Grant No. 174530).

Competing interests

Other research by G.P. and D.G.R. is funded by gifts from Google and Meta. D.G.R. was formerly on the advisory board of Twitter's Birdwatch (now X's Community Notes) while contributing to this research.

Supplementary Information
for
Harnessing Partisan Motives to Combat Misinformation

Table of Contents

1. Simple Model of Misinformation Flagging

Here we develop a simple model of crowdsourced misinformation flagging. In this model, accuracy motives drive people to report inaccurate content regardless of its political alignment, while partisan motives drive people to report counter-partisan content (and not to report co-partisan content) regardless of accuracy. Thus, the utility derived from flagging a piece of content with a given veracity and political alignment is:

$$U_{flagging} = \beta_1 I_{False} + \beta_2 I_{Discordant} - \beta_2 I_{Concordant} - 1,$$

Where β_1 reflects an individual's propensity to flag inaccurate content (i.e., their accuracy motivation), β_2 reflects an individual's propensity to flag politically discordant content and *not* flag politically concordant content (i.e., their partisan motivation), the cost of flagging (e.g., time, effort) is normalized to 1, and I is the indicator function.

Using this utility function and mapping utility into choice using a logistic function, we can examine the likelihood of different kinds of content being flagged as a function of the preferences of the flagger. First, content that is true and concordant is never flag, as neither motive drives the flagging of this content. Second, a stronger partisan motive β_2 always increases political bias (the tendency to flag counter-partisan content more than co-partisan content).

The effects on flagging discernment (the difference in flagging false versus true headlines), however, is more complex. The probability of flagging false discordant, false concordant, and true discordant content – along with flagging discernment, are shown as a function of the preference parameters β_1 and β_2 in Figure S1. When accuracy motives are strong ($\beta_1 > 1$), increasing partisan motives always reduces flagging discernment. This is because in the absence of any partisan motivation, accuracy motivated individuals are already flagging nearly all false content – but as partisan motives increase, these individuals flag fewer false co-partisan headlines and more true discordant headlines. When accuracy motives are weaker ($\beta_1 < 1$), on the other hand, adding partisan motivation (to a point) can actually increase flagging discernment. In this case, individuals are insufficiently motivated to flag content simply for being inaccurate, or simply for being counter-partisan – but will flag content that is both false and counter-partisan as partisan motives increase. Here, partisan motivation is needed to drive participation in flagging, and does so by increasing preferential flagging of false discordant content. Without partisan motivation, lower accuracy motivated individuals would flag very little – a problem that echoes contributions to similar crowdsourced public game provisioning systems such as online reviews (A. S. and M. Anderson, 2016; E. T. Anderson & Simester, 2014; Goes et al., 2016)

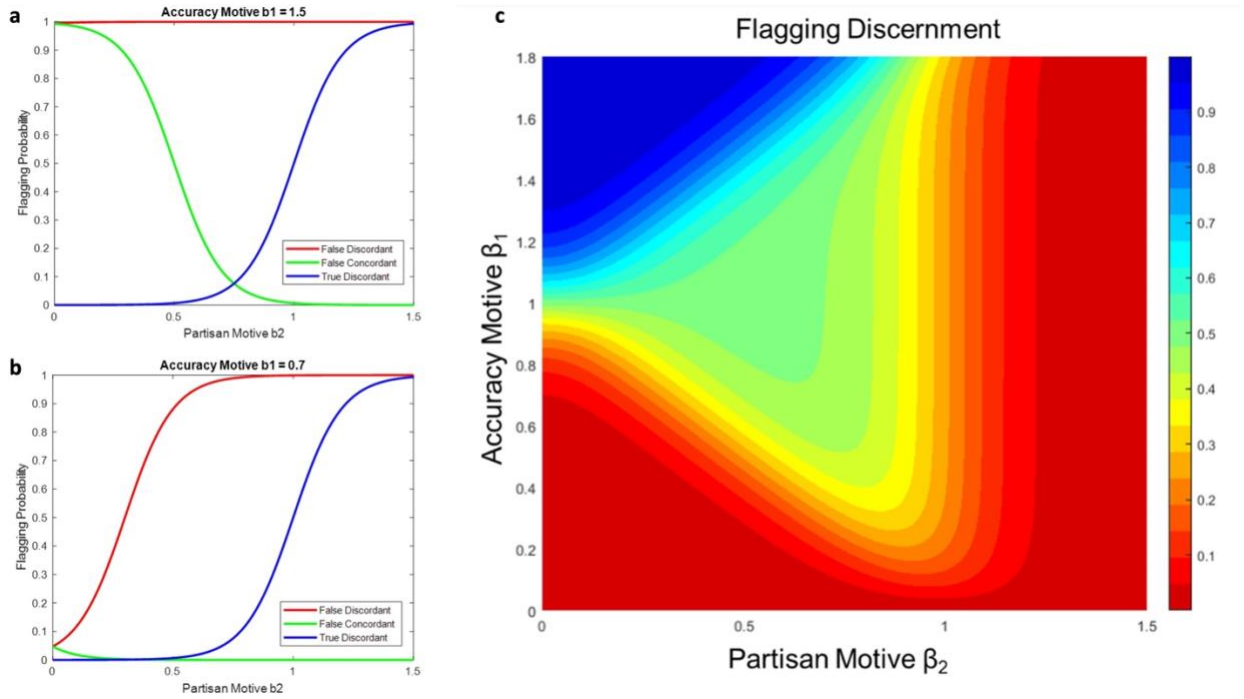

Figure S1. Partisan motives may increase discernment when accuracy motives are weak by increasing flagging of false discordant content. Shown are the results of a simplified formal model of a misinformation flagging system. In this model, the utility of flagging is predicted by the falsity of content (accuracy motivation), the political discordance of content (partisan motive to flag counter-partisan), and the political concordance of content (partisan motive *not* to flag co-partisan), as well as a standardized cost of flagging (e.g., time and effort to fact-check). Flagging implementation mapping utility to flagging decision is then modeled via a logistic function $P(flag) = \frac{1}{1+e^{-\lambda * U_{flagging}}}$, using λ=100. (a) For an individual with high accuracy motivation (β_1=1.5), flagging discernment in the absence of partisan motives (β_2=0) is high – but increased partisan motives result in decreased discernment by way of increased flagging of true discordant content and decreased flagging of false concordant content. (b) For an individual with weak accuracy motivation (β_1=0.7), increasing the strength of the partisan motive is actually beneficial – although it slightly decreases flagging of false concordant content, it substantially increases flagging of false discordant posts, and thereby increases sharing discernment. (b) A heatmap plot of flagging discernment (difference between false and true flagging) over the space of accuracy and partisan motives. Whenever β_1<1, discernment is maximized at an intermediate level of partisan motive.

Thus, our flagging model predicts that while partisan motives are necessarily deleterious for discernment when accuracy motives are strong, increasing partisan motives can substantially increase flagging of false content when accuracy motives are weaker (Figure S1b). Our model posits that the impact of partisan motives on flagging discernment is not straightforward – accuracy and partisan motives may be at odds in some contexts but work together to increase flagging discernment when accuracy motivations are weaker. Our survey and field empirical data suggests that these are contexts where partisan motives are helpful rather than harmful.

2. Study 1: Survey Study Supplement

a. Headline Pretesting

Our pretest asked participants (*N*=1,982, recruited from Lucid) to each rate 10 randomly selected news headlines from a corpus of 121 false and 157 true headlines on a number of dimensions. False and misleading headlines were selected from the fact-checking website Snopes.com and verified by third-party fact-checkers to be factually inaccurate. True headlines were selected from reputable mainstream news sources, and further assessed to verify veracity. All headlines were presented in 'Facebook format' with a headline, byline, and image (Pennycook, Binnendyk, et al., 2021b). Of primary interest, participants were asked the following question about partisan favorability: "Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans?" (1=More favorable for Democrats, 2=Moderately more favorable for Democrats, 3=Slightly more favorable for Democrats, 4=Slightly more favorable for Republicans, 5=Moderately more favorable for Republicans, 6=More favorable for Republicans). Participants were also asked to evaluate the plausibility of each headline: "What is the likelihood that the above headline is true? (1=Extremely unlikely, 7=Extremely likely). We used data from these questions to select the 60 news items used in our survey study, such that the pro-Democratic items were similarly partisan and plausible as the pro-Republican items within the true and false categories. The pretest was conducted on 22 April 2022.

The 20 non-news, entertainment items used in our survey were selected from a prior study on the effects of social cues on engagement (Epstein et al., 2022). These items were categorically non-political, non-news content, selected from sensational tabloid and clickbait outlets. These items were included in our study to (i) improve the ecological validity of our task structure by not only including political news items in participants' feeds, and (ii) provide a politically neutral comparison group for flagging rates of politically discordant and concordant headlines.

b. Principal Component Analysis Loadings

As preregistered, we conducted a principal component analysis (PCA) on our political motivation survey measures: out-party dislike, issue polarization, and political knowledge. The principal component loadings are reported below in Table S1. We find that all three measures have high weights in the first component, which we describe in the main text as overall political motivation. Out-party dislike has the greatest weight in the second component, and political knowledge has a strong negative weight in this component. Issue polarization is weighted highly in the third component, which also has a strong negative weight for political knowledge. Thus, the second and third principal components may be interested as increased out-party dislike and issue polarization, respectively, at the expense of political knowledge (rather than *independent* of political knowledge).

	PC1	**PC2**	**PC3**
Out-party Dislike	0.522	0.812	-0.263
Issue Polarization	0.626	-0.154	0.765

Political Knowledge	0.580	-0.563	-0.588

Table S1. *Principal component loadings from PCA of out-party dislike, issue polarization, and political knowledge.*

c. Regression Tables

The following subsections contain full regression tables and robustness checks for our survey study analyses. For all relevant tables, please note the following: 'Exclude True Independents' indicates that the analyses exclude individuals with partisanship scores equal to the scale midpoint (i.e., they did not report leaning towards the Democratic nor Republican party; remaining included *n*=1,994); 'Attention + Practice Filter' indicates that the analyses only include individuals who both (i) pass greater than the median number of pre-task instructional attention items (greater than 2 out of 4 attention items) and (ii) correctly answer the practice main task items upon at least their second attempt (total remaining included *n*=957).

Analysis plan. For each headline in our main flagging task, participants could either scroll past and not flag it as misleading (flag = 0) or click a flag button and write a free response justification of why they flagged the headline as false or misleading (flag = 1). From this, we preregistered constructing the following dependent variables. (i) At the headline-participant level, whether a participant flagged a given headline (0 = no flag, 1 = flag). (ii) At the participant level, how many flags a participant wrote (count variable). (iii) At the participant level, the proportion of flagged headlines that are politically discordant or (iv) that are not false. And (v) at the headline type-individual level (Discordant-False, Concordant-False, Discordant-True, Concordant-True, Sensational [non-political]), how many flags were written (count variable). Headline partisanship (pro-Democrat, pro-Republican, sensational) and headline veracity (true, false, sensational) were based on pretesting (see SI section 2a). Participant partisanship was assessed on a 7-point scale from our political identity item (1=Strong Democrat, 2=Weak Democrat, 3=Lean Democrat, 4=Independent/Other + Neither Lean, 5=Lean Republican, 6=Weak Republican, 7=Strong Republican). Headline political concordance was scored as 'concordant' if the headline and participant matched on Democrat or Republican lean, and 'discordant' if the headline and participant partisanship were opposed on Democrat or Republican lean. For true independent participants (partisanship = 4), as preregistered we imputed Democratic or Republican partisanship in order to define concordance via the MICE package (Van Buuren & Groothuis-Oudshoorn, 2011) using age, gender, race, education, belief in God(s), feelings towards Democratic and Republican Party voters, political issue items, political ideology, and economic and social conservatism as predictors.

Our preregistration also specified the following independent variable specifications. (i) Political knowledge specified as the sum of correct responses on our four-item measure, *z*-scored; (ii) Issue polarization as assessed by the absolute value of the difference between the standardized scaled sum of 11 issue items and the scale midpoint, *z*-scored (with 'Don't Know' responses scored as missing data); (iii) Out-party animosity scored from 0 to 100 as per our feeling thermometer item, with out-party determined by Democrat or Republican party identification (and minimum thermometer rating recorded for true independents), reverse-coded and *z*-scored; (iv) Participant partisanship assessed 1-7 as per above political identity item, then *z*-scored; (v) age, *z*-scored; (vi) gender recorded as 1 = female, 0 = not female, then *z*-scored; (vii) race recorded as 1 = White-

only, 0=not White-only, then *z*-scored; (viii) college recorded as 1 = at least attended college, 0 = did not attend any college, then *z*-scored; and (ix) attention, scored as the sum of correct responses across four instructional attention items, *z*-scored. For all analyses, we define participant partisanship, age, gender, race, college, and attention as our standard controls.

Our first analysis was preregistered to use logistic regression to predict flagging by headline type (5-level factor; baseline = Discordant-False), with robust standard errors clustered by both participant and headline. Our key prediction was that there would be negative simple effects for each headline type relative to the Discordant-False baseline, such that false, discordant headlines were most likely to be flagged as misleading relative to other headline types. Our second preregistered analysis specified that at the participant level, we would conduct three separate quasi-Poisson (QP) general linear models to predict total flag count by either political knowledge, issue polarization, or out-party animosity, with standard controls and HC2 robust standard errors. We also specified conducting this model with all three political motivation measures in the same model, as well as performing a similar model instead with the first three principal components from a principal components analysis (PCA) of political knowledge, issue polarization, and out-party animosity. Our key prediction was that there would be positive simple effects for our political motivation predictors on flag count. Our third preregistered analysis specified that for instances when a flag was written, we would conduct a linear model predicting the proportion of discordant flags a participant recorded (Discordant-False or Discordant-True, only) by either political knowledge, issue polarization, or out-party animosity, and standard controls; with analytic weighting by total flag count and HC2 robust standard errors. We also preregistered conducing a similar model with all three political motivation items in the same regression, as well as a model with the first three principal components from these political items as specified above. Our key prediction was that there would be positive simple effects for political motivation predictors on the proportion of discordant flags produced. Our fourth preregistered analysis set was nearly identical, except instead predicting the proportion of false-positive (Discordant-True, Concordant-True, sensational non-political) flags written. Our key prediction was that there would be negative simple effects for political motivation on the proportion of false-positive flags written. Finally, our fifth preregistered analysis set specified that looking at instances where a flag was written, we would conduct quasi-Poisson models to predict the number of flags by headline type at the individual level, by headline type (5-level factor, baseline = Discordant-False) and either political knowledge, issue polarization, or out-party animosity; allowing for interactions between headline type and each political motivation predictor plus standard controls (each interacted with headline type), and clustered standard errors by participant. We also preregistered performing similar models except including all political motivation predictors in the same model, as well as replacing them with the first three principal components of our PCA of these items. Our key prediction was that there would be negative interaction terms between political motivation predictors and headline type factor dummies relative to the Discordant-False baseline, such that more politically motivated participants flagged more false, discordant headlines than other headline types.

As secondary analyses and robustness checks we also preregistered the following. (i) All count analyses also conducted as linear models, in place of quasi-Poisson regressions; (ii) All relevant analyses excluding true independents, rather than imputing partisanship; (iii) Conducting additional analyses including quadratic terms for political knowledge, issue polarization, and out-party animosity to test for potential nonlinear relationships; (iv) Conducting Heckman 2-step models to predict total flag count, in order to model a potential two-stage flagging selection

process; and (v) Performing all analyses filtering by high attention check performance and correctly completing the practice task items.

All analysis code is available here: https://osf.io/3ngbt/.

i. Flagging Likelihood

	Dependent variable:		
	Headline-level flagging		
	Logit Model	Exclude True Independents	Attention + Practice Filter
	(1)	(2)	(3)
ItemTypeDT	-1.580***	-1.568***	-2.054***
	(0.140)	(0.143)	(0.188)
	t = -11.325	t = -11.002	t = -10.915
	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.234***	-0.253***	-0.284***
	(0.037)	(0.040)	(0.062)
	t = -6.268	t = -6.361	t = -4.582
	p = 0.000	p = 0.000	p = 0.00001
ItemTypeCT	-1.798***	-1.842***	-2.432***
	(0.118)	(0.120)	(0.158)
	t = -15.182	t = -15.315	t = -15.435
	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.688***	-1.709***	-2.195***
	(0.125)	(0.125)	(0.164)
	t = -13.553	t = -13.720	t = -13.416
	p = 0.000	p = 0.000	p = 0.000
Constant	-1.574***	-1.516***	-1.135***
	(0.087)	(0.086)	(0.106)
	t = -18.150	t = -17.540	t = -10.670
	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

Logit regression predicting flagging likelihood by item type (intercept=Discordant False), with two-way robust standard errors clustered by participant and headline

Table S2. *Headline-level flagging predicted by item type. Across specifications, we observe negative effects of headline type categories on flagging likelihood relative to the baseline level*

discordant, false – such that participants were less likely to flag all other headline types relative to discordant, false headlines.

ii. Flag Count

	Dependent variable:			
	Flag count			
	QP Model	OLS Model	QP Exclude True Independents	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)
PC1	0.228***	1.641***	0.224***	0.166***
	(0.023)	(0.166)	(0.025)	(0.033)
	t = 9.844	t = 9.888	t = 9.001	t = 5.102
	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
PC2	-0.101**	-0.739***	-0.106**	-0.062
	(0.032)	(0.190)	(0.034)	(0.047)
	t = -3.162	t = -3.884	t = -3.114	t = -1.309
	p = 0.002	p = 0.0002	p = 0.002	p = 0.191
PC3	-0.013	0.034	-0.012	-0.026
	(0.035)	(0.240)	(0.038)	(0.047)
	t = -0.383	t = 0.142	t = -0.314	t = -0.562
	p = 0.702	p = 0.888	p = 0.754	p = 0.574
zPartisan	-0.033	-0.198	-0.032	-0.064
	(0.027)	(0.179)	(0.027)	(0.039)
	t = -1.229	t = -1.106	t = -1.198	t = -1.648
	p = 0.220	p = 0.269	p = 0.232	p = 0.100
zage	0.104**	0.671**	0.086*	0.034
	(0.034)	(0.223)	(0.037)	(0.047)
	t = 3.032	t = 3.006	t = 2.342	t = 0.723
	p = 0.003	p = 0.003	p = 0.020	p = 0.470
zfemale	-0.077**	-0.560**	-0.089**	-0.101**
	(0.027)	(0.175)	(0.029)	(0.037)
	t = -2.908	t = -3.196	t = -3.116	t = -2.733
	p = 0.004	p = 0.002	p = 0.002	p = 0.007
zwhite	0.069*	0.389*	0.080*	0.106*
	(0.032)	(0.175)	(0.034)	(0.051)
	t = 2.154	t = 2.219	t = 2.323	t = 2.096
	p = 0.032	p = 0.027	p = 0.021	p = 0.037
zcollege	0.050	0.288	0.045	0.041

	(0.027)	(0.176)	(0.029)	(0.039)
	t = 1.844	t = 1.637	t = 1.546	t = 1.057
	p = 0.066	p = 0.102	p = 0.123	p = 0.291
zScreenTot	0.046	0.198	0.046	
	(0.031)	(0.196)	(0.034)	
	t = 1.464	t = 1.012	t = 1.335	
	p = 0.144	p = 0.312	p = 0.182	
Constant	1.808***	6.694***	1.824***	1.946***
	(0.028)	(0.179)	(0.031)	(0.047)
	t = 63.620	t = 37.313	t = 58.987	t = 41.795
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

Regression predicting flag count by first three PCs of political knowledge, issue polarization, and out-party dislike, and standard controls, with HC2 robust SEs

Table S3. *Total flag count predicted by political motivation principal components and standard controls. Across specifications, we find that overall political motivation (PC1) is associated with greater flag count – such that more politically motivated participants flag more headlines.*

	Dependent variable:				
	Flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
zPK	0.260***	1.767***	0.261***	0.272***	0.193***
	(0.032)	(0.210)	(0.035)	(0.033)	(0.043)
	t = 8.196	t = 8.421	t = 7.508	t = 8.140	t = 4.449
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00001
I(zPK2)				-0.027	
				(0.028)	
				t = -0.950	
				p = 0.342	
zPartisan	-0.054*	-0.339	-0.053	-0.055*	-0.086*
	(0.027)	(0.179)	(0.027)	(0.027)	(0.038)
	t = -1.982	t = -1.891	t = -1.953	t = -2.013	t = -2.237
	p = 0.048	p = 0.059	p = 0.051	p = 0.045	p = 0.026
zage	0.115**	0.761***	0.098**	0.115***	0.032
	(0.035)	(0.226)	(0.037)	(0.035)	(0.048)
	t = 3.284	t = 3.368	t = 2.609	t = 3.299	t = 0.668
	p = 0.002	p = 0.001	p = 0.010	p = 0.001	p = 0.505
zfemale	-0.066*	-0.456**	-0.074**	-0.068*	-0.092*
	(0.027)	(0.175)	(0.028)	(0.027)	(0.037)
	t = -2.468	t = -2.603	t = -2.619	t = -2.529	t = -2.496
	p = 0.014	p = 0.010	p = 0.009	p = 0.012	p = 0.013
zwhite	0.071*	0.395*	0.083*	0.072*	0.111*
	(0.032)	(0.177)	(0.035)	(0.032)	(0.051)
	t = 2.204	t = 2.233	t = 2.381	t = 2.229	t = 2.199
	p = 0.028	p = 0.026	p = 0.018	p = 0.026	p = 0.028
zcollege	0.053	0.314	0.047	0.053	0.041

	(0.028)	(0.178)	(0.029)	(0.028)	(0.039)
	t = 1.920	t = 1.768	t = 1.611	t = 1.916	t = 1.047
	p = 0.055	p = 0.078	p = 0.108	p = 0.056	p = 0.296
zScreenTot	0.072*	0.424*	0.069*	0.071*	
	(0.031)	(0.192)	(0.034)	(0.031)	
	t = 2.331	t = 2.209	t = 1.998	t = 2.299	
	p = 0.020	p = 0.028	p = 0.046	p = 0.022	
Constant	1.819***	6.705***	1.839***	1.844***	1.970***
	(0.028)	(0.181)	(0.031)	(0.039)	(0.045)
	t = 64.184	t = 36.994	t = 60.136	t = 46.698	t = 43.482
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

Regression predicting flag count by political knowledge and standard controls, with HC2 robust SEs

Table S4. *Total flag count predicted by political knowledge and standard controls. Across specifications, we find that political knowledge is associated with greater flag count – such that more politically knowledgeable participants flag more headlines.*

	Dependent variable:				
	Flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
zIssuePol	0.214***	1.606***	0.214***	0.246***	0.145***
	(0.026)	(0.205)	(0.027)	(0.038)	(0.034)
	t = 8.283	t = 7.818	t = 7.839	t = 6.509	t = 4.271
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00002
I(zIssuePol2)				-0.028	
				(0.023)	
				t = -1.190	
				p = 0.235	
zPartisan	-0.025	-0.123	-0.024	-0.026	-0.061
	(0.027)	(0.180)	(0.027)	(0.027)	(0.039)
	t = -0.934	t = -0.683	t = -0.896	t = -0.964	t = -1.582
	p = 0.351	p = 0.495	p = 0.371	p = 0.336	p = 0.114
zage	0.161***	1.053***	0.141***	0.159***	0.077
	(0.032)	(0.214)	(0.035)	(0.032)	(0.044)
	t = 4.985	t = 4.917	t = 4.044	t = 4.913	t = 1.747
	p = 0.00000	p = 0.00000	p = 0.0001	p = 0.00000	p = 0.081
zfemale	-0.094***	-0.631***	-0.104***	-0.096***	-0.119**
	(0.027)	(0.176)	(0.028)	(0.027)	(0.037)
	t = -3.529	t = -3.586	t = -3.688	t = -3.586	t = -3.239
	p = 0.0005	p = 0.0004	p = 0.0003	p = 0.0004	p = 0.002
zwhite	0.073*	0.395*	0.086*	0.073*	0.111*
	(0.032)	(0.176)	(0.034)	(0.032)	(0.050)
	t = 2.290	t = 2.240	t = 2.505	t = 2.282	t = 2.203
	p = 0.023	p = 0.026	p = 0.013	p = 0.023	p = 0.028
zcollege	0.061*	0.363*	0.055	0.061*	0.045
	(0.027)	(0.178)	(0.029)	(0.027)	(0.039)
	t = 2.237	t = 2.045	t = 1.885	t = 2.204	t = 1.142
	p = 0.026	p = 0.041	p = 0.060	p = 0.028	p = 0.254
zScreenTot	0.087**	0.474*	0.091**	0.085**	

	(0.031)	(0.192)	(0.033)	(0.031)	
	t = 2.823	t = 2.463	t = 2.727	t = 2.746	
	p = 0.005	p = 0.014	p = 0.007	p = 0.007	
Constant	1.825***	6.721***	1.840***	1.851***	1.994***
	(0.028)	(0.182)	(0.031)	(0.035)	(0.043)
	t = 65.050	t = 36.982	t = 60.044	t = 52.324	t = 45.966
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

Regression predicting flag count by issue polarization and standard controls, with HC2 robust SEs

Table S5. *Total flag count predicted by issue polarization and standard controls. Across specifications, we find that issue polarization is associated with greater flag count – such that more issue polarized participants flag more headlines.*

	Dependent variable:				
	Flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
zOutThermom	0.119***	0.719***	0.110***	0.097**	0.111*
	(0.029)	(0.169)	(0.031)	(0.032)	(0.044)
	t = 4.080	t = 4.249	t = 3.572	t = 3.059	t = 2.516
	p = 0.00005	p = 0.00003	p = 0.0004	p = 0.003	p = 0.012
I(zOutThermom2)				-0.050	
				(0.030)	
				t = -1.682	
				p = 0.093	
zPartisan	-0.057*	-0.358*	-0.056*	-0.059*	-0.092*
	(0.027)	(0.182)	(0.027)	(0.028)	(0.038)
	t = -2.092	t = -1.961	t = -2.070	t = -2.153	t = -2.423
	p = 0.037	p = 0.050	p = 0.039	p = 0.032	p = 0.016
zage	0.184***	1.221***	0.167***	0.182***	0.086
	(0.033)	(0.219)	(0.035)	(0.033)	(0.044)
	t = 5.562	t = 5.579	t = 4.702	t = 5.491	t = 1.944
	p = 0.00000	p = 0.00000	p = 0.00001	p = 0.00000	p = 0.052
zfemale	-0.097***	-0.650***	-0.108***	-0.102***	-0.118**
	(0.027)	(0.179)	(0.029)	(0.027)	(0.037)
	t = -3.637	t = -3.623	t = -3.771	t = -3.810	t = -3.192
	p = 0.0003	p = 0.0003	p = 0.0002	p = 0.0002	p = 0.002
zwhite	0.080*	0.430*	0.094**	0.080*	0.126*
	(0.033)	(0.179)	(0.035)	(0.033)	(0.051)
	t = 2.460	t = 2.401	t = 2.689	t = 2.445	t = 2.463
	p = 0.014	p = 0.017	p = 0.008	p = 0.015	p = 0.014
zcollege	0.069*	0.425*	0.062*	0.068*	0.050
	(0.028)	(0.179)	(0.030)	(0.028)	(0.040)
	t = 2.472	t = 2.366	t = 2.098	t = 2.439	t = 1.259
	p = 0.014	p = 0.018	p = 0.036	p = 0.015	p = 0.208

zScreenTot	0.128***	0.779***	0.132***	0.122***	
	(0.030)	(0.187)	(0.033)	(0.030)	
	t = 4.282	t = 4.174	t = 4.037	t = 4.064	
	p = 0.00002	p = 0.00003	p = 0.0001	p = 0.00005	
Constant	1.846***	6.764***	1.866***	1.895***	2.027***
	(0.028)	(0.185)	(0.031)	(0.039)	(0.042)
	t = 65.642	t = 36.633	t = 61.107	t = 48.563	t = 47.962
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

Regression predicting flag count by out-party dislike and standard controls, with HC2 robust SEs

Table S6. *Total flag count predicted by out-party dislike and standard controls. Across specifications, we find that out-party dislike is associated with greater flag count – such that participants who express greater dislike of counter-partisans flag more headlines.*

	Dependent variable:				
	Flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
zPK	0.197***	1.348***	0.196***	0.206***	0.147**
	(0.033)	(0.216)	(0.036)	(0.035)	(0.045)
	t = 5.937	t = 6.252	t = 5.393	t = 5.962	t = 3.242
	p = 0.000	p = 0.000	p = 0.00000	p = 0.000	p = 0.002
I(zPK2)				-0.034	
				(0.028)	
				t = -1.219	
				p = 0.223	
zIssuePol	0.148***	1.167***	0.147***	0.190***	0.094*
	(0.028)	(0.215)	(0.030)	(0.039)	(0.037)
	t = 5.269	t = 5.434	t = 4.985	t = 4.868	t = 2.536
	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.00001	p = 0.012
I(zIssuePol2)				-0.027	
				(0.023)	
				t = -1.159	
				p = 0.247	
zOutThermom	0.040	0.247	0.034	0.012	0.043
	(0.030)	(0.169)	(0.031)	(0.033)	(0.045)
	t = 1.354	t = 1.461	t = 1.073	t = 0.361	t = 0.958
	p = 0.176	p = 0.144	p = 0.284	p = 0.719	p = 0.338
I(zOutThermom2)				-0.063*	
				(0.030)	
				t = -2.138	
				p = 0.033	
zPartisan	-0.033	-0.198	-0.032	-0.036	-0.064
	(0.027)	(0.179)	(0.027)	(0.027)	(0.039)
	t = -1.229	t = -1.106	t = -1.198	t = -1.312	t = -1.648
	p = 0.220	p = 0.269	p = 0.232	p = 0.190	p = 0.100
zage	0.104**	0.671**	0.086*	0.099**	0.034
	(0.034)	(0.223)	(0.037)	(0.034)	(0.047)

	t = 3.032	t = 3.006	t = 2.342	t = 2.915	t = 0.723
	p = 0.003	p = 0.003	p = 0.020	p = 0.004	p = 0.470
zfemale	-0.077**	-0.560**	-0.089**	-0.089***	-0.101**
	(0.027)	(0.175)	(0.029)	(0.027)	(0.037)
	t = -2.908	t = -3.196	t = -3.116	t = -3.303	t = -2.733
	p = 0.004	p = 0.002	p = 0.002	p = 0.001	p = 0.007
zwhite	0.069*	0.389*	0.080*	0.069*	0.106*
	(0.032)	(0.175)	(0.034)	(0.032)	(0.051)
	t = 2.154	t = 2.219	t = 2.323	t = 2.150	t = 2.096
	p = 0.032	p = 0.027	p = 0.021	p = 0.032	p = 0.037
zcollege	0.050	0.288	0.045	0.047	0.041
	(0.027)	(0.176)	(0.029)	(0.027)	(0.039)
	t = 1.844	t = 1.637	t = 1.546	t = 1.750	t = 1.057
	p = 0.066	p = 0.102	p = 0.123	p = 0.081	p = 0.291
zScreenTot	0.046	0.198	0.046	0.033	
	(0.031)	(0.196)	(0.034)	(0.031)	
	t = 1.464	t = 1.012	t = 1.335	t = 1.054	
	p = 0.144	p = 0.312	p = 0.182	p = 0.292	
Constant	1.808***	6.694***	1.824***	1.927***	1.946***
	(0.028)	(0.179)	(0.031)	(0.050)	(0.047)
	t = 63.620	t = 37.313	t = 58.987	t = 38.871	t = 41.795
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

Regression predicting flag count by political knowledge, issue polarization, out-party dislike and standard controls, with HC2 robust SEs

Table S7. *Total flag count predicted by political knowledge, issue polarization, out-party dislike and standard controls. Across specifications, we find that political knowledge and issue polarization are associated with increased headline flagging; out-party dislike is not significantly associated with flag count in these models.*

	Dependent variable:	
	AnyFlag	FlagCount
	probit	*OLS*
	(1)	(2)
(Intercept)	0.492***	-20.101
	(0.027)	(11.721)
	t = 17.898	t = -1.715
	p = 0.000	p = 0.087
PC1	0.105***	4.716***
	(0.025)	(1.170)
	t = 4.220	t = 4.030
	p = 0.00003	p = 0.0001
PC2	0.001	-1.041***
	(0.031)	(0.239)
	t = 0.023	t = -4.351
	p = 0.982	p = 0.00002
PC3	0.064	1.552*
	(0.036)	(0.741)
	t = 1.752	t = 2.095
	p = 0.080	p = 0.037
zPartisan	-0.028	-0.942*
	(0.028)	(0.374)
	t = -0.983	t = -2.518
	p = 0.326	p = 0.012
zage	-0.218***	-4.260
	(0.033)	(2.459)
	t = -6.530	t = -1.732
	p = 0.000	p = 0.084
zfemale	-0.079**	-2.675**
	(0.028)	(0.919)
	t = -2.814	t = -2.911
	p = 0.005	p = 0.004
zwhite	0.077**	2.389**
	(0.029)	(0.882)
	t = 2.608	t = 2.707

	p = 0.010	p = 0.007
zcollege	0.004	0.588**
	(0.028)	(0.212)
	t = 0.148	t = 2.781
	p = 0.883	p = 0.006
zScreenTot	0.030	1.120**
	(0.031)	(0.409)
	t = 0.974	t = 2.740
	p = 0.330	p = 0.007
IMR1		57.577*
		(22.565)
		t = 2.552
		p = 0.011

Note: *p**p***p<0.001

2-step Heckman model predicting flag count (any flag; flag count) by political knowledge, issue polarization, out-party dislike, and standard controls

Table S8. *As preregistered, we also conducted a 2-step Heckman analysis to model a two-stage decision process (any flagging participation; flag count conditional on participation). We find that overall political motivation (PC1) predicts both flagging participation and flag count conditional on flagging any headline.*

	Dependent variable:	
	AnyFlag	FlagCount
	probit	*OLS*
	(1)	(2)
(Intercept)	0.492***	-20.101
	(0.027)	(11.721)
	t = 17.898	t = -1.715
	p = 0.000	p = 0.087
zPK	0.023	2.409***
	(0.033)	(0.364)
	t = 0.690	t = 6.618
	p = 0.491	p = 0.000
zIssuePol	0.114***	4.298***
	(0.032)	(1.270)
	t = 3.570	t = 3.385
	p = 0.0004	p = 0.001
zOutThermom	0.038	1.207*
	(0.029)	(0.480)
	t = 1.349	t = 2.517
	p = 0.178	p = 0.012
zPartisan	-0.028	-0.942*
	(0.028)	(0.374)
	t = -0.983	t = -2.518
	p = 0.326	p = 0.012
zage	-0.218***	-4.260
	(0.033)	(2.459)
	t = -6.530	t = -1.732
	p = 0.000	p = 0.084
zfemale	-0.079**	-2.675**
	(0.028)	(0.919)
	t = -2.814	t = -2.911
	p = 0.005	p = 0.004
zwhite	0.077**	2.389**
	(0.029)	(0.882)
	t = 2.608	t = 2.707

	p = 0.010	p = 0.007
zcollege	0.004	0.588**
	(0.028)	(0.212)
	t = 0.148	t = 2.781
	p = 0.883	p = 0.006
zScreenTot	0.030	1.120**
	(0.031)	(0.409)
	t = 0.974	t = 2.740
	p = 0.330	p = 0.007
IMR1		57.577*
		(22.565)
		t = 2.552
		p = 0.011

Note: *p**p***p<0.001

2-step Heckman model predicting flag count (any flag; flag count) by political knowledge, issue polarization, out-party dislike, and standard controls

Table S9. *As preregistered, we also conducted a 2-step Heckman analysis to model a two-stage decision process (any flagging participation; flag count conditional on participation). We find that political knowledge and out-party dislike predict flag count conditional on flagging any headline. We also find that issue polarization predicts both participation and conditional flag count.*

iii. Discordant Flagging

	Dependent variable:		
	Proportion discordant flags		
	OLS Model	Exclude True Independents	Attention + Practice Filter
	(1)	(2)	(3)
PC1	0.034***	0.036***	0.027***
	(0.003)	(0.003)	(0.005)
	t = 10.634	t = 10.815	t = 5.830
	p = 0.000	p = 0.000	p = 0.000
PC2	0.005	0.009	0.010
	(0.005)	(0.005)	(0.007)
	t = 1.032	t = 1.698	t = 1.396
	p = 0.303	p = 0.090	p = 0.163
PC3	0.009	0.008	0.010
	(0.005)	(0.005)	(0.007)
	t = 1.825	t = 1.596	t = 1.335
	p = 0.069	p = 0.111	p = 0.182
zPartisan	-0.011**	-0.010**	-0.012*
	(0.004)	(0.004)	(0.006)
	t = -2.661	t = -2.611	t = -2.155
	p = 0.008	p = 0.010	p = 0.032
zage	0.018***	0.018***	0.014*
	(0.005)	(0.005)	(0.006)
	t = 3.870	t = 4.044	t = 2.190
	p = 0.0002	p = 0.0001	p = 0.029
zfemale	0.002	0.003	-0.001
	(0.004)	(0.004)	(0.005)
	t = 0.402	t = 0.663	t = -0.152
	p = 0.688	p = 0.508	p = 0.879
zwhite	0.001	0.001	0.0001
	(0.005)	(0.005)	(0.007)
	t = 0.180	t = 0.282	t = 0.015
	p = 0.857	p = 0.778	p = 0.989
zcollege	0.010*	0.009*	0.007

	(0.004)	(0.004)	(0.006)
	t = 2.572	t = 2.248	t = 1.193
	p = 0.011	p = 0.025	p = 0.233
zScreenTot	0.005	0.004	
	(0.005)	(0.005)	
	t = 1.154	t = 0.716	
	p = 0.249	p = 0.475	
Constant	0.468***	0.471***	0.488***
	(0.004)	(0.005)	(0.007)
	t = 110.833	t = 103.814	t = 68.817
	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant flags by the first three PCs of political knowledge, issue polarization, and out-party dislike, and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S10. *Proportion discordant flags predicted by political motivation principal components. Across specifications, we find that overall political motivation (PC1) is associated with flagging an increased proportion of politically discordant headlines.*

	Dependent variable:			
	Proportion discordant flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zPK	0.027***	0.028***	0.028***	0.018**
	(0.004)	(0.005)	(0.005)	(0.006)
	t = 6.168	t = 6.111	t = 5.737	t = 2.913
	p = 0.000	p = 0.000	p = 0.000	p = 0.004
I(zPK2)			-0.002	
			(0.004)	
			t = -0.490	
			p = 0.625	
zPartisan	-0.015***	-0.015***	-0.015***	-0.019**
	(0.004)	(0.004)	(0.004)	(0.006)
	t = -3.778	t = -3.685	t = -3.780	t = -3.247
	p = 0.0002	p = 0.0003	p = 0.0002	p = 0.002
zage	0.019***	0.019***	0.019***	0.010
	(0.005)	(0.005)	(0.005)	(0.006)
	t = 4.087	t = 3.902	t = 4.080	t = 1.603
	p = 0.00005	p = 0.0001	p = 0.00005	p = 0.109
zfemale	0.005	0.007	0.005	0.002
	(0.004)	(0.004)	(0.004)	(0.006)
	t = 1.174	t = 1.627	t = 1.139	t = 0.325
	p = 0.241	p = 0.104	p = 0.255	p = 0.746
zwhite	0.002	0.004	0.002	0.003
	(0.005)	(0.005)	(0.005)	(0.007)
	t = 0.482	t = 0.760	t = 0.497	t = 0.442
	p = 0.630	p = 0.448	p = 0.620	p = 0.659
zcollege	0.010*	0.009*	0.010*	0.004
	(0.004)	(0.004)	(0.004)	(0.006)
	t = 2.412	t = 2.160	t = 2.406	t = 0.718
	p = 0.016	p = 0.031	p = 0.017	p = 0.473
zScreenTot	0.009*	0.008	0.009*	
	(0.005)	(0.005)	(0.005)	

	t = 2.005	t = 1.523	t = 1.976	
	p = 0.045	p = 0.128	p = 0.049	
Constant	0.473***	0.477***	0.475***	0.497***
	(0.004)	(0.005)	(0.006)	(0.007)
	t = 110.196	t = 102.955	t = 76.905	t = 69.460
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant flags by political knowledge and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S11. *Proportion discordant flags predicted by political knowledge. Across specifications, we find that political knowledge is associated with flagging an increased proportion of politically discordant headlines.*

	Dependent variable:			
	Proportion discordant flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zIssuePol	0.037***	0.038***	0.035***	0.030***
	(0.004)	(0.004)	(0.005)	(0.005)
	t = 9.318	t = 9.549	t = 6.342	t = 6.231
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
I(zIssuePol2)			0.002	
			(0.003)	
			t = 0.539	
			p = 0.590	
zPartisan	-0.010*	-0.009*	-0.010*	-0.012*
	(0.004)	(0.004)	(0.004)	(0.006)
	t = -2.372	t = -2.254	t = -2.369	t = -2.040
	p = 0.018	p = 0.025	p = 0.018	p = 0.042
zage	0.022***	0.023***	0.022***	0.015**
	(0.004)	(0.005)	(0.005)	(0.006)
	t = 4.943	t = 4.968	t = 4.896	t = 2.579
	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.010
zfemale	0.001	0.003	0.001	-0.001
	(0.004)	(0.004)	(0.004)	(0.005)
	t = 0.331	t = 0.781	t = 0.355	t = -0.214
	p = 0.741	p = 0.435	p = 0.723	p = 0.831
zwhite	0.001	0.002	0.001	0.0004
	(0.005)	(0.005)	(0.005)	(0.007)
	t = 0.291	t = 0.414	t = 0.280	t = 0.053
	p = 0.772	p = 0.679	p = 0.780	p = 0.958
zcollege	0.011**	0.010*	0.011**	0.006
	(0.004)	(0.004)	(0.004)	(0.006)
	t = 2.697	t = 2.354	t = 2.702	t = 1.118
	p = 0.007	p = 0.019	p = 0.007	p = 0.264
zScreenTot	0.008	0.007	0.008	
	(0.005)	(0.005)	(0.005)	

	t = 1.863	t = 1.541	t = 1.873	
	p = 0.063	p = 0.124	p = 0.062	
Constant	0.470***	0.473***	0.469***	0.491***
	(0.004)	(0.004)	(0.005)	(0.007)
	t = 113.271	t = 105.468	t = 91.708	t = 73.990
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant flags by issue polarization and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S12. *Proportion discordant flags predicted by issue polarization. Across specifications, we find that issue polarization is associated with flagging an increased proportion of politically discordant headlines.*

	Dependent variable:			
	Proportion discordant flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zOutThermom	0.032***	0.036***	0.034***	0.033***
	(0.004)	(0.005)	(0.005)	(0.007)
	t = 7.348	t = 7.901	t = 6.933	t = 4.799
	p = 0.000	p = 0.000	p = 0.000	p = 0.00001
I(zOutThermom2)			0.005	
			(0.004)	
			t = 1.294	
			p = 0.196	
zPartisan	-0.016***	-0.016***	-0.016***	-0.018**
	(0.004)	(0.004)	(0.004)	(0.006)
	t = -4.026	t = -3.967	t = -3.968	t = -3.241
	p = 0.0001	p = 0.0001	p = 0.0001	p = 0.002
zage	0.024***	0.024***	0.024***	0.015*
	(0.004)	(0.005)	(0.004)	(0.006)
	t = 5.439	t = 5.308	t = 5.458	t = 2.523
	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.012
zfemale	-0.001	0.001	-0.0001	-0.002
	(0.004)	(0.004)	(0.004)	(0.006)
	t = -0.127	t = 0.192	t = -0.023	t = -0.354
	p = 0.899	p = 0.848	p = 0.982	p = 0.724
zwhite	0.003	0.004	0.003	0.002
	(0.005)	(0.005)	(0.005)	(0.007)
	t = 0.597	t = 0.806	t = 0.610	t = 0.277
	p = 0.551	p = 0.421	p = 0.542	p = 0.782
zcollege	0.011**	0.010*	0.011**	0.006
	(0.004)	(0.004)	(0.004)	(0.006)
	t = 2.707	t = 2.317	t = 2.786	t = 0.986
	p = 0.007	p = 0.021	p = 0.006	p = 0.325
zScreenTot	0.012**	0.010*	0.013**	
	(0.004)	(0.005)	(0.004)	

	t = 2.769	t = 2.193	t = 2.884	
	p = 0.006	p = 0.029	p = 0.004	
Constant	0.476***	0.479***	0.471***	0.498***
	(0.004)	(0.004)	(0.005)	(0.006)
	t = 116.981	t = 110.215	t = 87.558	t = 79.688
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant flags by out-party dislike and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S13. *Proportion discordant flags predicted by out-party dislike. Across specifications, we find that out-party dislike is associated with flagging an increased proportion of politically discordant headlines.*

	Dependent variable:			
	Proportion discordant flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zPK	0.012**	0.011*	0.013**	0.004
	(0.004)	(0.005)	(0.005)	(0.006)
	t = 2.626	t = 2.429	t = 2.642	t = 0.646
	p = 0.009	p = 0.016	p = 0.009	p = 0.519
I(zPK2)			-0.003	
			(0.004)	
			t = -0.591	
			p = 0.555	
zIssuePol	0.028***	0.028***	0.026***	0.023***
	(0.004)	(0.004)	(0.006)	(0.006)
	t = 6.479	t = 6.410	t = 4.484	t = 4.014
	p = 0.000	p = 0.000	p = 0.00001	p = 0.0001
I(zIssuePol2)			0.002	
			(0.003)	
			t = 0.492	
			p = 0.623	
zOutThermom	0.020***	0.024***	0.020***	0.020**
	(0.005)	(0.005)	(0.005)	(0.008)
	t = 4.274	t = 4.903	t = 4.002	t = 2.650
	p = 0.00002	p = 0.00000	p = 0.0001	p = 0.009
I(zOutThermom2)			0.002	
			(0.004)	
			t = 0.511	
			p = 0.610	
zPartisan	-0.011**	-0.010**	-0.011**	-0.012*
	(0.004)	(0.004)	(0.004)	(0.006)
	t = -2.661	t = -2.611	t = -2.648	t = -2.155
	p = 0.008	p = 0.010	p = 0.009	p = 0.032
zage	0.018***	0.018***	0.018***	0.014*
	(0.005)	(0.005)	(0.005)	(0.006)

	t = 3.870	t = 4.044	t = 3.836	t = 2.190
	p = 0.0002	p = 0.0001	p = 0.0002	p = 0.029
zfemale	0.002	0.003	0.002	-0.001
	(0.004)	(0.004)	(0.004)	(0.005)
	t = 0.402	t = 0.663	t = 0.433	t = -0.152
	p = 0.688	p = 0.508	p = 0.665	p = 0.879
zwhite	0.001	0.001	0.001	0.0001
	(0.005)	(0.005)	(0.005)	(0.007)
	t = 0.180	t = 0.282	t = 0.194	t = 0.015
	p = 0.857	p = 0.778	p = 0.847	p = 0.989
zcollege	0.010*	0.009*	0.010**	0.007
	(0.004)	(0.004)	(0.004)	(0.006)
	t = 2.572	t = 2.248	t = 2.580	t = 1.193
	p = 0.011	p = 0.025	p = 0.010	p = 0.233
zScreenTot	0.005	0.004	0.005	
	(0.005)	(0.005)	(0.005)	
	t = 1.154	t = 0.716	t = 1.172	
	p = 0.249	p = 0.475	p = 0.242	
Constant	0.468***	0.471***	0.467***	0.488***
	(0.004)	(0.005)	(0.007)	(0.007)
	t = 110.833	t = 103.814	t = 63.887	t = 68.817
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant flags by political knowledge, issue polarization, out-party dislike, and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S14. *Proportion discordant flags predicted by political knowledge, issue polarization, and out-party dislike. Across specifications, we find that each of these measures are associated with flagging an increased proportion of politically discordant headlines (political knowledge no longer a significant predictor when filtering for attention and practice task accuracy).*

iv. Flagging Quality

	Dependent variable:		
	Proportion false-positive flags		
	OLS Model	Exclude True Independents	Attention + Practice Filter
	(1)	(2)	(3)
PC1	-0.054***	-0.058***	-0.034***
	(0.005)	(0.005)	(0.007)
	t = -10.673	t = -10.972	t = -4.862
	p = 0.000	p = 0.000	p = 0.00001
PC2	0.030***	0.032***	0.041***
	(0.007)	(0.007)	(0.010)
	t = 4.411	t = 4.389	t = 3.944
	p = 0.00002	p = 0.00002	p = 0.0001
PC3	0.021**	0.022*	0.030*
	(0.008)	(0.008)	(0.012)
	t = 2.599	t = 2.551	t = 2.552
	p = 0.010	p = 0.011	p = 0.011
zPartisan	0.024***	0.024***	0.034***
	(0.006)	(0.006)	(0.008)
	t = 4.070	t = 4.090	t = 4.079
	p = 0.00005	p = 0.00005	p = 0.00005
zage	-0.014*	-0.014*	0.003
	(0.006)	(0.007)	(0.009)
	t = -2.114	t = -2.069	t = 0.285
	p = 0.035	p = 0.039	p = 0.776
zfemale	-0.022***	-0.024***	-0.015
	(0.006)	(0.006)	(0.008)
	t = -3.676	t = -3.784	t = -1.903
	p = 0.0003	p = 0.0002	p = 0.058
zwhite	-0.006	-0.006	-0.007
	(0.007)	(0.007)	(0.010)
	t = -0.884	t = -0.922	t = -0.735
	p = 0.377	p = 0.357	p = 0.463
zcollege	-0.016**	-0.011	-0.013

	(0.006)	(0.006)	(0.008)
	t = -2.800	t = -1.835	t = -1.618
	p = 0.006	p = 0.067	p = 0.106
zScreenTot	-0.053***	-0.048***	
	(0.007)	(0.007)	
	t = -8.025	t = -6.994	
	p = 0.000	p = 0.000	
Constant	0.326***	0.330***	0.247***
	(0.006)	(0.006)	(0.011)
	t = 53.455	t = 51.136	t = 23.208
	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

OLS regression predicting proportion false-positive flags by the first three PCs of political knowledge, issue polarization, and out-party dislike, and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S15. *Proportion false-positive (i.e., incorrect) flags predicted by political motivation principal components. Across specifications, we find that overall political motivation (PC1) is associated with flagging a decreased proportion of non-false headlines.*

	Dependent variable:			
	Proportion false-positive flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zPK	-0.072***	-0.077***	-0.074***	-0.061***
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -10.759	t = -10.947	t = -10.245	t = -6.378
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
I(zPK2)			0.004	
			(0.006)	
			t = 0.672	
			p = 0.502	
zPartisan	0.028***	0.028***	0.028***	0.035***
	(0.006)	(0.006)	(0.006)	(0.008)
	t = 4.848	t = 4.924	t = 4.851	t = 4.299
	p = 0.00001	p = 0.00000	p = 0.00001	p = 0.00002
zage	-0.015*	-0.014*	-0.014*	0.003
	(0.007)	(0.007)	(0.007)	(0.009)
	t = -2.208	t = -2.017	t = -2.191	t = 0.326
	p = 0.028	p = 0.044	p = 0.029	p = 0.745
zfemale	-0.024***	-0.026***	-0.024***	-0.015
	(0.006)	(0.006)	(0.006)	(0.008)
	t = -3.963	t = -4.121	t = -3.901	t = -1.898
	p = 0.0001	p = 0.00004	p = 0.0001	p = 0.058
zwhite	-0.007	-0.008	-0.007	-0.007
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -1.033	t = -1.149	t = -1.052	t = -0.749
	p = 0.302	p = 0.251	p = 0.293	p = 0.454
zcollege	-0.016**	-0.011	-0.016**	-0.013
	(0.006)	(0.006)	(0.006)	(0.008)
	t = -2.753	t = -1.821	t = -2.745	t = -1.616
	p = 0.006	p = 0.069	p = 0.007	p = 0.107
zScreenTot	-0.056***	-0.051***	-0.056***	
	(0.007)	(0.007)	(0.007)	
	t = -8.509	t = -7.425	t = -8.420	

	p = 0.000	p = 0.000	p = 0.000	
Constant	0.322***	0.325***	0.318***	0.246***
	(0.006)	(0.006)	(0.008)	(0.010)
	t = 52.920	t = 50.260	t = 37.492	t = 23.877
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

OLS regression predicting proportion false-positive flags by political knowledge and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S16. *Proportion false-positive (i.e., incorrect) flags predicted by political knowledge. Across specifications, we find that overall political knowledge is associated with flagging a decreased proportion of non-false headlines.*

	Dependent variable:			
	Proportion false-positive flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zIssuePol	-0.043***	-0.047***	-0.046***	-0.018*
	(0.006)	(0.006)	(0.009)	(0.008)
	t = -6.863	t = -7.343	t = -5.333	t = -2.325
	p = 0.000	p = 0.000	p = 0.00000	p = 0.021
I(zIssuePol2)			0.003	
			(0.006)	
			t = 0.441	
			p = 0.660	
zPartisan	0.023***	0.022***	0.023***	0.035***
	(0.006)	(0.006)	(0.006)	(0.009)
	t = 3.565	t = 3.517	t = 3.568	t = 3.851
	p = 0.0004	p = 0.0005	p = 0.0004	p = 0.0002
zage	-0.031***	-0.031***	-0.030***	-0.017
	(0.007)	(0.007)	(0.007)	(0.009)
	t = -4.690	t = -4.594	t = -4.637	t = -1.872
	p = 0.00001	p = 0.00001	p = 0.00001	p = 0.062
zfemale	-0.015*	-0.017*	-0.015*	-0.005
	(0.006)	(0.006)	(0.006)	(0.008)
	t = -2.389	t = -2.555	t = -2.363	t = -0.546
	p = 0.017	p = 0.011	p = 0.019	p = 0.586
zwhite	-0.008	-0.008	-0.008	-0.007
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -1.124	t = -1.165	t = -1.133	t = -0.673
	p = 0.261	p = 0.245	p = 0.258	p = 0.501
zcollege	-0.019**	-0.013*	-0.019**	-0.014
	(0.006)	(0.006)	(0.006)	(0.009)
	t = -3.144	t = -2.069	t = -3.135	t = -1.643
	p = 0.002	p = 0.039	p = 0.002	p = 0.101
zScreenTot	-0.065***	-0.062***	-0.065***	
	(0.007)	(0.007)	(0.007)	
	t = -9.618	t = -8.867	t = -9.610	

	p = 0.000	p = 0.000	p = 0.000	
Constant	0.316***	0.321***	0.314***	0.226***
	(0.006)	(0.007)	(0.008)	(0.010)
	t = 51.034	t = 48.457	t = 39.898	t = 22.306
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

OLS regression predicting proportion false-positive flags by issue polarization and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S17. *Proportion false-positive (i.e., incorrect) flags predicted by issue polarization. Across specifications, we find issue polarization is associated with flagging a decreased proportion of non-false headlines.*

	Dependent variable:			
	Proportion false-positive flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zOutThermom	-0.029***	-0.032***	-0.016*	-0.005
	(0.007)	(0.008)	(0.007)	(0.010)
	t = -4.080	t = -4.187	t = -2.259	t = -0.528
	p = 0.00005	p = 0.00003	p = 0.024	p = 0.598
I(zOutThermom2)			0.030***	
			(0.006)	
			t = 4.817	
			p = 0.00001	
zPartisan	0.030***	0.030***	0.031***	0.040***
	(0.006)	(0.006)	(0.006)	(0.009)
	t = 4.867	t = 4.966	t = 5.191	t = 4.609
	p = 0.00001	p = 0.00000	p = 0.00000	p = 0.00001
zage	-0.034***	-0.034***	-0.033***	-0.017
	(0.007)	(0.007)	(0.007)	(0.009)
	t = -5.009	t = -4.777	t = -4.919	t = -1.867
	p = 0.00000	p = 0.00001	p = 0.00000	p = 0.062
zfemale	-0.013*	-0.015*	-0.011	-0.004
	(0.006)	(0.007)	(0.006)	(0.008)
	t = -2.128	t = -2.228	t = -1.748	t = -0.522
	p = 0.034	p = 0.026	p = 0.081	p = 0.602
zwhite	-0.010	-0.011	-0.009	-0.008
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -1.406	t = -1.546	t = -1.364	t = -0.839
	p = 0.160	p = 0.123	p = 0.173	p = 0.402
zcollege	-0.019**	-0.013*	-0.017**	-0.013
	(0.006)	(0.007)	(0.006)	(0.009)
	t = -3.129	t = -2.032	t = -2.787	t = -1.515
	p = 0.002	p = 0.043	p = 0.006	p = 0.130
zScreenTot	-0.070***	-0.067***	-0.067***	
	(0.007)	(0.007)	(0.007)	
	t = -10.532	t = -9.583	t = -10.114	

	p = 0.000	p = 0.000	p = 0.000	
Constant	0.308***	0.311***	0.280***	0.218***
	(0.006)	(0.006)	(0.008)	(0.009)
	t = 50.618	t = 47.913	t = 34.178	t = 23.824
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p**p***p<0.001

OLS regression predicting proportion false-positive flags by out-party dislike and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S18. *Proportion false-positive (i.e., incorrect) flags predicted by out-party dislike. Across specifications, we largely find out-party dislike is associated with flagging a decreased proportion of non-false headlines (no longer a significant predictor when filtering for attention and practice task accuracy).*

	Dependent variable:			
	Proportion false-positive flags			
	OLS Model	Exclude True Independents	With Quadratic	Attention + Practice Filter
	(1)	(2)	(3)	(4)
zPK	-0.061***	-0.065***	-0.059***	-0.060***
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -8.744	t = -8.792	t = -8.168	t = -5.897
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
I(zPK2)			0.003	
			(0.006)	
			t = 0.528	
			p = 0.598	
zIssuePol	-0.023***	-0.025***	-0.027**	-0.005
	(0.007)	(0.007)	(0.008)	(0.009)
	t = -3.466	t = -3.671	t = -3.248	t = -0.543
	p = 0.001	p = 0.0003	p = 0.002	p = 0.588
I(zIssuePol2)			-0.0003	
			(0.005)	
			t = -0.052	
			p = 0.959	
zOutThermom	-0.009	-0.010	0.005	0.008
	(0.007)	(0.007)	(0.007)	(0.011)
	t = -1.331	t = -1.338	t = 0.663	t = 0.758
	p = 0.184	p = 0.181	p = 0.508	p = 0.449
I(zOutThermom2)			0.031***	
			(0.006)	
			t = 5.134	
			p = 0.00000	
zPartisan	0.024***	0.024***	0.024***	0.034***
	(0.006)	(0.006)	(0.006)	(0.008)
	t = 4.070	t = 4.090	t = 4.265	t = 4.079
	p = 0.00005	p = 0.00005	p = 0.00003	p = 0.00005
zage	-0.014*	-0.014*	-0.013*	0.003
	(0.006)	(0.007)	(0.006)	(0.009)
	t = -2.114	t = -2.069	t = -2.078	t = 0.285

	p = 0.035	p = 0.039	p = 0.038	p = 0.776
zfemale	-0.022***	-0.024***	-0.019**	-0.015
	(0.006)	(0.006)	(0.006)	(0.008)
	t = -3.676	t = -3.784	t = -3.204	t = -1.903
	p = 0.0003	p = 0.0002	p = 0.002	p = 0.058
zwhite	-0.006	-0.006	-0.006	-0.007
	(0.007)	(0.007)	(0.007)	(0.010)
	t = -0.884	t = -0.922	t = -0.832	t = -0.735
	p = 0.377	p = 0.357	p = 0.406	p = 0.463
zcollege	-0.016**	-0.011	-0.014*	-0.013
	(0.006)	(0.006)	(0.006)	(0.008)
	t = -2.800	t = -1.835	t = -2.411	t = -1.618
	p = 0.006	p = 0.067	p = 0.016	p = 0.106
zScreenTot	-0.053***	-0.048***	-0.049***	
	(0.007)	(0.007)	(0.007)	
	t = -8.025	t = -6.994	t = -7.587	
	p = 0.000	p = 0.000	p = 0.000	
Constant	0.326***	0.330***	0.294***	0.247***
	(0.006)	(0.006)	(0.010)	(0.011)
	t = 53.455	t = 51.136	t = 30.104	t = 23.208
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion false-positive flags by political knowledge, issue polarization, out-party dislike, and standard controls; with analytic weighting by flag count and HC2 robust SEs

Table S19. *Proportion false-positive (i.e., incorrect) flags predicted by political knowledge, issue polarization, and out-party dislike. Across specifications, we largely find that political knowledge and issue polarization are associated with flagging a decreased proportion of non-false headlines. In these models, we do not find that out-party dislike is a significant predictor (though the quadratic term is, indicating a significant nonlinear relationship between out-party dislike and false-positive flagging – i.e., out-party dislike may be associated with higher quality flagging up until a point, as our formal model also predicts).*

v. Flag Count by Headline Type

	Dependent variable:			
	Item-type flag count			
	QP Model	OLS Model	QP Exclude True Independents	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)
ItemTypeDT	-1.324***	-2.013***	-1.290***	-1.777***
	(0.040)	(0.065)	(0.043)	(0.081)
	t = -32.962	t = -30.825	t = -29.814	t = -22.048
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.156***	-0.479***	-0.163***	-0.181***
	(0.018)	(0.040)	(0.020)	(0.028)
	t = -8.420	t = -11.881	t = -8.130	t = -6.512
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT	-1.495***	-2.146***	-1.502***	-1.935***
	(0.039)	(0.067)	(0.042)	(0.073)
	t = -37.868	t = -31.924	t = -35.423	t = -26.491
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.099***	-1.912***	-1.082***	-1.398***
	(0.038)	(0.067)	(0.040)	(0.067)
	t = -28.771	t = -28.659	t = -26.741	t = -20.859
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
PC1	0.327***	0.941***	0.330***	0.231***
	(0.024)	(0.072)	(0.026)	(0.033)
	t = 13.672	t = 13.060	t = 12.906	t = 6.967
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
PC2	-0.113***	-0.363***	-0.115**	-0.072
	(0.034)	(0.075)	(0.036)	(0.048)
	t = -3.368	t = -4.869	t = -3.192	t = -1.505
	p = 0.001	p = 0.00001	p = 0.002	p = 0.133
PC3	-0.020	0.060	-0.029	-0.037
	(0.035)	(0.097)	(0.038)	(0.048)
	t = -0.571	t = 0.619	t = -0.755	t = -0.777

	p = 0.568	p = 0.536	p = 0.451	p = 0.438
zPartisan	-0.121***	-0.317***	-0.121***	-0.157***
	(0.027)	(0.073)	(0.027)	(0.038)
	t = -4.424	t = -4.350	t = -4.454	t = -4.168
	p = 0.00001	p = 0.00002	p = 0.00001	p = 0.00004
zage	0.135***	0.324***	0.116**	0.051
	(0.036)	(0.091)	(0.039)	(0.048)
	t = 3.732	t = 3.554	t = 2.997	t = 1.055
	p = 0.0002	p = 0.0004	p = 0.003	p = 0.292
zfemale	-0.056*	-0.169*	-0.063*	-0.078*
	(0.028)	(0.072)	(0.030)	(0.038)
	t = -2.001	t = -2.329	t = -2.106	t = -2.077
	p = 0.046	p = 0.020	p = 0.036	p = 0.038
zwhite	0.085*	0.190**	0.106**	0.112*
	(0.034)	(0.070)	(0.037)	(0.051)
	t = 2.497	t = 2.700	t = 2.866	t = 2.172
	p = 0.013	p = 0.007	p = 0.005	p = 0.030
zcollege	0.076**	0.148*	0.066*	0.053
	(0.029)	(0.073)	(0.031)	(0.040)
	t = 2.646	t = 2.041	t = 2.176	t = 1.348
	p = 0.009	p = 0.042	p = 0.030	p = 0.178
zScreenTot	0.107**	0.191*	0.102**	
	(0.033)	(0.079)	(0.036)	
	t = 3.253	t = 2.432	t = 2.882	
	p = 0.002	p = 0.016	p = 0.004	
ItemTypeDT:PC1	-0.168***	-0.831***	-0.181***	-0.081
	(0.034)	(0.066)	(0.036)	(0.051)
	t = -4.982	t = -12.612	t = -5.099	t = -1.583
	p = 0.00000	p = 0.000	p = 0.00000	p = 0.114
ItemTypeCF:PC1	-0.062***	-0.334***	-0.063***	-0.051**
	(0.014)	(0.038)	(0.015)	(0.018)
	t = -4.563	t = -8.859	t = -4.345	t = -2.920
	p = 0.00001	p = 0.000	p = 0.00002	p = 0.004
ItemTypeCT:PC1	-0.339***	-0.948***	-0.367***	-0.339***

	(0.039)	(0.068)	(0.042)	(0.058)
	t = -8.742	t = -14.000	t = -8.821	t = -5.790
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens:PC1	-0.345***	-0.954***	-0.371***	-0.275***
	(0.034)	(0.067)	(0.036)	(0.052)
	t = -10.245	t = -14.158	t = -10.422	t = -5.315
	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeDT:PC2	0.131**	0.370***	0.148**	0.233**
	(0.044)	(0.066)	(0.045)	(0.081)
	t = 3.006	t = 5.603	t = 3.282	t = 2.893
	p = 0.003	p = 0.00000	p = 0.002	p = 0.004
ItemTypeCF:PC2	-0.054**	-0.020	-0.068**	-0.086**
	(0.020)	(0.042)	(0.021)	(0.029)
	t = -2.679	t = -0.485	t = -3.280	t = -2.981
	p = 0.008	p = 0.628	p = 0.002	p = 0.003
ItemTypeCT:PC2	0.131**	0.371***	0.136**	0.163*
	(0.047)	(0.067)	(0.050)	(0.081)
	t = 2.774	t = 5.569	t = 2.741	t = 2.011
	p = 0.006	p = 0.00000	p = 0.007	p = 0.045
ItemTypeSens:PC2	0.106*	0.357***	0.099*	0.175*
	(0.043)	(0.066)	(0.044)	(0.075)
	t = 2.492	t = 5.396	t = 2.231	t = 2.340
	p = 0.013	p = 0.00000	p = 0.026	p = 0.020
ItemTypeDT:PC3	0.090	-0.005	0.125*	0.194*
	(0.049)	(0.085)	(0.052)	(0.082)
	t = 1.842	t = -0.057	t = 2.406	t = 2.376
	p = 0.066	p = 0.955	p = 0.017	p = 0.018
ItemTypeCF:PC3	-0.066**	-0.198***	-0.049*	-0.060*
	(0.020)	(0.052)	(0.021)	(0.028)
	t = -3.282	t = -3.791	t = -2.285	t = -2.161
	p = 0.002	p = 0.0002	p = 0.023	p = 0.031
ItemTypeCT:PC3	0.088	-0.025	0.083	0.168*
	(0.057)	(0.089)	(0.060)	(0.079)
	t = 1.555	t = -0.283	t = 1.379	t = 2.134

	p = 0.120	p = 0.777	p = 0.168	p = 0.033
ItemTypeSens:PC3	0.048	-0.039	0.052	0.065
	(0.053)	(0.088)	(0.055)	(0.083)
	t = 0.906	t = -0.446	t = 0.939	t = 0.785
	p = 0.365	p = 0.656	p = 0.348	p = 0.433
ItemTypeDT:zPartisan	0.319***	0.448***	0.316***	0.479***
	(0.040)	(0.064)	(0.039)	(0.069)
	t = 7.968	t = 6.968	t = 8.193	t = 6.957
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF:zPartisan	0.110***	0.294***	0.112***	0.109***
	(0.016)	(0.041)	(0.016)	(0.021)
	t = 6.907	t = 7.261	t = 7.030	t = 5.168
	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeCT:zPartisan	0.097*	0.305***	0.102*	0.154*
	(0.042)	(0.066)	(0.042)	(0.072)
	t = 2.309	t = 4.596	t = 2.432	t = 2.131
	p = 0.021	p = 0.00001	p = 0.016	p = 0.034
ItemTypeSens:zPartisan	0.151***	0.339***	0.153***	0.152*
	(0.040)	(0.066)	(0.039)	(0.065)
	t = 3.796	t = 5.119	t = 3.910	t = 2.343
	p = 0.0002	p = 0.00000	p = 0.0001	p = 0.020
ItemTypeDT:zage	0.017	-0.227**	0.021	0.037
	(0.048)	(0.080)	(0.051)	(0.073)
	t = 0.354	t = -2.823	t = 0.416	t = 0.501
	p = 0.724	p = 0.005	p = 0.678	p = 0.617
ItemTypeCF:zage	-0.040*	-0.124**	-0.038	-0.043
	(0.019)	(0.046)	(0.020)	(0.024)
	t = -2.101	t = -2.719	t = -1.895	t = -1.771
	p = 0.036	p = 0.007	p = 0.059	p = 0.077
ItemTypeCT:zage	-0.038	-0.276***	-0.062	0.005
	(0.052)	(0.082)	(0.056)	(0.082)
	t = -0.724	t = -3.372	t = -1.109	t = 0.065
	p = 0.470	p = 0.001	p = 0.268	p = 0.949
ItemTypeSens:zage	-0.131**	-0.321***	-0.115*	-0.054
	(0.047)	(0.081)	(0.050)	(0.071)

	t = -2.779	t = -3.947	t = -2.296	t = -0.757
	p = 0.006	p = 0.0001	p = 0.022	p = 0.449
ItemTypeDT:zfemale	-0.117**	0.053	-0.130**	-0.177**
	(0.039)	(0.065)	(0.041)	(0.064)
	t = -3.000	t = 0.817	t = -3.147	t = -2.769
	p = 0.003	p = 0.414	p = 0.002	p = 0.006
ItemTypeCF:zfemale	0.007	0.041	0.003	-0.014
	(0.016)	(0.039)	(0.017)	(0.021)
	t = 0.458	t = 1.052	t = 0.206	t = -0.662
	p = 0.648	p = 0.293	p = 0.837	p = 0.508
ItemTypeCT:zfemale	-0.073	0.104	-0.091*	-0.093
	(0.042)	(0.067)	(0.045)	(0.065)
	t = -1.753	t = 1.551	t = -2.021	t = -1.429
	p = 0.080	p = 0.121	p = 0.044	p = 0.153
ItemTypeSens:zfemale	-0.056	0.086	-0.062	-0.006
	(0.039)	(0.066)	(0.042)	(0.063)
	t = -1.425	t = 1.299	t = -1.478	t = -0.096
	p = 0.155	p = 0.195	p = 0.140	p = 0.924
ItemTypeDT:zwhite	-0.047	-0.172**	-0.070	-0.024
	(0.045)	(0.064)	(0.048)	(0.082)
	t = -1.041	t = -2.701	t = -1.454	t = -0.294
	p = 0.298	p = 0.007	p = 0.147	p = 0.769
ItemTypeCF:zwhite	0.0004	-0.043	-0.012	-0.001
	(0.020)	(0.038)	(0.021)	(0.029)
	t = 0.019	t = -1.140	t = -0.585	t = -0.049
	p = 0.985	p = 0.255	p = 0.559	p = 0.962
ItemTypeCT:zwhite	-0.076	-0.185**	-0.094	0.056
	(0.046)	(0.064)	(0.050)	(0.085)
	t = -1.649	t = -2.893	t = -1.897	t = 0.658
	p = 0.100	p = 0.004	p = 0.058	p = 0.511
ItemTypeSens:zwhite	-0.044	-0.160*	-0.059	-0.064
	(0.042)	(0.063)	(0.045)	(0.075)
	t = -1.046	t = -2.535	t = -1.325	t = -0.854
	p = 0.296	p = 0.012	p = 0.186	p = 0.394
ItemTypeDT:zcollege	-0.039	-0.123	-0.018	-0.066
	(0.040)	(0.065)	(0.043)	(0.064)

	t = -0.985	t = -1.899	t = -0.429	t = -1.032
	p = 0.325	p = 0.058	p = 0.669	p = 0.302
ItemTypeCF:zcollege	-0.017	-0.040	-0.025	-0.001
	(0.017)	(0.039)	(0.018)	(0.022)
	t = -0.987	t = -1.025	t = -1.376	t = -0.061
	p = 0.324	p = 0.306	p = 0.169	p = 0.952
ItemTypeCT:zcollege	-0.104*	-0.162*	-0.074	-0.065
	(0.042)	(0.066)	(0.045)	(0.071)
	t = -2.463	t = -2.457	t = -1.647	t = -0.919
	p = 0.014	p = 0.014	p = 0.100	p = 0.359
ItemTypeSens:zcollege	-0.048	-0.128	-0.019	-0.042
	(0.039)	(0.067)	(0.042)	(0.064)
	t = -1.220	t = -1.914	t = -0.450	t = -0.656
	p = 0.223	p = 0.056	p = 0.653	p = 0.513
ItemTypeDT:zScreenTot	-0.280***	-0.305***	-0.263***	
	(0.046)	(0.071)	(0.049)	
	t = -6.126	t = -4.302	t = -5.397	
	p = 0.000	p = 0.00002	p = 0.00000	
ItemTypeCF:zScreenTot	0.016	0.016	0.015	
	(0.019)	(0.044)	(0.021)	
	t = 0.809	t = 0.358	t = 0.705	
	p = 0.419	p = 0.721	p = 0.481	
ItemTypeCT:zScreenTot	-0.244***	-0.261***	-0.206***	
	(0.050)	(0.072)	(0.055)	
	t = -4.858	t = -3.631	t = -3.733	
	p = 0.00001	p = 0.0003	p = 0.0002	
ItemTypeSens:zScreenTot	-0.128**	-0.207**	-0.120*	
	(0.044)	(0.072)	(0.047)	
	t = -2.912	t = -2.882	t = -2.538	
	p = 0.004	p = 0.004	p = 0.012	
Constant	0.785***	2.649***	0.796***	1.054***
	(0.031)	(0.074)	(0.033)	(0.049)
	t = 25.692	t = 35.903	t = 23.824	t = 21.522
	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: $^{*}p$ $^{**}p$ $^{***}p<0.001$

Regression predicting item-type flag count by item type, the first three PCs of political knowledge, issue polarization, and out-party dislike; and standard controls, and all interactions between covariates and item type; with clustered SEs by participant

Table S20. *Item-type flag count predicted by item type and political motivation principal components. Across nearly all specifications, we find that increased overall political motivation (PC1) is associated with increased flag count of false, discordant headlines, and decreased flag count for all item types relative to the increased flag count for false, discordant headlines.*

	Dependent variable:				
	Item-type flag count				
	QP Model (1)	OLS Model (2)	QP Exclude True Independents (3)	QP With Quadratic (4)	QP Attention + Practice Filter (5)
ItemTypeDT	-1.334***	-2.018***	-1.300***	-1.333***	-1.742***
	(0.039)	(0.066)	(0.042)	(0.055)	(0.076)
	t = -33.840	t = -30.469	t = -30.820	t = -24.073	t = -22.801
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.171***	-0.481***	-0.180***	-0.173***	-0.208***
	(0.019)	(0.041)	(0.020)	(0.025)	(0.028)
	t = -9.216	t = -11.716	t = -8.984	t = -6.880	t = -7.503
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT	-1.517***	-2.152***	-1.532***	-1.516***	-1.971***
	(0.040)	(0.068)	(0.043)	(0.057)	(0.073)
	t = -38.059	t = -31.424	t = -35.645	t = -26.409	t = -26.946
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.122***	-1.917***	-1.112***	-1.098***	-1.435***
	(0.039)	(0.068)	(0.041)	(0.055)	(0.066)
	t = -29.024	t = -28.182	t = -26.914	t = -19.994	t = -21.691
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
zPK	0.364***	0.970***	0.374***	0.392***	0.263***
	(0.034)	(0.089)	(0.037)	(0.038)	(0.045)
	t = 10.740	t = 10.890	t = 10.171	t = 10.180	t = 5.813
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
I(zPK2)				-0.047	
				(0.031)	
				t = -1.518	
				p = 0.129	
zPartisan	-0.155***	-0.399***	-0.153***	-0.157***	-0.189***
	(0.028)	(0.075)	(0.028)	(0.028)	(0.038)
	t = -5.491	t = -5.342	t = -5.504	t = -5.549	t = -5.025

	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.00000
zage	0.148***	0.380***	0.129**	0.148***	0.047
	(0.037)	(0.093)	(0.040)	(0.037)	(0.050)
	t = 3.963	t = 4.093	t = 3.236	t = 3.987	t = 0.942
	p = 0.0001	p = 0.00005	p = 0.002	p = 0.0001	p = 0.347
zfemale	-0.038	-0.103	-0.039	-0.042	-0.066
	(0.028)	(0.073)	(0.030)	(0.028)	(0.038)
	t = -1.345	t = -1.416	t = -1.307	t = -1.486	t = -1.737
	p = 0.179	p = 0.157	p = 0.192	p = 0.138	p = 0.083
zwhite	0.089*	0.193**	0.110**	0.090**	0.119*
	(0.035)	(0.072)	(0.038)	(0.035)	(0.052)
	t = 2.554	t = 2.680	t = 2.911	t = 2.601	t = 2.292
	p = 0.011	p = 0.008	p = 0.004	p = 0.010	p = 0.022
zcollege	0.078**	0.164*	0.068*	0.078**	0.052
	(0.029)	(0.074)	(0.031)	(0.029)	(0.040)
	t = 2.650	t = 2.206	t = 2.163	t = 2.638	t = 1.280
	p = 0.009	p = 0.028	p = 0.031	p = 0.009	p = 0.201
zScreenTot	0.144***	0.328***	0.134***	0.142***	
	(0.033)	(0.078)	(0.036)	(0.033)	
	t = 4.406	t = 4.205	t = 3.768	t = 4.350	
	p = 0.00002	p = 0.00003	p = 0.0002	p = 0.00002	
ItemTypeDT:zPK	-0.256***	-0.898***	-0.289***	-0.271***	-0.254***
	(0.043)	(0.079)	(0.045)	(0.046)	(0.068)
	t = -6.000	t = -11.297	t = -6.374	t = -5.879	t = -3.707
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.0003
ItemTypeCF:zPK	-0.013	-0.206***	-0.015	-0.016	0.011
	(0.018)	(0.044)	(0.019)	(0.022)	(0.023)
	t = -0.712	t = -4.651	t = -0.799	t = -0.712	t = 0.480
	p = 0.477	p = 0.00001	p = 0.425	p = 0.477	p = 0.632
ItemTypeCT:zPK	-0.409***	-0.992***	-0.437***	-0.433***	-0.449***
	(0.050)	(0.082)	(0.055)	(0.052)	(0.071)
	t = -8.143	t = -12.094	t = -7.920	t = -8.349	t = -6.300

	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens:zPK	-0.387***	-0.987***	-0.412***	-0.407***	-0.362***
	(0.043)	(0.082)	(0.047)	(0.046)	(0.068)
	t = -8.900	t = -12.007	t = -8.754	t = -8.803	t = -5.285
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeDT:I(zPK2)				0.003	
				(0.039)	
				t = 0.073	
				p = 0.942	
ItemTypeCF:I(zPK2)				0.002	
				(0.018)	
				t = 0.133	
				p = 0.895	
ItemTypeCT:I(zPK2)				0.004	
				(0.043)	
				t = 0.087	
				p = 0.931	
ItemTypeSens:I(zPK2)				-0.020	
				(0.042)	
				t = -0.478	
				p = 0.633	
ItemTypeDT:zPartisan	0.343***	0.518***	0.337***	0.344***	0.480***
	(0.040)	(0.066)	(0.039)	(0.040)	(0.069)
	t = 8.505	t = 7.845	t = 8.725	t = 8.486	t = 6.937
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF:zPartisan	0.125***	0.335***	0.123***	0.125***	0.127***
	(0.016)	(0.041)	(0.016)	(0.016)	(0.020)
	t = 7.793	t = 8.228	t = 7.748	t = 7.785	t = 6.223
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT:zPartisan	0.126**	0.384***	0.132**	0.127**	0.183**
	(0.042)	(0.069)	(0.042)	(0.042)	(0.070)
	t = 3.008	t = 5.605	t = 3.157	t = 3.005	t = 2.603
	p = 0.003	p = 0.00000	p = 0.002	p = 0.003	p = 0.010
ItemTypeSens:zPartisan	0.183***	0.420***	0.184***	0.183***	0.188**
	(0.040)	(0.069)	(0.039)	(0.040)	(0.064)

	t = 4.606	t = 6.111	t = 4.738	t = 4.573	t = 2.931
	p = 0.00001	p = 0.000	p = 0.00001	p = 0.00001	p = 0.004
ItemTypeDT:zage	0.022	-0.273***	0.030	0.023	0.043
	(0.048)	(0.082)	(0.051)	(0.048)	(0.075)
	t = 0.466	t = -3.348	t = 0.587	t = 0.472	t = 0.577
	p = 0.642	p = 0.001	p = 0.558	p = 0.638	p = 0.564
ItemTypeCF:zage	-0.045*	-0.157***	-0.043*	-0.045*	-0.040
	(0.019)	(0.047)	(0.021)	(0.019)	(0.025)
	t = -2.330	t = -3.363	t = -2.102	t = -2.338	t = -1.622
	p = 0.020	p = 0.001	p = 0.036	p = 0.020	p = 0.105
ItemTypeCT:zage	-0.049	-0.331***	-0.076	-0.047	0.009
	(0.054)	(0.083)	(0.057)	(0.053)	(0.083)
	t = -0.906	t = -3.973	t = -1.329	t = -0.881	t = 0.109
	p = 0.366	p = 0.0001	p = 0.184	p = 0.379	p = 0.914
ItemTypeSens:zage	-0.145**	-0.378***	-0.132**	-0.143**	-0.048
	(0.048)	(0.083)	(0.050)	(0.047)	(0.072)
	t = -3.039	t = -4.562	t = -2.619	t = -3.017	t = -0.672
	p = 0.003	p = 0.00001	p = 0.009	p = 0.003	p = 0.502
ItemTypeDT:zfemale	-0.123**	0.001	-0.136***	-0.122**	-0.177**
	(0.039)	(0.065)	(0.041)	(0.039)	(0.064)
	t = -3.151	t = 0.009	t = -3.307	t = -3.106	t = -2.748
	p = 0.002	p = 0.994	p = 0.001	p = 0.002	p = 0.006
ItemTypeCF:zfemale	-0.003	0.001	-0.010	-0.003	-0.022
	(0.016)	(0.039)	(0.017)	(0.016)	(0.021)
	t = -0.181	t = 0.031	t = -0.615	t = -0.161	t = -1.047
	p = 0.857	p = 0.976	p = 0.539	p = 0.873	p = 0.296
ItemTypeCT:zfemale	-0.091*	0.038	-0.116**	-0.090*	-0.105
	(0.042)	(0.067)	(0.045)	(0.042)	(0.065)
	t = -2.168	t = 0.573	t = -2.610	t = -2.139	t = -1.611
	p = 0.031	p = 0.567	p = 0.010	p = 0.033	p = 0.108
ItemTypeSens:zfemale	-0.076	0.019	-0.093*	-0.078*	-0.018
	(0.039)	(0.067)	(0.042)	(0.040)	(0.063)
	t = -1.941	t = 0.280	t = -2.235	t = -1.964	t = -0.280
	p = 0.053	p = 0.780	p = 0.026	p = 0.050	p = 0.780

ItemTypeDT:zwhite	-0.049	-0.174**	-0.073	-0.049	-0.024
	(0.045)	(0.065)	(0.048)	(0.045)	(0.082)
	t = -1.103	t = -2.683	t = -1.507	t = -1.096	t = -0.292
	p = 0.271	p = 0.008	p = 0.132	p = 0.274	p = 0.771
ItemTypeCF:zwhite	-0.001	-0.043	-0.013	-0.001	-0.004
	(0.020)	(0.038)	(0.021)	(0.020)	(0.030)
	t = -0.048	t = -1.134	t = -0.626	t = -0.051	t = -0.133
	p = 0.963	p = 0.257	p = 0.532	p = 0.960	p = 0.894
ItemTypeCT:zwhite	-0.080	-0.187**	-0.098	-0.079	0.050
	(0.047)	(0.066)	(0.050)	(0.047)	(0.085)
	t = -1.699	t = -2.858	t = -1.932	t = -1.688	t = 0.590
	p = 0.090	p = 0.005	p = 0.054	p = 0.092	p = 0.556
ItemTypeSens:zwhite	-0.047	-0.163*	-0.062	-0.046	-0.074
	(0.043)	(0.065)	(0.046)	(0.043)	(0.076)
	t = -1.105	t = -2.506	t = -1.372	t = -1.079	t = -0.972
	p = 0.270	p = 0.013	p = 0.170	p = 0.281	p = 0.331
ItemTypeDT:zcollege	-0.036	-0.136*	-0.014	-0.036	-0.056
	(0.040)	(0.066)	(0.043)	(0.040)	(0.065)
	t = -0.904	t = -2.060	t = -0.327	t = -0.894	t = -0.853
	p = 0.367	p = 0.040	p = 0.744	p = 0.372	p = 0.394
ItemTypeCF:zcollege	-0.016	-0.048	-0.024	-0.016	0.001
	(0.017)	(0.040)	(0.018)	(0.017)	(0.023)
	t = -0.965	t = -1.203	t = -1.302	t = -0.965	t = 0.056
	p = 0.335	p = 0.230	p = 0.193	p = 0.335	p = 0.956
ItemTypeCT:zcollege	-0.105*	-0.177**	-0.075	-0.104*	-0.063
	(0.042)	(0.068)	(0.046)	(0.042)	(0.071)
	t = -2.476	t = -2.618	t = -1.654	t = -2.456	t = -0.887
	p = 0.014	p = 0.009	p = 0.099	p = 0.015	p = 0.376
ItemTypeSens:zcollege	-0.050	-0.143*	-0.020	-0.049	-0.041
	(0.040)	(0.068)	(0.042)	(0.040)	(0.064)
	t = -1.253	t = -2.091	t = -0.477	t = -1.225	t = -0.643
	p = 0.211	p = 0.037	p = 0.634	p = 0.221	p = 0.521
ItemTypeDT:zScreenTot	-0.289***	-0.419***	-0.268***	-0.288***	
	(0.045)	(0.070)	(0.048)	(0.045)	
	t = -6.452	t = -5.977	t = -5.586	t = -6.443	
	p = 0.000	p = 0.000	p = 0.00000	p = 0.000	

ItemTypeCF:zScreenTot	-0.001	-0.059	0.001	-0.001	
	(0.019)	(0.043)	(0.021)	(0.019)	
	t = -0.055	t = -1.380	t = 0.046	t = -0.041	
	p = 0.957	p = 0.168	p = 0.964	p = 0.968	
ItemTypeCT:zScreenTot	-0.276***	-0.395***	-0.238***	-0.274***	
	(0.049)	(0.071)	(0.054)	(0.049)	
	t = -5.657	t = -5.531	t = -4.416	t = -5.648	
	p = 0.00000	p = 0.00000	p = 0.00002	p = 0.00000	
ItemTypeSens:zScreenTot	-0.166***	-0.345***	-0.157***	-0.165***	
	(0.043)	(0.071)	(0.047)	(0.043)	
	t = -3.824	t = -4.830	t = -3.352	t = -3.820	
	p = 0.0002	p = 0.00001	p = 0.001	p = 0.0002	
Constant	0.809***	2.655***	0.826***	0.850***	1.092***
	(0.031)	(0.075)	(0.033)	(0.042)	(0.048)
	t = 26.419	t = 35.346	t = 24.861	t = 20.434	t = 22.769
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

Regression predicting item-type flag count by item type, political knowledge, their interaction, and standard controls (each interacted with item type); with clustered SEs by participant

Table S21. *Item-type flag count predicted by item type and political knowledge. Across nearly all specifications, we find that increased political knowledge is associated with increased flagging of discordant false content – as well as increased flagging of concordant false content. We do not observe similar increases by political knowledge in flagging of other headline types.*

	Dependent variable:				
	Item-type flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
ItemTypeDT	-1.353***	-2.026***	-1.321***	-1.356***	-1.853***
	(0.040)	(0.067)	(0.044)	(0.054)	(0.080)
	t = -33.644	t = -30.287	t = -30.308	t = -25.129	t = -23.284
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.151***	-0.478***	-0.160***	-0.147***	-0.167***
	(0.018)	(0.040)	(0.019)	(0.022)	(0.025)
	t = -8.532	t = -11.804	t = -8.316	t = -6.799	t = -6.589
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT	-1.527***	-2.160***	-1.535***	-1.477***	-2.039***
	(0.040)	(0.069)	(0.043)	(0.055)	(0.073)
	t = -37.753	t = -31.275	t = -35.380	t = -26.807	t = -27.861
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.131***	-1.925***	-1.114***	-1.107***	-1.485***
	(0.039)	(0.069)	(0.041)	(0.052)	(0.068)
	t = -28.964	t = -28.104	t = -26.889	t = -21.230	t = -21.759
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
zIssuePol	0.304***	0.938***	0.305***	0.346***	0.200***
	(0.027)	(0.091)	(0.029)	(0.042)	(0.035)
	t = 11.305	t = 10.302	t = 10.604	t = 8.192	t = 5.743
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
I(zIssuePol2)				-0.032	
				(0.025)	
				t = -1.313	
				p = 0.190	
zPartisan	-0.112***	-0.273***	-0.111***	-0.113***	-0.154***
	(0.028)	(0.074)	(0.028)	(0.028)	(0.038)

	t = -4.029	t = -3.717	t = -4.007	t = -4.064	t = -4.038
	p = 0.0001	p = 0.0003	p = 0.0001	p = 0.00005	p = 0.0001
zage	0.214***	0.531***	0.194***	0.210***	0.109*
	(0.034)	(0.089)	(0.037)	(0.034)	(0.046)
	t = 6.194	t = 5.973	t = 5.241	t = 6.099	t = 2.382
	p = 0.000	p = 0.000	p = 0.00000	p = 0.000	p = 0.018
zfemale	-0.079**	-0.201**	-0.083**	-0.081**	-0.102**
	(0.028)	(0.073)	(0.030)	(0.028)	(0.038)
	t = -2.794	t = -2.752	t = -2.776	t = -2.874	t = -2.699
	p = 0.006	p = 0.006	p = 0.006	p = 0.005	p = 0.007
zwhite	0.092**	0.192**	0.116**	0.092**	0.118*
	(0.034)	(0.071)	(0.037)	(0.034)	(0.051)
	t = 2.702	t = 2.695	t = 3.124	t = 2.695	t = 2.285
	p = 0.007	p = 0.008	p = 0.002	p = 0.008	p = 0.023
zcollege	0.090**	0.188*	0.079*	0.089**	0.058
	(0.029)	(0.074)	(0.031)	(0.029)	(0.040)
	t = 3.083	t = 2.551	t = 2.550	t = 3.031	t = 1.443
	p = 0.003	p = 0.011	p = 0.011	p = 0.003	p = 0.150
zScreenTot	0.162***	0.341***	0.165***	0.159***	
	(0.032)	(0.078)	(0.035)	(0.033)	
	t = 4.999	t = 4.398	t = 4.749	t = 4.884	
	p = 0.00000	p = 0.00002	p = 0.00001	p = 0.00001	
ItemTypeDT:zIssuePol	-0.122**	-0.806***	-0.118**	-0.137**	0.015
	(0.039)	(0.084)	(0.042)	(0.053)	(0.060)
	t = -3.094	t = -9.658	t = -2.844	t = -2.585	t = 0.246
	p = 0.002	p = 0.000	p = 0.005	p = 0.010	p = 0.806
ItemTypeCF:zIssuePol	-0.090***	-0.413***	-0.080***	-0.091***	-0.074***
	(0.016)	(0.049)	(0.017)	(0.025)	(0.020)
	t = -5.634	t = -8.411	t = -4.768	t = -3.660	t = -3.621
	p = 0.00000	p = 0.000	p = 0.00001	p = 0.0003	p = 0.0003
ItemTypeCT:zIssuePol	-0.281***	-0.926***	-0.313***	-0.256***	-0.240***

	(0.048)	(0.086)	(0.049)	(0.057)	(0.067)
	t = -5.919	t = -10.769	t = -6.377	t = -4.501	t = -3.594
	p = 0.000	p = 0.000	p = 0.000	p = 0.00001	p = 0.0004
ItemTypeSens:zIssuePol	-0.306***	-0.938***	-0.329***	-0.304***	-0.235***
	(0.042)	(0.085)	(0.043)	(0.054)	(0.060)
	t = -7.261	t = -10.990	t = -7.584	t = -5.669	t = -3.886
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.0002
ItemTypeDT:I(zIssuePol2)				0.006	
				(0.038)	
				t = 0.167	
				p = 0.868	
ItemTypeCF:I(zIssuePol2)				-0.003	
				(0.014)	
				t = -0.190	
				p = 0.850	
ItemTypeCT:I(zIssuePol2)				-0.050	
				(0.046)	
				t = -1.090	
				p = 0.276	
ItemTypeSens:I(zIssuePol2)				-0.022	
				(0.039)	
				t = -0.560	
				p = 0.576	
ItemTypeDT:zPartisan	0.317***	0.410***	0.313***	0.317***	0.485***
	(0.041)	(0.065)	(0.039)	(0.041)	(0.071)
	t = 7.809	t = 6.321	t = 7.959	t = 7.784	t = 6.849
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF:zPartisan	0.109***	0.281***	0.110***	0.109***	0.107***
	(0.016)	(0.040)	(0.016)	(0.016)	(0.021)
	t = 6.896	t = 6.959	t = 6.935	t = 6.911	t = 5.069
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeCT:zPartisan	0.085*	0.260***	0.087*	0.083	0.148*
	(0.043)	(0.067)	(0.043)	(0.043)	(0.074)
	t = 1.974	t = 3.866	t = 2.036	t = 1.940	t = 1.989

	p = 0.049	p = 0.0002	p = 0.042	p = 0.053	p = 0.047
ItemTypeSens:zPartisan	0.140***	0.294***	0.138***	0.139***	0.149*
	(0.040)	(0.067)	(0.039)	(0.040)	(0.066)
	t = 3.478	t = 4.374	t = 3.513	t = 3.473	t = 2.256
	p = 0.001	p = 0.00002	p = 0.0005	p = 0.001	p = 0.025
ItemTypeDT:zage	-0.045	-0.423***	-0.048	-0.044	-0.046
	(0.046)	(0.079)	(0.049)	(0.046)	(0.073)
	t = -0.977	t = -5.345	t = -0.968	t = -0.956	t = -0.630
	p = 0.329	p = 0.00000	p = 0.334	p = 0.340	p = 0.529
ItemTypeCF:zage	-0.035	-0.154***	-0.036	-0.035	-0.029
	(0.019)	(0.045)	(0.020)	(0.019)	(0.024)
	t = -1.884	t = -3.410	t = -1.808	t = -1.870	t = -1.205
	p = 0.060	p = 0.001	p = 0.071	p = 0.062	p = 0.229
ItemTypeCT:zage	-0.133**	-0.491***	-0.159**	-0.134**	-0.111
	(0.049)	(0.081)	(0.053)	(0.049)	(0.081)
	t = -2.684	t = -6.070	t = -2.986	t = -2.718	t = -1.379
	p = 0.008	p = 0.000	p = 0.003	p = 0.007	p = 0.168
ItemTypeSens:zage	-0.218***	-0.534***	-0.204***	-0.218***	-0.140*
	(0.046)	(0.080)	(0.049)	(0.046)	(0.070)
	t = -4.727	t = -6.663	t = -4.175	t = -4.750	t = -1.987
	p = 0.00001	p = 0.000	p = 0.00003	p = 0.00001	p = 0.047
ItemTypeDT:zfemale	-0.092*	0.089	-0.102*	-0.090*	-0.137*
	(0.039)	(0.066)	(0.041)	(0.039)	(0.064)
	t = -2.353	t = 1.351	t = -2.467	t = -2.308	t = -2.131
	p = 0.019	p = 0.177	p = 0.014	p = 0.021	p = 0.034
ItemTypeCF:zfemale	0.001	0.031	-0.005	0.001	-0.022
	(0.016)	(0.039)	(0.017)	(0.016)	(0.021)
	t = 0.058	t = 0.800	t = -0.323	t = 0.061	t = -1.040
	p = 0.954	p = 0.424	p = 0.747	p = 0.952	p = 0.299

ItemTypeCT:zfemale	-0.048	0.138*	-0.068	-0.048	-0.044
	(0.042)	(0.068)	(0.045)	(0.042)	(0.067)
	t = -1.139	t = 2.034	t = -1.521	t = -1.148	t = -0.655
	p = 0.255	p = 0.042	p = 0.129	p = 0.251	p = 0.513
ItemTypeSens:zfemale	-0.034	0.118	-0.045	-0.034	0.032
	(0.039)	(0.067)	(0.042)	(0.039)	(0.062)
	t = -0.860	t = 1.756	t = -1.085	t = -0.853	t = 0.522
	p = 0.390	p = 0.080	p = 0.279	p = 0.394	p = 0.602
ItemTypeDT:zwhite	-0.053	-0.174**	-0.079	-0.053	-0.039
	(0.045)	(0.064)	(0.048)	(0.045)	(0.081)
	t = -1.185	t = -2.704	t = -1.647	t = -1.185	t = -0.475
	p = 0.237	p = 0.007	p = 0.100	p = 0.236	p = 0.635
ItemTypeCF:zwhite	0.001	-0.040	-0.011	0.001	0.004
	(0.020)	(0.037)	(0.020)	(0.020)	(0.029)
	t = 0.063	t = -1.073	t = -0.518	t = 0.058	t = 0.128
	p = 0.950	p = 0.284	p = 0.605	p = 0.955	p = 0.898
ItemTypeCT:zwhite	-0.084	-0.187**	-0.104*	-0.084	0.042
	(0.047)	(0.065)	(0.050)	(0.047)	(0.085)
	t = -1.774	t = -2.870	t = -2.067	t = -1.780	t = 0.486
	p = 0.077	p = 0.005	p = 0.039	p = 0.076	p = 0.627
ItemTypeSens:zwhite	-0.051	-0.162*	-0.068	-0.051	-0.076
	(0.043)	(0.064)	(0.045)	(0.043)	(0.075)
	t = -1.197	t = -2.519	t = -1.525	t = -1.200	t = -1.016
	p = 0.232	p = 0.012	p = 0.128	p = 0.231	p = 0.310
ItemTypeDT:zcollege	-0.050	-0.162*	-0.029	-0.048	-0.073
	(0.040)	(0.066)	(0.043)	(0.040)	(0.065)
	t = -1.236	t = -2.451	t = -0.681	t = -1.201	t = -1.128
	p = 0.217	p = 0.015	p = 0.496	p = 0.230	p = 0.260
ItemTypeCF:zcollege	-0.014	-0.044	-0.022	-0.014	0.002
	(0.017)	(0.039)	(0.018)	(0.017)	(0.023)

	t = -0.833	t = -1.117	t = -1.231	t = -0.807	t = 0.089
	p = 0.405	p = 0.265	p = 0.219	p = 0.420	p = 0.930
ItemTypeCT:zcollege	-0.121**	-0.203**	-0.091*	-0.120**	-0.074
	(0.043)	(0.067)	(0.046)	(0.043)	(0.072)
	t = -2.837	t = -3.024	t = -1.994	t = -2.815	t = -1.027
	p = 0.005	p = 0.003	p = 0.047	p = 0.005	p = 0.305
ItemTypeSens:zcollege	-0.063	-0.168*	-0.034	-0.062	-0.050
	(0.040)	(0.068)	(0.043)	(0.040)	(0.065)
	t = -1.589	t = -2.475	t = -0.789	t = -1.562	t = -0.778
	p = 0.113	p = 0.014	p = 0.431	p = 0.119	p = 0.437
ItemTypeDT:zScreenTot	-0.321***	-0.446***	-0.316***	-0.319***	
	(0.045)	(0.070)	(0.047)	(0.045)	
	t = -7.200	t = -6.344	t = -6.680	t = -7.157	
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	
ItemTypeCF:zScreenTot	0.020	-0.007	0.019	0.021	
	(0.019)	(0.043)	(0.020)	(0.019)	
	t = 1.065	t = -0.171	t = 0.921	t = 1.077	
	p = 0.288	p = 0.865	p = 0.357	p = 0.282	
ItemTypeCT:zScreenTot	-0.311***	-0.417***	-0.284***	-0.312***	
	(0.049)	(0.071)	(0.053)	(0.049)	
	t = -6.350	t = -5.859	t = -5.319	t = -6.369	
	p = 0.000	p = 0.000	p = 0.00000	p = 0.000	
ItemTypeSens:zScreenTot	-0.189***	-0.362***	-0.192***	-0.189***	
	(0.044)	(0.071)	(0.046)	(0.044)	
	t = -4.344	t = -5.069	t = -4.145	t = -4.332	
	p = 0.00002	p = 0.00000	p = 0.00004	p = 0.00002	
Constant	0.817***	2.662***	0.829***	0.845***	1.123***
	(0.030)	(0.075)	(0.033)	(0.038)	(0.045)
	t = 27.060	t = 35.327	t = 25.110	t = 22.515	t = 24.765

	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: $^{*}p$ $^{**}p$ $^{***}p<0.001$

Regression predicting item-type flag count by item type, issue polarization, their interaction, and standard controls (each interacted with item type); with clustered SEs by participant

Table S22. *Item-type flag count predicted by item type and issue polarization. Across nearly all specifications, we find that increased issue polarization is associated with decreased flag count for all item types relative to the observed increase in flag count for false, discordant headlines.*

	Dependent variable:				
	Item-type flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
ItemTypeDT	-1.374***	-2.048***	-1.344***	-1.562***	-1.870***
	(0.040)	(0.069)	(0.043)	(0.057)	(0.076)
	t = -34.452	t = -29.815	t = -31.286	t = -27.639	t = -24.555
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.163***	-0.489***	-0.170***	-0.143***	-0.179***
	(0.017)	(0.041)	(0.019)	(0.023)	(0.024)
	t = -9.441	t = -11.879	t = -9.049	t = -6.342	t = -7.374
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT	-1.560***	-2.185***	-1.575***	-1.639***	-2.092***
	(0.041)	(0.071)	(0.044)	(0.056)	(0.072)
	t = -38.370	t = -30.684	t = -35.984	t = -29.174	t = -28.851
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.165***	-1.951***	-1.156***	-1.236***	-1.546***
	(0.039)	(0.071)	(0.042)	(0.053)	(0.066)
	t = -29.821	t = -27.587	t = -27.821	t = -23.471	t = -23.557
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
zOutThermom	0.201***	0.448***	0.200***	0.173***	0.163***
	(0.030)	(0.066)	(0.032)	(0.034)	(0.045)
	t = 6.610	t = 6.751	t = 6.238	t = 5.093	t = 3.639
	p = 0.000	p = 0.000	p = 0.000	p = 0.00000	p = 0.0003
I(zOutThermom2)				-0.091**	
				(0.032)	
				t = -2.841	

				p = 0.005	
zPartisan	-0.161***	-0.411***	-0.160***	-0.164***	-0.198***
	(0.029)	(0.077)	(0.028)	(0.029)	(0.038)
	t = -5.612	t = -5.342	t = -5.658	t = -5.683	t = -5.244
	p = 0.00000	p = 0.00000	p = 0.00000	p = 0.000	p = 0.00000
zage	0.241***	0.626***	0.223***	0.236***	0.119**
	(0.035)	(0.091)	(0.038)	(0.035)	(0.046)
	t = 6.808	t = 6.909	t = 5.907	t = 6.703	t = 2.578
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.010
zfemale	-0.082**	-0.215**	-0.089**	-0.090**	-0.099**
	(0.028)	(0.076)	(0.030)	(0.028)	(0.038)
	t = -2.889	t = -2.842	t = -2.924	t = -3.142	t = -2.594
	p = 0.004	p = 0.005	p = 0.004	p = 0.002	p = 0.010
zwhite	0.105**	0.213**	0.131***	0.104**	0.141**
	(0.035)	(0.074)	(0.038)	(0.035)	(0.053)
	t = 2.979	t = 2.894	t = 3.401	t = 2.938	t = 2.653
	p = 0.003	p = 0.004	p = 0.001	p = 0.004	p = 0.008
zcollege	0.099***	0.224**	0.088**	0.097**	0.064
	(0.030)	(0.075)	(0.032)	(0.030)	(0.041)
	t = 3.304	t = 2.966	t = 2.749	t = 3.252	t = 1.581
	p = 0.001	p = 0.004	p = 0.006	p = 0.002	p = 0.114
zScreenTot	0.222***	0.515***	0.225***	0.212***	
	(0.032)	(0.077)	(0.034)	(0.032)	
	t = 6.961	t = 6.660	t = 6.521	t = 6.585	
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	
ItemTypeDT:zOutThermom	-0.075	-0.369***	-0.077	0.003	0.075
	(0.042)	(0.059)	(0.044)	(0.044)	(0.076)
	t = -1.793	t = -6.224	t = -1.766	t = 0.075	t = 0.994
	p = 0.073	p = 0.000	p = 0.078	p = 0.941	p = 0.321

ItemTypeCF:zOutThermom	-0.085***	-0.234***	-0.095***	-0.099***	-0.099***
	(0.018)	(0.037)	(0.019)	(0.020)	(0.026)
	t = -4.765	t = -6.280	t = -5.100	t = -4.839	t = -3.818
	p = 0.00001	p = 0.000	p = 0.00000	p = 0.00001	p = 0.0002
ItemTypeCT:zOutThermom	-0.209***	-0.453***	-0.222***	-0.186***	-0.207**
	(0.042)	(0.061)	(0.044)	(0.050)	(0.078)
	t = -4.957	t = -7.476	t = -5.000	t = -3.749	t = -2.641
	p = 0.00000	p = 0.000	p = 0.00000	p = 0.0002	p = 0.009
ItemTypeSens:zOutThermom	-0.224***	-0.465***	-0.245***	-0.206***	-0.137*
	(0.040)	(0.060)	(0.041)	(0.045)	(0.069)
	t = -5.598	t = -7.743	t = -5.934	t = -4.584	t = -1.984
	p = 0.00000	p = 0.000	p = 0.000	p = 0.00001	p = 0.048
ItemTypeDT:I(zOutThermom2)				0.188***	
				(0.038)	
				t = 4.934	
				p = 0.00000	
ItemTypeCF:I(zOutThermom2)				-0.021	
				(0.018)	
				t = -1.157	
				p = 0.248	
ItemTypeCT:I(zOutThermom2)				0.085*	
				(0.039)	
				t = 2.167	
				p = 0.031	
ItemTypeSens:I(zOutThermom2)				0.077*	
				(0.038)	
				t = 2.019	
				p = 0.044	
ItemTypeDT:zPartisan	0.344***	0.528***	0.339***	0.349***	0.489***
	(0.040)	(0.069)	(0.039)	(0.039)	(0.068)
	t = 8.562	t = 7.693	t = 8.781	t = 8.843	t = 7.213

	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF:zPartisan	0.125***	0.343***	0.124***	0.125***	0.123***
	(0.016)	(0.041)	(0.016)	(0.016)	(0.020)
	t = 7.954	t = 8.432	t = 7.966	t = 7.850	t = 6.055
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCT:zPartisan	0.132**	0.396***	0.139**	0.135**	0.199**
	(0.043)	(0.071)	(0.043)	(0.043)	(0.072)
	t = 3.092	t = 5.571	t = 3.269	t = 3.167	t = 2.776
	p = 0.002	p = 0.00000	p = 0.002	p = 0.002	p = 0.006
ItemTypeSens:zPartisan	0.190***	0.432***	0.193***	0.193***	0.202**
	(0.041)	(0.071)	(0.040)	(0.041)	(0.064)
	t = 4.684	t = 6.055	t = 4.853	t = 4.723	t = 3.136
	p = 0.00001	p = 0.000	p = 0.00001	p = 0.00001	p = 0.002
ItemTypeDT:zage	-0.052	-0.506***	-0.054	-0.042	-0.038
	(0.046)	(0.080)	(0.049)	(0.046)	(0.073)
	t = -1.135	t = -6.340	t = -1.117	t = -0.914	t = -0.518
	p = 0.257	p = 0.000	p = 0.265	p = 0.361	p = 0.605
ItemTypeCF:zage	-0.039*	-0.191***	-0.037	-0.040*	-0.029
	(0.018)	(0.045)	(0.019)	(0.018)	(0.024)
	t = -2.118	t = -4.234	t = -1.894	t = -2.181	t = -1.205
	p = 0.035	p = 0.00003	p = 0.059	p = 0.030	p = 0.229
ItemTypeCT:zage	-0.155**	-0.583***	-0.187***	-0.150**	-0.123
	(0.051)	(0.082)	(0.054)	(0.051)	(0.082)
	t = -3.054	t = -7.100	t = -3.453	t = -2.970	t = -1.491
	p = 0.003	p = 0.000	p = 0.001	p = 0.003	p = 0.136
ItemTypeSens:zage	-0.242***	-0.626***	-0.231***	-0.238***	-0.156*
	(0.047)	(0.081)	(0.050)	(0.047)	(0.071)
	t = -5.142	t = -7.691	t = -4.644	t = -5.059	t = -2.195

	p = 0.00000	p = 0.000	p = 0.00001	p = 0.00000	p = 0.029
ItemTypeDT:zfemale	-0.100*	0.099	-0.110**	-0.083*	-0.147*
	(0.039)	(0.068)	(0.041)	(0.039)	(0.064)
	t = -2.571	t = 1.454	t = -2.662	t = -2.129	t = -2.298
	p = 0.011	p = 0.146	p = 0.008	p = 0.034	p = 0.022
ItemTypeCF:zfemale	0.002	0.042	-0.002	-0.0004	-0.024
	(0.016)	(0.040)	(0.017)	(0.016)	(0.021)
	t = 0.110	t = 1.043	t = -0.117	t = -0.028	t = -1.130
	p = 0.913	p = 0.298	p = 0.907	p = 0.978	p = 0.259
ItemTypeCT:zfemale	-0.042	0.153*	-0.060	-0.036	-0.047
	(0.042)	(0.071)	(0.046)	(0.042)	(0.067)
	t = -0.994	t = 2.167	t = -1.306	t = -0.849	t = -0.701
	p = 0.321	p = 0.031	p = 0.192	p = 0.396	p = 0.484
ItemTypeSens:zfemale	-0.027	0.134	-0.034	-0.022	0.027
	(0.040)	(0.070)	(0.043)	(0.040)	(0.063)
	t = -0.688	t = 1.916	t = -0.790	t = -0.542	t = 0.433
	p = 0.492	p = 0.056	p = 0.430	p = 0.588	p = 0.666
ItemTypeDT:zwhite	-0.064	-0.192**	-0.091	-0.064	-0.045
	(0.045)	(0.067)	(0.049)	(0.045)	(0.082)
	t = -1.408	t = -2.883	t = -1.871	t = -1.412	t = -0.547
	p = 0.160	p = 0.004	p = 0.062	p = 0.158	p = 0.585
ItemTypeCF:zwhite	-0.005	-0.050	-0.017	-0.005	-0.007
	(0.020)	(0.038)	(0.021)	(0.020)	(0.030)
	t = -0.244	t = -1.310	t = -0.841	t = -0.248	t = -0.240
	p = 0.808	p = 0.191	p = 0.401	p = 0.805	p = 0.811
ItemTypeCT:zwhite	-0.097*	-0.208**	-0.120*	-0.095*	0.015
	(0.047)	(0.067)	(0.051)	(0.047)	(0.085)

	t = -2.039	t = -3.086	t = -2.341	t = -2.013	t = 0.172
	p = 0.042	p = 0.003	p = 0.020	p = 0.045	p = 0.864
ItemTypeSens:zwhite	-0.065	-0.184**	-0.085	-0.063	-0.102
	(0.043)	(0.067)	(0.046)	(0.043)	(0.076)
	t = -1.491	t = -2.747	t = -1.836	t = -1.455	t = -1.341
	p = 0.136	p = 0.007	p = 0.067	p = 0.146	p = 0.180
ItemTypeDT:zcollege	-0.052	-0.192**	-0.031	-0.047	-0.071
	(0.040)	(0.067)	(0.043)	(0.040)	(0.065)
	t = -1.283	t = -2.857	t = -0.706	t = -1.176	t = -1.081
	p = 0.200	p = 0.005	p = 0.481	p = 0.240	p = 0.280
ItemTypeCF:zcollege	-0.016	-0.059	-0.023	-0.016	-0.0004
	(0.017)	(0.040)	(0.018)	(0.017)	(0.022)
	t = -0.936	t = -1.488	t = -1.297	t = -0.941	t = -0.017
	p = 0.350	p = 0.137	p = 0.195	p = 0.347	p = 0.987
ItemTypeCT:zcollege	-0.129**	-0.238***	-0.099*	-0.127**	-0.082
	(0.043)	(0.069)	(0.046)	(0.043)	(0.072)
	t = -2.996	t = -3.450	t = -2.151	t = -2.971	t = -1.135
	p = 0.003	p = 0.001	p = 0.032	p = 0.003	p = 0.257
ItemTypeSens:zcollege	-0.072	-0.204**	-0.042	-0.070	-0.058
	(0.040)	(0.070)	(0.043)	(0.040)	(0.065)
	t = -1.777	t = -2.920	t = -0.973	t = -1.747	t = -0.884
	p = 0.076	p = 0.004	p = 0.331	p = 0.081	p = 0.377
ItemTypeDT:zScreenTot	-0.356***	-0.599***	-0.351***	-0.331***	
	(0.044)	(0.071)	(0.047)	(0.043)	
	t = -8.113	t = -8.486	t = -7.494	t = -7.613	
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	

ItemTypeCF:zScreenTot	0.002	-0.079	0.004	-0.001	
	(0.018)	(0.042)	(0.020)	(0.018)	
	t = 0.101	t = -1.872	t = 0.192	t = -0.065	
	p = 0.920	p = 0.062	p = 0.848	p = 0.949	
ItemTypeCT:zScreenTot	-0.365***	-0.587***	-0.342***	-0.355***	
	(0.047)	(0.071)	(0.051)	(0.046)	
	t = -7.841	t = -8.224	t = -6.734	t = -7.656	
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	
ItemTypeSens:zScreenTot	-0.247***	-0.534***	-0.251***	-0.238***	
	(0.042)	(0.071)	(0.045)	(0.042)	
	t = -5.907	t = -7.493	t = -5.610	t = -5.646	
	p = 0.000	p = 0.000	p = 0.00000	p = 0.00000	
Constant	0.851***	2.688***	0.869***	0.937***	1.170***
	(0.030)	(0.078)	(0.033)	(0.042)	(0.045)
	t = 27.974	t = 34.679	t = 26.176	t = 22.041	t = 26.266
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

Regression predicting item-type flag count by item type, out-party dislike, their interaction, and standard controls (each interacted with item type); with clustered SEs by participant

Table S23. *Item-type flag count predicted by item type and out-party dislike. We find that out-party dislike is associated with increased flagging of false discordant headlines – as well as true discordant headlines, in multiplicative space (though additively, out-party dislike is associated with a greater number of false than true discordant flags). We also find robust evidence for an interaction between quadratic out-party dislike and true, discordant headline flagging (relative to false, discordant headline flagging), suggesting a nonlinear relationship between out-party dislike and true (versus false) discordant headline flagging.*

	Dependent variable:				
	Item-type flag count				
	QP Model	OLS Model	QP Exclude True Independents	QP With Quadratic	QP Attention + Practice Filter
	(1)	(2)	(3)	(4)	(5)
ItemTypeDT	-1.324***	-2.013***	-1.290***	-1.495***	-1.777***
	(0.040)	(0.065)	(0.043)	(0.072)	(0.081)
	t = -32.962	t = -30.825	t = -29.814	t = -20.643	t = -22.048
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF	-0.156***	-0.479***	-0.163***	-0.146***	-0.181***
	(0.018)	(0.040)	(0.020)	(0.031)	(0.028)
	t = -8.420	t = -11.881	t = -8.130	t = -4.783	t = -6.512
	p = 0.000	p = 0.000	p = 0.000	p = 0.00001	p = 0.000
ItemTypeCT	-1.495***	-2.146***	-1.502***	-1.543***	-1.935***
	(0.039)	(0.067)	(0.042)	(0.071)	(0.073)
	t = -37.868	t = -31.924	t = -35.423	t = -21.625	t = -26.491
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeSens	-1.099***	-1.912***	-1.082***	-1.149***	-1.398***
	(0.038)	(0.067)	(0.040)	(0.069)	(0.067)
	t = -28.771	t = -28.659	t = -26.741	t = -16.586	t = -20.859
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
zPK	0.266***	0.715***	0.273***	0.288***	0.197***
	(0.035)	(0.088)	(0.038)	(0.039)	(0.047)
	t = 7.610	t = 8.107	t = 7.196	t = 7.364	t = 4.190
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00003
I(zPK2)				-0.055	
				(0.031)	
				t = -1.812	

				p = 0.070	
zIssuePol	0.207***	0.691***	0.202***	0.258***	0.127***
	(0.029)	(0.092)	(0.031)	(0.043)	(0.038)
	t = 7.209	t = 7.532	t = 6.616	t = 6.055	t = 3.372
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.001
I(zIssuePol2)				-0.028	
				(0.024)	
				t = -1.156	
				p = 0.248	
zOutThermom	0.084**	0.180**	0.087**	0.048	0.072
	(0.030)	(0.063)	(0.032)	(0.035)	(0.046)
	t = 2.755	t = 2.855	t = 2.693	t = 1.397	t = 1.567
	p = 0.006	p = 0.005	p = 0.008	p = 0.163	p = 0.118
I(zOutThermom2)				-0.108***	
				(0.031)	
				t = -3.457	
				p = 0.001	
zPartisan	-0.121***	-0.317***	-0.121***	-0.124***	-0.157***
	(0.027)	(0.073)	(0.027)	(0.028)	(0.038)
	t = -4.424	t = -4.350	t = -4.454	t = -4.500	t = -4.168
	p = 0.00001	p = 0.00002	p = 0.00001	p = 0.00001	p = 0.00004
zage	0.135***	0.324***	0.116**	0.128***	0.051
	(0.036)	(0.091)	(0.039)	(0.035)	(0.048)
	t = 3.732	t = 3.554	t = 2.997	t = 3.608	t = 1.055
	p = 0.0002	p = 0.0004	p = 0.003	p = 0.0004	p = 0.292
zfemale	-0.056*	-0.169*	-0.063*	-0.073**	-0.078*
	(0.028)	(0.072)	(0.030)	(0.028)	(0.038)
	t = -2.001	t = -2.329	t = -2.106	t = -2.597	t = -2.077
	p = 0.046	p = 0.020	p = 0.036	p = 0.010	p = 0.038
zwhite	0.085*	0.190**	0.106**	0.084*	0.112*
	(0.034)	(0.070)	(0.037)	(0.034)	(0.051)

	t = 2.497	t = 2.700	t = 2.866	t = 2.474	t = 2.172
	p = 0.013	p = 0.007	p = 0.005	p = 0.014	p = 0.030
zcollege	0.076**	0.148*	0.066*	0.072*	0.053
	(0.029)	(0.073)	(0.031)	(0.029)	(0.040)
	t = 2.646	t = 2.041	t = 2.176	t = 2.500	t = 1.348
	p = 0.009	p = 0.042	p = 0.030	p = 0.013	p = 0.178
zScreenTot	0.107**	0.191*	0.102**	0.088**	
	(0.033)	(0.079)	(0.036)	(0.033)	
	t = 3.253	t = 2.432	t = 2.882	t = 2.670	
	p = 0.002	p = 0.016	p = 0.004	p = 0.008	
ItemTypeDT:zPK	-0.224***	-0.687***	-0.262***	-0.221***	-0.293***
	(0.044)	(0.077)	(0.047)	(0.047)	(0.072)
	t = -5.077	t = -8.874	t = -5.578	t = -4.702	t = -4.050
	p = 0.00000	p = 0.000	p = 0.00000	p = 0.00001	p = 0.0001
ItemTypeCF:zPK	0.033	-0.066	0.030	0.027	0.054*
	(0.019)	(0.044)	(0.020)	(0.022)	(0.026)
	t = 1.740	t = -1.485	t = 1.476	t = 1.195	t = 2.101
	p = 0.082	p = 0.138	p = 0.141	p = 0.233	p = 0.036
ItemTypeCT:zPK	-0.322***	-0.744***	-0.338***	-0.339***	-0.387***
	(0.052)	(0.080)	(0.057)	(0.053)	(0.075)
	t = -6.221	t = -9.329	t = -5.952	t = -6.370	t = -5.159
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeSens:zPK	-0.288***	-0.732***	-0.301***	-0.303***	-0.296***
	(0.045)	(0.080)	(0.049)	(0.047)	(0.073)
	t = -6.425	t = -9.137	t = -6.207	t = -6.422	t = -4.056
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00005
ItemTypeDT:I(zPK2)				-0.007	
				(0.040)	

				t = -0.178 p = 0.859	
ItemTypeCF:I(zPK2)				0.007 (0.018) t = 0.381 p = 0.704	
ItemTypeCT:I(zPK2)				0.011 (0.044) t = 0.260 p = 0.795	
ItemTypeSens:I(zPK2)				-0.011 (0.042) t = -0.256 p = 0.799	
ItemTypeDT:zIssuePol	-0.057 (0.041) t = -1.373 p = 0.170	-0.581*** (0.083) t = -7.002 p = 0.000	-0.041 (0.044) t = -0.927 p = 0.354	-0.090 (0.053) t = -1.684 p = 0.093	0.061 (0.066) t = 0.930 p = 0.353
ItemTypeCF:zIssuePol	-0.081*** (0.017) t = -4.734 p = 0.00001	-0.357*** (0.050) t = -7.072 p = 0.000	-0.066*** (0.018) t = -3.734 p = 0.0002	-0.078** (0.025) t = -3.111 p = 0.002	-0.064** (0.023) t = -2.862 p = 0.005
ItemTypeCT:zIssuePol	-0.165*** (0.050) t = -3.304 p = 0.001	-0.669*** (0.086) t = -7.824 p = 0.000	-0.188*** (0.051) t = -3.652 p = 0.0003	-0.148* (0.058) t = -2.553 p = 0.011	-0.108 (0.068) t = -1.599 p = 0.110
ItemTypeSens:zIssuePol	-0.195*** (0.046) t = -4.279 p = 0.00002	-0.682*** (0.085) t = -8.072 p = 0.000	-0.208*** (0.047) t = -4.450 p = 0.00001	-0.198*** (0.055) t = -3.571 p = 0.0004	-0.150* (0.067) t = -2.222 p = 0.027

ItemTypeDT:I(zIssuePol2)				-0.002 (0.037) t = -0.050 p = 0.960	
ItemTypeCF:I(zIssuePol2)				-0.002 (0.014) t = -0.135 p = 0.893	
ItemTypeCT:I(zIssuePol2)				-0.050 (0.045) t = -1.120 p = 0.263	
ItemTypeSens:I(zIssuePol2)				-0.023 (0.039) t = -0.595 p = 0.552	
ItemTypeDT:zOutThermom	-0.005 (0.042) t = -0.117 p = 0.907	-0.132* (0.056) t = -2.362 p = 0.019	-0.008 (0.043) t = -0.178 p = 0.860	0.080 (0.046) t = 1.741 p = 0.082	0.095 (0.079) t = 1.212 p = 0.226
ItemTypeCF:zOutThermom	-0.059** (0.018) t = -3.203 p = 0.002	-0.139*** (0.037) t = -3.780 p = 0.0002	-0.075*** (0.019) t = -3.955 p = 0.0001	-0.072*** (0.022) t = -3.308 p = 0.001	-0.081** (0.027) t = -2.969 p = 0.003
ItemTypeCT:zOutThermom	-0.094* (0.042) t = -2.231 p = 0.026	-0.187*** (0.056) t = -3.313 p = 0.001	-0.103* (0.044) t = -2.339 p = 0.020	-0.062 (0.050) t = -1.250 p = 0.212	-0.089 (0.077) t = -1.149 p = 0.251
ItemTypeSens:zOutThermom	-0.106** (0.041) t = -2.622	-0.197*** (0.055) t = -3.560	-0.127** (0.042) t = -3.044	-0.077 (0.046) t = -1.664	-0.019 (0.072) t = -0.262

	p = 0.009	p = 0.0004	p = 0.003	p = 0.097	p = 0.794
ItemTypeDT:I(zOutThermom2)				0.189***	
				(0.038)	
				t = 4.946	
				p = 0.00000	
ItemTypeCF:I(zOutThermom2)				-0.012	
				(0.018)	
				t = -0.679	
				p = 0.498	
ItemTypeCT:I(zOutThermom2)				0.098*	
				(0.038)	
				t = 2.569	
				p = 0.011	
ItemTypeSens:I(zOutThermom2)				0.095*	
				(0.037)	
				t = 2.560	
				p = 0.011	
ItemTypeDT:zPartisan	0.319***	0.448***	0.316***	0.318***	0.479***
	(0.040)	(0.064)	(0.039)	(0.040)	(0.069)
	t = 7.968	t = 6.968	t = 8.193	t = 7.996	t = 6.957
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
ItemTypeCF:zPartisan	0.110***	0.294***	0.112***	0.111***	0.109***
	(0.016)	(0.041)	(0.016)	(0.016)	(0.021)
	t = 6.907	t = 7.261	t = 7.030	t = 6.881	t = 5.168
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.00000
ItemTypeCT:zPartisan	0.097*	0.305***	0.102*	0.096*	0.154*
	(0.042)	(0.066)	(0.042)	(0.042)	(0.072)
	t = 2.309	t = 4.596	t = 2.432	t = 2.289	t = 2.131
	p = 0.021	p = 0.00001	p = 0.016	p = 0.023	p = 0.034
ItemTypeSens:zPartisan	0.151***	0.339***	0.153***	0.152***	0.152*
	(0.040)	(0.066)	(0.039)	(0.040)	(0.065)
	t = 3.796	t = 5.119	t = 3.910	t = 3.777	t = 2.343

	p = 0.0002	p = 0.00000	p = 0.0001	p = 0.0002	p = 0.020
ItemTypeDT:zage	0.017	-0.227**	0.021	0.026	0.037
	(0.048)	(0.080)	(0.051)	(0.047)	(0.073)
	t = 0.354	t = -2.823	t = 0.416	t = 0.539	t = 0.501
	p = 0.724	p = 0.005	p = 0.678	p = 0.591	p = 0.617
ItemTypeCF:zage	-0.040*	-0.124**	-0.038	-0.040*	-0.043
	(0.019)	(0.046)	(0.020)	(0.019)	(0.024)
	t = -2.101	t = -2.719	t = -1.895	t = -2.154	t = -1.771
	p = 0.036	p = 0.007	p = 0.059	p = 0.032	p = 0.077
ItemTypeCT:zage	-0.038	-0.276***	-0.062	-0.036	0.005
	(0.052)	(0.082)	(0.056)	(0.052)	(0.082)
	t = -0.724	t = -3.372	t = -1.109	t = -0.693	t = 0.065
	p = 0.470	p = 0.001	p = 0.268	p = 0.489	p = 0.949
ItemTypeSens:zage	-0.131**	-0.321***	-0.115*	-0.127**	-0.054
	(0.047)	(0.081)	(0.050)	(0.047)	(0.071)
	t = -2.779	t = -3.947	t = -2.296	t = -2.723	t = -0.757
	p = 0.006	p = 0.0001	p = 0.022	p = 0.007	p = 0.449
ItemTypeDT:zfemale	-0.117**	0.053	-0.130**	-0.098*	-0.177**
	(0.039)	(0.065)	(0.041)	(0.040)	(0.064)
	t = -3.000	t = 0.817	t = -3.147	t = -2.471	t = -2.769
	p = 0.003	p = 0.414	p = 0.002	p = 0.014	p = 0.006
ItemTypeCF:zfemale	0.007	0.041	0.003	0.006	-0.014
	(0.016)	(0.039)	(0.017)	(0.016)	(0.021)
	t = 0.458	t = 1.052	t = 0.206	t = 0.395	t = -0.662
	p = 0.648	p = 0.293	p = 0.837	p = 0.693	p = 0.508
ItemTypeCT:zfemale	-0.073	0.104	-0.091*	-0.065	-0.093
	(0.042)	(0.067)	(0.045)	(0.042)	(0.065)

	t = -1.753	t = 1.551	t = -2.021	t = -1.556	t = -1.429
	p = 0.080	p = 0.121	p = 0.044	p = 0.120	p = 0.153
ItemTypeSens:zfemale	-0.056	0.086	-0.062	-0.049	-0.006
	(0.039)	(0.066)	(0.042)	(0.040)	(0.063)
	t = -1.425	t = 1.299	t = -1.478	t = -1.225	t = -0.096
	p = 0.155	p = 0.195	p = 0.140	p = 0.221	p = 0.924
ItemTypeDT:zwhite	-0.047	-0.172**	-0.070	-0.044	-0.024
	(0.045)	(0.064)	(0.048)	(0.045)	(0.082)
	t = -1.041	t = -2.701	t = -1.454	t = -0.985	t = -0.294
	p = 0.298	p = 0.007	p = 0.147	p = 0.325	p = 0.769
ItemTypeCF:zwhite	0.0004	-0.043	-0.012	-0.00005	-0.001
	(0.020)	(0.038)	(0.021)	(0.020)	(0.029)
	t = 0.019	t = -1.140	t = -0.585	t = -0.002	t = -0.049
	p = 0.985	p = 0.255	p = 0.559	p = 0.999	p = 0.962
ItemTypeCT:zwhite	-0.076	-0.185**	-0.094	-0.074	0.056
	(0.046)	(0.064)	(0.050)	(0.046)	(0.085)
	t = -1.649	t = -2.893	t = -1.897	t = -1.599	t = 0.658
	p = 0.100	p = 0.004	p = 0.058	p = 0.110	p = 0.511
ItemTypeSens:zwhite	-0.044	-0.160*	-0.059	-0.041	-0.064
	(0.042)	(0.063)	(0.045)	(0.042)	(0.075)
	t = -1.046	t = -2.535	t = -1.325	t = -0.967	t = -0.854
	p = 0.296	p = 0.012	p = 0.186	p = 0.334	p = 0.394
ItemTypeDT:zcollege	-0.039	-0.123	-0.018	-0.032	-0.066
	(0.040)	(0.065)	(0.043)	(0.040)	(0.064)
	t = -0.985	t = -1.899	t = -0.429	t = -0.805	t = -1.032
	p = 0.325	p = 0.058	p = 0.669	p = 0.421	p = 0.302

ItemTypeCF:zcollege	-0.017	-0.040	-0.025	-0.016	-0.001
	(0.017)	(0.039)	(0.018)	(0.017)	(0.022)
	t = -0.987	t = -1.025	t = -1.376	t = -0.947	t = -0.061
	p = 0.324	p = 0.306	p = 0.169	p = 0.344	p = 0.952
ItemTypeCT:zcollege	-0.104*	-0.162*	-0.074	-0.100*	-0.065
	(0.042)	(0.066)	(0.045)	(0.042)	(0.071)
	t = -2.463	t = -2.457	t = -1.647	t = -2.383	t = -0.919
	p = 0.014	p = 0.014	p = 0.100	p = 0.018	p = 0.359
ItemTypeSens:zcollege	-0.048	-0.128	-0.019	-0.044	-0.042
	(0.039)	(0.067)	(0.042)	(0.039)	(0.064)
	t = -1.220	t = -1.914	t = -0.450	t = -1.118	t = -0.656
	p = 0.223	p = 0.056	p = 0.653	p = 0.264	p = 0.513
ItemTypeDT:zScreenTot	-0.280***	-0.305***	-0.263***	-0.253***	
	(0.046)	(0.071)	(0.049)	(0.045)	
	t = -6.126	t = -4.302	t = -5.397	t = -5.595	
	p = 0.000	p = 0.00002	p = 0.00000	p = 0.00000	
ItemTypeCF:zScreenTot	0.016	0.016	0.015	0.014	
	(0.019)	(0.044)	(0.021)	(0.020)	
	t = 0.809	t = 0.358	t = 0.705	t = 0.725	
	p = 0.419	p = 0.721	p = 0.481	p = 0.469	
ItemTypeCT:zScreenTot	-0.244***	-0.261***	-0.206***	-0.233***	
	(0.050)	(0.072)	(0.055)	(0.050)	
	t = -4.858	t = -3.631	t = -3.733	t = -4.665	
	p = 0.00001	p = 0.0003	p = 0.0002	p = 0.00001	
ItemTypeSens:zScreenTot	-0.128**	-0.207**	-0.120*	-0.117**	
	(0.044)	(0.072)	(0.047)	(0.044)	
	t = -2.912	t = -2.882	t = -2.538	t = -2.639	

	p = 0.004	p = 0.004	p = 0.012	p = 0.009	
Constant	0.785***	2.649***	0.796***	0.961***	1.054***
	(0.031)	(0.074)	(0.033)	(0.054)	(0.049)
	t = 25.692	t = 35.903	t = 23.824	t = 17.959	t = 21.522
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

Regression predicting item-type flag count by item type, political knowledge, issue polarization, out-party dislike, and standard controls, and all interactions between covariates and item type; with clustered SEs by participant

Table S24. *Item-type flag count predicted by item type and political knowledge, issue polarization, and out-party dislike. Results indicate similar findings as shown and described in Tables S21-S23. Political knowledge is associated with increased flagging of both discordant and concordant false headlines relative to other headline types. Issue polarization and out-party dislike are associated with increased flagging of discordant false headlines relative to other headline types in additive space. Proportionally, issue polarization and out-party dislike are associated similarly with increased flagging of both discordant true and false headlines.*

d. Supplemental Figures

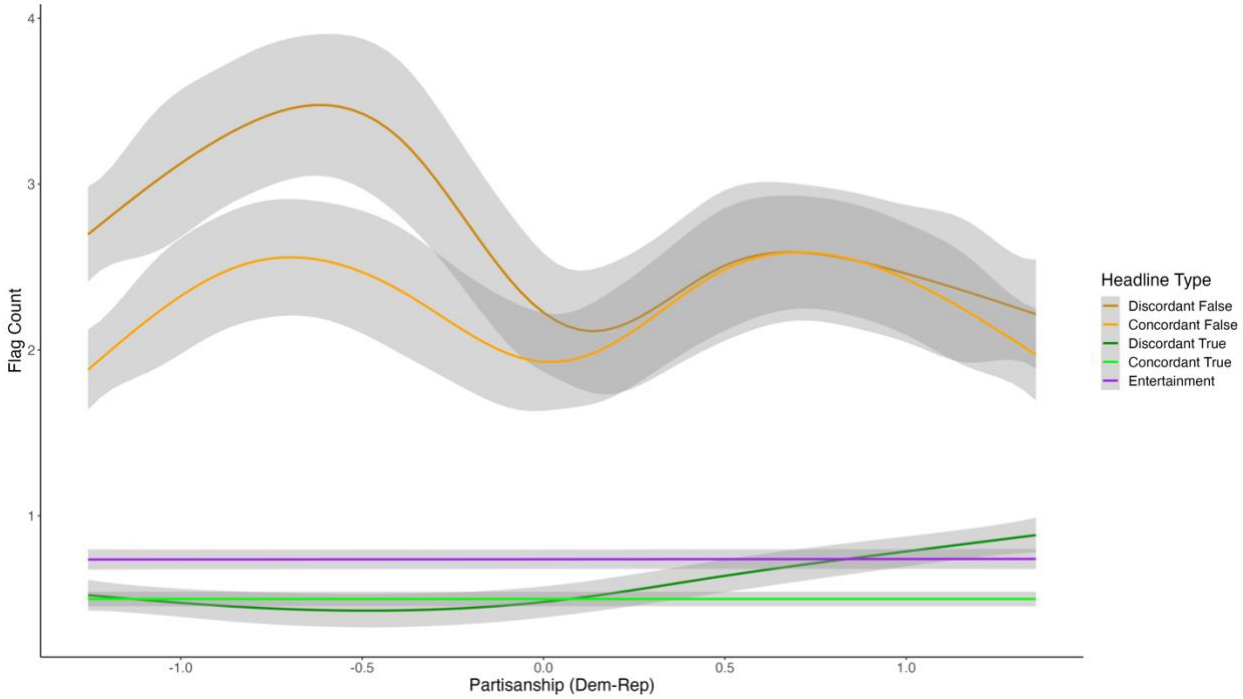

Figure S2. Republican-leaning survey participants contribute a lower proportion of discordant flags, but also a greater proportion of incorrect flags. Shown are the predicted number of flags by post type and political motivation. Error bars reflect 95% confidence intervals.

3. Study 2: Field Study Supplement

a. Fact-checker Veracity Evaluation Survey

To estimate a "ground truth" for tweets flagged by Community Notes users, we sent a subset of 461 original tweets (collected by sampling all tweets with at least three flags, and then randomly sampling 300 additional tweets with at least one note) to two professional fact-checkers recruited via Upwork. The fact-checkers were told they were to research a series of tweets. For each tweet, the fact-checkers were asked to assess whether the tweet was "Not Misleading" or "Misinformed or Potentially Misleading" (same wording as on Community Notes). Fact-checkers also evaluated each tweet on a series of other characteristics, including harm, unbiasedness, objectivity, whether it describes an event that actually happened, accuracy reliability, trustworthiness, and truth (Allen et al., 2021b). Fact-checkers were also asked to paste the URLs to any sources they used while researching each tweet. Both fact-checkers evaluated all 461 tweets. Our full fact-checker evaluation survey is available here: https://osf.io/3ngbt/.

b. Mechanical Turk Partisan Favorability Survey

To assess the partisan favorability of our 461 fact-checker evaluated tweets, we recruited Amazon Mechanical Turk workers (N=355; mean age=39.91; 124 female, 226 male, 5 reported another gender or did not respond to this question) to evaluate these original tweet texts on several dimensions. Each rater evaluated 30 unique tweets. On average, each tweet received 18.68 ratings (SD=4.14; min=9, max=33, median=19).

Raters first completed a trivial attention captcha – those who failed were excluded from continuing the task. Raters then answered demographic questions (age, gender, race) and political ideology and identity measures. Next, participants were told they would see a series of 30 tweets. For each tweet, they were told they would be asked (i) what category is the main topic of the tweet?; (ii) how favorable is the tweet to Democrats versus Republicans, assuming the content of the tweet is entirely accurate?; and (iii) how controversial is the main topic of the tweet? Participants were told that if any of those questions were unclear from the tweet text alone, they should click the included hyperlinks (to the original tweets) and open the tweets in a new browser to view their full content and context.

Tweet Topic. Participants were first asked: "What category best describes the main topic of the above tweet?" followed by choices: Politics, Sports, Celebrity news, Science/technology, Business, Health/COVID-19, Other (please specify).

Tweet Partisan Favorability. Then, participants were asked: "Assuming the tweet is entirely accurate, how favorable would it be to Democrats versus Republicans?" followed by choices: 1=More favorable for Democrats, 2=Moderately more favorable for Democrats, 3=Slightly more favorable for Democrats, 4=Neither/equally favorable for Democrats and Republicans, 5=Slightly more favorable for Republicans, 6=Moderately more favorable for Republicans, 7=More favorable for Republicans.

Tweet Controversy. Finally, participants were asked: "How controversial is the main topic of the above tweet?" followed by choices: 1=Not at all controversial, 2=Slightly controversial, 3=Somewhat controversial, 4=Moderately controversial, 5=Controversial, 6=Very controversial, 7=Extremely controversial.

As a measure of interrater reliability, we computed intraclass correlations (ICC) for partisan favorability and controversy. We assessed a one-way random effects ICC model, since each tweet was assessed by a different set of randomly selected raters; and average ratings were used as the final evaluation criterion. We observed high ICC(1,k) scores for both partisan favorability (ICC=0.885, 95% CI: [0.869, 0.899], $p<.001$) and controversy (ICC=0.790, 95% CI: [0.761, 0.817], $p<.001$). We observed lower agreement for tweet category categorization (assessed via Krippendorff's alpha = 0.357); however, this measure was just exploratory. The average partisan favorability of our tweet set is 3.96 (*SD*=1.00). The average controversy of our tweet set is 4.00 (0.84). Modal tweet categories are as follows: 344 Politics (74.6%), 4 Sports (0.87%), 12 Celebrity news (2.60%), 15 Science/technology (3.25%), 17 Business (3.69%, 57 Health/Covid-19 (12.36%), 12 Other (2.60%).

Our full Mechanical Turk evaluation survey is available here: https://osf.io/3ngbt/.

c. Regression Tables

Analysis plan. Our field study analyses were not preregistered – however, we attempted to match our analyses as closely as possible to our preregistered survey study analyses for comparability purposes. We first conduct both a quasi-Poisson regression and OLS regression of note count (number of total notes written by each individual) predicted by inferred partisanship (Barberá et al., 2015), political extremity (absolute value of difference between partisanship score and 0.5, where scores less than 0.5 reflect pro-Democratic party affinity and scores greater than 0.5 reflect pro-Republican affinity, then piecewise scaled 0 to 1 by partisanship, and *z*-scored), and control variables follower count, friends count, statuses count, inferred gender, inferred age, feed quality score, misinformation-exposure score, and toxicity (all controls *z*-scores; imputed means for missing toxicity, feed quality, and misinformation-score values) (Lin et al., 2023; Mosleh & Rand, 2022; *Perspective API - How It Works*, n.d.; Wang et al., 2019); with HC2 robust standard errors. Our key prediction was that more politically extreme users would write more notes overall. As robustness checks, we also conduct similar analyses except (i) including a quadratic political extremity term, and (ii) predicting note count via a Heckman 2-step model, respectively. Second, for original tweets evaluated by Mechanical Turk workers, we evaluate the proportion of notes written by individuals on politically discordant tweets (where discordance is evaluated as a mismatch between binarized user partisanship and binarized tweet average Republican favorability) by the same predictors as the previous analysis, with analytic weighting by total note count and HC2 robust standard errors. To test for nonlinearities, we also conduct this analysis with a quadratic term for political extremity. Our main prediction was that more politically extreme users would write a higher proportion of notes on politically discordant tweets. Third, we conduct nearly identical analyses, except predicting the proportion of notes written that agree with at least one professional fact-checker (i.e., at least one fact-checker evaluated the original tweet as potentially misleading or misinformed). Our prediction was that more politically extreme users would write a higher proportion of notes agreeing with fact-checkers; or at least would not write a significantly lower proportion of notes agreeing with fact-checkers. Fourth, we conduct both a

quasi-Poisson regression (for multiplicative differences) and an OLS regression (for additive differences) to predict the number of notes users wrote on each tweet type by tweet type (6-level factor, baseline = Discordant-False), political extremity, inferred partisanship, control variables, and interactions between tweet type and all other predictors, clustering standard errors by user. We also conduct this analysis including a quadratic term for political extremity. Our key prediction was that there would be negative interaction terms between political extremity and tweet type factors dummies, such that more politically extreme users wrote more notes on Discordant-False tweets in particular.

All analysis code is available here: https://osf.io/3ngbt/.

i. Flag Count

	Dependent variable:		
	Note count		
	QP Model	OLS Model	QP With Quadratic
	(1)	(2)	(3)
polextrem	0.369***	0.453***	0.368***
	(0.095)	(0.129)	(0.091)
	t = 3.880	t = 3.506	t = 4.029
	p = 0.0002	p = 0.0005	p = 0.0001
I(polextrem2)			0.005
			(0.072)
			t = 0.069
			p = 0.946
politics	0.222***	0.303**	0.222**
	(0.067)	(0.110)	(0.068)
	t = 3.298	t = 2.749	t = 3.258
	p = 0.001	p = 0.006	p = 0.002
followers_count	0.060	0.154	0.060
	(0.042)	(0.194)	(0.042)
	t = 1.431	t = 0.794	t = 1.435
	p = 0.153	p = 0.428	p = 0.152
friends_count	0.006	0.017	0.006
	(0.051)	(0.218)	(0.051)
	t = 0.115	t = 0.080	t = 0.116
	p = 0.909	p = 0.937	p = 0.908
statuses_count	0.102**	0.217*	0.102**
	(0.035)	(0.099)	(0.036)

	t = 2.902	t = 2.185	t = 2.872
	p = 0.004	p = 0.029	p = 0.005
is_female_agg	-0.016	-0.025	-0.016
	(0.089)	(0.111)	(0.090)
	t = -0.176	t = -0.228	t = -0.177
	p = 0.861	p = 0.820	p = 0.860
age_agg30-39	-0.060	-0.046	-0.060
	(0.196)	(0.237)	(0.195)
	t = -0.305	t = -0.192	t = -0.307
	p = 0.761	p = 0.848	p = 0.759
age_agg19-29	0.031	0.070	0.030
	(0.187)	(0.236)	(0.185)
	t = 0.164	t = 0.297	t = 0.161
	p = 0.871	p = 0.767	p = 0.872
age_agg< =18	0.103	0.172	0.102
	(0.267)	(0.343)	(0.273)
	t = 0.386	t = 0.501	t = 0.373
	p = 0.700	p = 0.617	p = 0.710
domain_quality_score	0.256**	0.286**	0.256**
	(0.092)	(0.106)	(0.092)
	t = 2.772	t = 2.686	t = 2.771
	p = 0.006	p = 0.008	p = 0.006
elite_politifact_score	0.158*	0.212*	0.158*
	(0.081)	(0.108)	(0.078)
	t = 1.962	t = 1.970	t = 2.013
	p = 0.050	p = 0.049	p = 0.045
toxic_score	0.162	0.200	0.162
	(0.118)	(0.161)	(0.117)
	t = 1.374	t = 1.239	t = 1.379
	p = 0.170	p = 0.216	p = 0.168
Constant	0.081	1.188***	0.077
	(0.140)	(0.159)	(0.161)
	t = 0.578	t = 7.463	t = 0.477
	p = 0.564	p = 0.000	p = 0.634

Note: *p**p***p<0.001

Regression predicting note count by political extremity, partisanship, and controls, with HC2 robust SEs

Table S25. *Flag count predicted by political extremity and standard controls. Across specifications, we find that more politically extreme users contribute more flags.*

	Dependent variable:	
	AnyFlag	FlagCount
	probit	*OLS*
	(1)	(2)
(Intercept)	-0.587***	2.558
	(0.044)	(21.101)
	t = -13.408	t = 0.121
	p = 0.000	p = 0.904
polextrem	0.109***	1.158
	(0.026)	(1.471)
	t = 4.109	t = 0.787
	p = 0.00004	p = 0.432
politics	0.057*	0.838
	(0.024)	(0.787)
	t = 2.403	t = 1.065
	p = 0.017	p = 0.288
followers_count	-0.019	0.617
	(0.031)	(0.401)
	t = -0.621	t = 1.541
	p = 0.535	p = 0.124
friends_count	0.095	-0.262
	(0.065)	(0.612)
	t = 1.462	t = -0.428
	p = 0.144	p = 0.669
statuses_count	0.110***	0.118
	(0.024)	(1.326)
	t = 4.661	t = 0.089
	p = 0.00001	p = 0.930
is_female_agg	-0.042	0.119
	(0.024)	(0.615)
	t = -1.775	t = 0.193
	p = 0.076	p = 0.847
age_agg30-39	-0.005	-0.153
	(0.063)	(0.837)
	t = -0.080	t = -0.183
	p = 0.937	p = 0.855

age_agg19-29	0.034	-0.063
	(0.061)	(0.897)
	t = 0.553	t = -0.070
	p = 0.581	p = 0.945
age_agg< =18	0.033	0.698
	(0.075)	(1.061)
	t = 0.442	t = 0.658
	p = 0.659	p = 0.511
domain_quality_score	0.047	0.986
	(0.024)	(0.677)
	t = 1.958	t = 1.457
	p = 0.051	p = 0.146
elite_politifact_score	0.055	0.525
	(0.030)	(0.841)
	t = 1.827	t = 0.624
	p = 0.068	p = 0.533
toxic_score	0.087***	0.408
	(0.023)	(1.147)
	t = 3.746	t = 0.356
	p = 0.0002	p = 0.722
IMR1		1.228
		(17.490)
		t = 0.070
		p = 0.945

Note: *p**p***p<0.001

2-step Heckman model predicting note count (any note; note count) by political extremity, partisanship, and controls

Table S26. *Flag count predicted by political extremity and standard controls via a 2-step Heckman model, modeling a two-stage decision process (any participation in community notes; number of notes conditional on participation). This analysis suggests that on Community Notes, more politically extreme users are more likely to participate at all on the platform (via writing a note) – then, conditional on participation, we do not see a significant relationship between political extremity and note count (perhaps due to overall low contribution levels).*

ii. Discordant Flagging

	Dependent variable:	
	Proportion discordant notes	
	OLS Model	With Quadratic
	(1)	(2)
polextrem	0.051*	0.050*
	(0.020)	(0.021)
	t = 2.564	t = 2.415
	p = 0.011	p = 0.016
I(polextrem2)		0.007
		(0.015)
		t = 0.434
		p = 0.665
politics	0.026*	0.025
	(0.013)	(0.013)
	t = 2.059	t = 1.871
	p = 0.040	p = 0.062
followers_count	0.010	0.010
	(0.015)	(0.015)
	t = 0.693	t = 0.707
	p = 0.489	p = 0.480
friends_count	-0.014	-0.016
	(0.038)	(0.038)
	t = -0.356	t = -0.410
	p = 0.723	p = 0.682
statuses_count	0.002	0.002
	(0.006)	(0.007)
	t = 0.309	t = 0.351
	p = 0.758	p = 0.726
is_female_agg	0.016	0.016
	(0.013)	(0.013)
	t = 1.220	t = 1.231
	p = 0.223	p = 0.219

age_agg30-39	-0.017	-0.017
	(0.041)	(0.042)
	t = -0.409	t = -0.416
	p = 0.683	p = 0.678
age_agg19-29	-0.038	-0.040
	(0.044)	(0.045)
	t = -0.867	t = -0.884
	p = 0.387	p = 0.377
age_agg< =18	-0.048	-0.048
	(0.048)	(0.048)
	t = -0.996	t = -0.998
	p = 0.320	p = 0.319
domain_quality_score	0.020	0.020
	(0.017)	(0.017)
	t = 1.196	t = 1.134
	p = 0.232	p = 0.257
elite_politifact_score	0.023	0.021
	(0.017)	(0.017)
	t = 1.377	t = 1.248
	p = 0.169	p = 0.213
toxic_score	0.018	0.017
	(0.017)	(0.016)
	t = 1.081	t = 1.058
	p = 0.280	p = 0.290
Constant	0.826***	0.822***
	(0.033)	(0.030)
	t = 24.716	t = 27.043
	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion discordant notes by political extremity, partisanship, and controls; with analytic weighting by note count and HC2 robust SEs

Table S27. *Proportion of flags written on politically discordant tweets predicted by political extremity and standard controls. Across specifications, we find that more politically extreme users write a greater proportion of flags on politically discordant tweets.*

iii. Flagging Quality

	Dependent variable:	
	Proportion notes agreeing with FCers	
	OLS Model	With Quadratic
	(1)	(2)
polextrem	-0.021	-0.024
	(0.014)	(0.014)
	t = -1.505	t = -1.726
	p = 0.133	p = 0.085
I(polextrem2)		0.023*
		(0.011)
		t = 1.967
		p = 0.050
politics	-0.066***	-0.069***
	(0.010)	(0.011)
	t = -6.672	t = -6.381
	p = 0.000	p = 0.000
followers_count	-0.005	-0.004
	(0.007)	(0.008)
	t = -0.669	t = -0.555
	p = 0.504	p = 0.580
friends_count	-0.016	-0.023*
	(0.010)	(0.011)
	t = -1.629	t = -2.057
	p = 0.104	p = 0.040
statuses_count	0.016	0.017
	(0.010)	(0.010)
	t = 1.670	t = 1.792
	p = 0.096	p = 0.074
is_female_agg	-0.037**	-0.036*
	(0.014)	(0.014)
	t = -2.598	t = -2.558
	p = 0.010	p = 0.011
age_agg30-39	0.014	0.012

	(0.034)	(0.034)
	t = 0.425	t = 0.365
	p = 0.671	p = 0.716
age_agg19-29	-0.003	-0.009
	(0.040)	(0.038)
	t = -0.071	t = -0.231
	p = 0.944	p = 0.818
age_agg< =18	-0.010	-0.010
	(0.043)	(0.045)
	t = -0.237	t = -0.226
	p = 0.813	p = 0.822
domain_quality_score	0.001	-0.001
	(0.016)	(0.016)
	t = 0.044	t = -0.087
	p = 0.966	p = 0.932
elite_politifact_score	0.029	0.022
	(0.016)	(0.017)
	t = 1.747	t = 1.292
	p = 0.081	p = 0.197
toxic_score	-0.015	-0.018
	(0.014)	(0.013)
	t = -1.075	t = -1.329
	p = 0.283	p = 0.184
Constant	0.833***	0.817***
	(0.027)	(0.029)
	t = 30.413	t = 28.448
	p = 0.000	p = 0.000

Note: *p **p ***p<0.001

OLS regression predicting proportion notes agreeing with FCers by political extremity, partisanship, and controls; with analytic weighting by note count and HC2 robust SEs

Table S28. *Proportion of notes agreeing with fact-checker evaluations (i.e., true-positive, or 'correct,' notes) predicted by political extremity and standard controls. Across specifications, we do not find evidence for a positive or negative relationship between political extremity and proportion agreement with fact-checkers. We do observe a significant relationship between quadratic political extremity and agreement with fact-checkers – such that at intermediate levels, political extremity is associated with an increase in agreement with fact-checkers; but at extreme levels, political extremity is associated with a decrease in agreement with fact-checkers.*

iv. Flag Count by Tweet Type

	Dependent variable:		
	Item-type note count		
	QP Model	OLS Model	QP With Quadratic
	(1)	(2)	(3)
ItemTypeFalse Concord	-1.716***	-0.706***	-1.724***
	(0.215)	(0.102)	(0.222)
	t = -7.993	t = -6.912	t = -7.778
	p = 0.000	p = 0.000	p = 0.000
ItemTypeFalse NA	-19.061***	-0.850***	-19.102***
	(0.135)	(0.103)	(0.163)
	t = -141.516	t = -8.220	t = -117.137
	p = 0.000	p = 0.000	p = 0.000
ItemTypeTrue Discord	-2.021***	-0.652***	-1.902***
	(0.203)	(0.075)	(0.222)
	t = -9.979	t = -8.721	t = -8.563
	p = 0.000	p = 0.000	p = 0.000
ItemTypeTrue Concord	-3.166***	-0.789***	-2.920***
	(0.311)	(0.097)	(0.330)
	t = -10.191	t = -8.146	t = -8.841
	p = 0.000	p = 0.000	p = 0.000
ItemTypeTrue NA	-19.061***	-0.850***	-19.102***
	(0.135)	(0.103)	(0.163)
	t = -141.516	t = -8.220	t = -117.137
	p = 0.000	p = 0.000	p = 0.000
polextrem	0.366***	0.320***	0.383***
	(0.088)	(0.091)	(0.083)
	t = 4.158	t = 3.497	t = 4.614
	p = 0.00004	p = 0.0005	p = 0.00001
I(polextrem2)			-0.050
			(0.066)
			t = -0.757
			p = 0.450
politics	0.235***	0.241**	0.238***
	(0.067)	(0.085)	(0.066)

	t = 3.524	t = 2.823	t = 3.594
	p = 0.0005	p = 0.005	p = 0.0004
followers_count	0.105***	0.115**	0.104***
	(0.031)	(0.039)	(0.031)
	t = 3.351	t = 2.951	t = 3.323
	p = 0.001	p = 0.004	p = 0.001
friends_count	-0.170*	-0.075***	-0.160*
	(0.071)	(0.022)	(0.074)
	t = -2.406	t = -3.452	t = -2.176
	p = 0.017	p = 0.001	p = 0.030
statuses_count	0.048	0.039	0.046
	(0.041)	(0.052)	(0.041)
	t = 1.154	t = 0.748	t = 1.125
	p = 0.249	p = 0.455	p = 0.261
is_female_agg	0.075	0.066	0.077
	(0.073)	(0.063)	(0.074)
	t = 1.024	t = 1.038	t = 1.046
	p = 0.306	p = 0.300	p = 0.296
age_agg30-39	-0.020	-0.018	-0.025
	(0.180)	(0.163)	(0.183)
	t = -0.109	t = -0.112	t = -0.136
	p = 0.914	p = 0.911	p = 0.892
age_agg19-29	-0.253	-0.195	-0.250
	(0.149)	(0.125)	(0.148)
	t = -1.696	t = -1.561	t = -1.691
	p = 0.090	p = 0.119	p = 0.091
age_agg< =18	-0.021	0.007	-0.011
	(0.183)	(0.161)	(0.181)
	t = -0.113	t = 0.046	t = -0.061
	p = 0.910	p = 0.964	p = 0.952
domain_quality_score	0.142*	0.111*	0.141*
	(0.065)	(0.053)	(0.065)
	t = 2.196	t = 2.113	t = 2.164
	p = 0.029	p = 0.035	p = 0.031
elite_politifact_score	0.083	0.100	0.090
	(0.085)	(0.086)	(0.081)

	t = 0.976	t = 1.167	t = 1.112
	p = 0.330	p = 0.244	p = 0.267
toxic_score	0.020	0.016	0.020
	(0.062)	(0.050)	(0.062)
	t = 0.331	t = 0.317	t = 0.321
	p = 0.741	p = 0.751	p = 0.749
ItemTypeFalse Concord:polextrem	-0.280**	-0.307***	-0.299**
	(0.106)	(0.086)	(0.111)
	t = -2.637	t = -3.573	t = -2.703
	p = 0.009	p = 0.0004	p = 0.007
ItemTypeFalse NA:polextrem	-0.366***	-0.320***	-0.383***
	(0.095)	(0.091)	(0.092)
	t = -3.842	t = -3.497	t = -4.168
	p = 0.0002	p = 0.0005	p = 0.00004
ItemTypeTrue Discord:polextrem	-0.020	-0.174**	0.140
	(0.107)	(0.054)	(0.146)
	t = -0.191	t = -3.234	t = 0.960
	p = 0.849	p = 0.002	p = 0.338
ItemTypeTrue Concord:polextrem	-0.358*	-0.315***	-0.419*
	(0.145)	(0.091)	(0.197)
	t = -2.473	t = -3.485	t = -2.129
	p = 0.014	p = 0.0005	p = 0.034
ItemTypeTrue NA:polextrem	-0.366***	-0.320***	-0.383***
	(0.095)	(0.091)	(0.092)
	t = -3.842	t = -3.497	t = -4.168
	p = 0.0002	p = 0.0005	p = 0.00004
ItemTypeFalse Concord:I(polextrem2)			0.015
			(0.089)
			t = 0.168
			p = 0.867
ItemTypeFalse NA:I(polextrem2)			0.050
			(0.072)
			t = 0.693
			p = 0.489
ItemTypeTrue Discord:I(polextrem2)			-0.216

			(0.120) t = -1.798 p = 0.073
ItemTypeTrue Concord:I(polextrem2)			-0.272
			(0.156) t = -1.738 p = 0.083
ItemTypeTrue NA:I(polextrem2)			0.050
			(0.072) t = 0.693 p = 0.489
ItemTypeFalse Concord:politics	-0.123	-0.224**	-0.126
	(0.085)	(0.080)	(0.086)
	t = -1.437	t = -2.808	t = -1.464
	p = 0.151	p = 0.005	p = 0.144
ItemTypeFalse NA:politics	-0.235**	-0.241**	-0.238***
	(0.072)	(0.085)	(0.072)
	t = -3.250	t = -2.823	t = -3.299
	p = 0.002	p = 0.005	p = 0.001
ItemTypeTrue Discord:politics	0.589***	-0.055	0.626***
	(0.089)	(0.049)	(0.093)
	t = 6.595	t = -1.131	t = 6.706
	p = 0.000	p = 0.259	p = 0.000
ItemTypeTrue Concord:politics	0.157	-0.218**	0.183
	(0.144)	(0.084)	(0.158)
	t = 1.089	t = -2.598	t = 1.157
	p = 0.277	p = 0.010	p = 0.248
ItemTypeTrue NA:politics	-0.235**	-0.241**	-0.238***
	(0.072)	(0.085)	(0.072)
	t = -3.250	t = -2.823	t = -3.299
	p = 0.002	p = 0.005	p = 0.001
ItemTypeFalse Concord:followers_count	-0.082	-0.112**	-0.081
	(0.062)	(0.036)	(0.062)
	t = -1.314	t = -3.117	t = -1.304
	p = 0.189	p = 0.002	p = 0.193

ItemTypeFalse NA:followers_count	-0.105*	-0.115**	-0.104*
	(0.049)	(0.039)	(0.049)
	t = -2.137	t = -2.951	t = -2.122
	p = 0.033	p = 0.004	p = 0.034
ItemTypeTrue Discord:followers_count	0.059	-0.087**	0.068
	(0.055)	(0.028)	(0.065)
	t = 1.088	t = -3.173	t = 1.051
	p = 0.277	p = 0.002	p = 0.294
ItemTypeTrue Concord:followers_count	-4.998*	-0.120**	-4.680*
	(2.537)	(0.038)	(2.345)
	t = -1.970	t = -3.195	t = -1.996
	p = 0.049	p = 0.002	p = 0.046
ItemTypeTrue NA:followers_count	-0.105*	-0.115**	-0.104*
	(0.049)	(0.039)	(0.049)
	t = -2.137	t = -2.951	t = -2.122
	p = 0.033	p = 0.004	p = 0.034
ItemTypeFalse Concord:friends_count	0.110	0.071***	0.105
	(0.103)	(0.019)	(0.095)
	t = 1.064	t = 3.668	t = 1.102
	p = 0.288	p = 0.0003	p = 0.271
ItemTypeFalse NA:friends_count	0.170*	0.075***	0.160*
	(0.070)	(0.022)	(0.073)
	t = 2.433	t = 3.452	t = 2.208
	p = 0.015	p = 0.001	p = 0.028
ItemTypeTrue Discord:friends_count	0.077	0.062***	0.077
	(0.067)	(0.017)	(0.073)
	t = 1.147	t = 3.686	t = 1.061
	p = 0.252	p = 0.0003	p = 0.289
ItemTypeTrue Concord:friends_count	0.858	0.077***	0.769
	(0.744)	(0.021)	(0.748)
	t = 1.154	t = 3.660	t = 1.027

	p = 0.249	p = 0.0003	p = 0.305
ItemTypeTrue NA:friends_count	0.170*	0.075***	0.160*
	(0.070)	(0.022)	(0.073)
	t = 2.433	t = 3.452	t = 2.208
	p = 0.015	p = 0.001	p = 0.028
ItemTypeFalse Concord:statuses_count	-0.029	-0.036	-0.028
	(0.048)	(0.047)	(0.048)
	t = -0.602	t = -0.772	t = -0.583
	p = 0.548	p = 0.440	p = 0.560
ItemTypeFalse NA:statuses_count	-0.048	-0.039	-0.046
	(0.052)	(0.052)	(0.052)
	t = -0.920	t = -0.748	t = -0.893
	p = 0.358	p = 0.455	p = 0.373
ItemTypeTrue Discord:statuses_count	-0.061	-0.045	-0.079
	(0.095)	(0.036)	(0.101)
	t = -0.641	t = -1.234	t = -0.787
	p = 0.522	p = 0.218	p = 0.432
ItemTypeTrue Concord:statuses_count	0.045	-0.040	0.036
	(0.094)	(0.052)	(0.097)
	t = 0.479	t = -0.771	t = 0.372
	p = 0.632	p = 0.441	p = 0.710
ItemTypeTrue NA:statuses_count	-0.048	-0.039	-0.046
	(0.052)	(0.052)	(0.052)
	t = -0.920	t = -0.748	t = -0.893
	p = 0.358	p = 0.455	p = 0.373
ItemTypeFalse Concord:is_female_agg	-0.252	-0.088	-0.252
	(0.137)	(0.064)	(0.138)
	t = -1.839	t = -1.370	t = -1.836
	p = 0.066	p = 0.171	p = 0.067
ItemTypeFalse NA:is_female_agg	-0.075	-0.066	-0.077
	(0.081)	(0.063)	(0.082)
	t = -0.921	t = -1.038	t = -0.941
	p = 0.357	p = 0.300	p = 0.347

ItemTypeTrue Discord:is_female_agg	0.209*	-0.005	0.215*
	(0.106)	(0.042)	(0.106)
	t = 1.977	t = -0.131	t = 2.035
	p = 0.049	p = 0.896	p = 0.042
ItemTypeTrue Concord:is_female_agg	0.317*	-0.039	0.333*
	(0.134)	(0.056)	(0.136)
	t = 2.366	t = -0.701	t = 2.451
	p = 0.018	p = 0.484	p = 0.015
ItemTypeTrue NA:is_female_agg	-0.075	-0.066	-0.077
	(0.081)	(0.063)	(0.082)
	t = -0.921	t = -1.038	t = -0.941
	p = 0.357	p = 0.300	p = 0.347
ItemTypeFalse Concord:age_agg30-39	-0.123	-0.002	-0.121
	(0.280)	(0.157)	(0.279)
	t = -0.441	t = -0.010	t = -0.433
	p = 0.660	p = 0.992	p = 0.666
ItemTypeFalse NA:age_agg30-39	0.020	0.018	0.025
	(0.201)	(0.163)	(0.204)
	t = 0.098	t = 0.112	t = 0.123
	p = 0.923	p = 0.911	p = 0.903
ItemTypeTrue Discord:age_agg30-39	-0.132	-0.005	-0.156
	(0.303)	(0.115)	(0.307)
	t = -0.437	t = -0.043	t = -0.508
	p = 0.662	p = 0.966	p = 0.612
ItemTypeTrue Concord:age_agg30-39	0.073	0.027	0.042
	(0.380)	(0.157)	(0.377)
	t = 0.193	t = 0.169	t = 0.113
	p = 0.847	p = 0.866	p = 0.911
ItemTypeTrue NA:age_agg30-39	0.020	0.018	0.025
	(0.201)	(0.163)	(0.204)
	t = 0.098	t = 0.112	t = 0.123
	p = 0.923	p = 0.911	p = 0.903

ItemTypeFalse Concord:age_agg19-29	0.185	0.186	0.185
	(0.298)	(0.125)	(0.298)
	t = 0.623	t = 1.486	t = 0.619
	p = 0.534	p = 0.138	p = 0.537
ItemTypeFalse NA:age_agg19-29	0.253	0.195	0.250
	(0.172)	(0.125)	(0.171)
	t = 1.471	t = 1.561	t = 1.466
	p = 0.142	p = 0.119	p = 0.143
ItemTypeTrue Discord:age_agg19-29	0.129	0.176	0.132
	(0.257)	(0.096)	(0.249)
	t = 0.501	t = 1.831	t = 0.528
	p = 0.617	p = 0.068	p = 0.598
ItemTypeTrue Concord:age_agg19-29	-0.004	0.184	0.017
	(0.418)	(0.117)	(0.410)
	t = -0.011	t = 1.573	t = 0.041
	p = 0.992	p = 0.116	p = 0.968
ItemTypeTrue NA:age_agg19-29	0.253	0.195	0.250
	(0.172)	(0.125)	(0.171)
	t = 1.471	t = 1.561	t = 1.466
	p = 0.142	p = 0.119	p = 0.143
ItemTypeFalse Concord:age_agg< =18	0.250	0.030	0.246
	(0.337)	(0.155)	(0.339)
	t = 0.743	t = 0.194	t = 0.727
	p = 0.458	p = 0.846	p = 0.468
ItemTypeFalse NA:age_agg< =18	0.021	-0.007	0.011
	(0.213)	(0.161)	(0.211)
	t = 0.097	t = -0.046	t = 0.053
	p = 0.923	p = 0.964	p = 0.958
ItemTypeTrue Discord:age_agg< =18	0.025	0.003	0.037
	(0.269)	(0.135)	(0.269)
	t = 0.092	t = 0.024	t = 0.136
	p = 0.927	p = 0.981	p = 0.892

ItemTypeTrue Concord:age_agg< =18	-0.094	-0.010	-0.084
	(0.475)	(0.158)	(0.480)
	t = -0.198	t = -0.062	t = -0.175
	p = 0.843	p = 0.951	p = 0.862
ItemTypeTrue NA:age_agg< =18	0.021	-0.007	0.011
	(0.213)	(0.161)	(0.211)
	t = 0.097	t = -0.046	t = 0.053
	p = 0.923	p = 0.964	p = 0.958
ItemTypeFalse Concord:domain_quality_score	-0.100	-0.105*	-0.100
	(0.110)	(0.051)	(0.109)
	t = -0.914	t = -2.047	t = -0.916
	p = 0.361	p = 0.041	p = 0.360
ItemTypeFalse NA:domain_quality_score	-0.142	-0.111*	-0.141
	(0.074)	(0.053)	(0.074)
	t = -1.928	t = -2.113	t = -1.900
	p = 0.054	p = 0.035	p = 0.058
ItemTypeTrue Discord:domain_quality_score	0.072	-0.077	0.066
	(0.098)	(0.044)	(0.097)
	t = 0.729	t = -1.764	t = 0.676
	p = 0.466	p = 0.078	p = 0.500
ItemTypeTrue Concord:domain_quality_score	-0.008	-0.106*	-0.019
	(0.147)	(0.051)	(0.148)
	t = -0.055	t = -2.057	t = -0.130
	p = 0.956	p = 0.040	p = 0.897
ItemTypeTrue NA:domain_quality_score	-0.142	-0.111*	-0.141
	(0.074)	(0.053)	(0.074)
	t = -1.928	t = -2.113	t = -1.900
	p = 0.054	p = 0.035	p = 0.058
ItemTypeFalse Concord:elite_politifact_score	0.004	-0.088	-0.001
	(0.105)	(0.081)	(0.103)
	t = 0.034	t = -1.084	t = -0.014

	p = 0.973	p = 0.279	p = 0.989
ItemTypeFalse NA:elite_politifact_score	-0.083	-0.100	-0.090
	(0.093)	(0.086)	(0.089)
	t = -0.897	t = -1.167	t = -1.016
	p = 0.370	p = 0.244	p = 0.310
ItemTypeTrue Discord:elite_politifact_score	-0.126	-0.073	-0.064
	(0.092)	(0.056)	(0.088)
	t = -1.364	t = -1.288	t = -0.729
	p = 0.173	p = 0.198	p = 0.466
ItemTypeTrue Concord:elite_politifact_score	-0.129	-0.103	-0.073
	(0.157)	(0.084)	(0.159)
	t = -0.822	t = -1.229	t = -0.460
	p = 0.412	p = 0.219	p = 0.646
ItemTypeTrue NA:elite_politifact_score	-0.083	-0.100	-0.090
	(0.093)	(0.086)	(0.089)
	t = -0.897	t = -1.167	t = -1.016
	p = 0.370	p = 0.244	p = 0.310
ItemTypeFalse Concord:toxic_score	0.007	-0.012	0.007
	(0.111)	(0.051)	(0.112)
	t = 0.062	t = -0.234	t = 0.066
	p = 0.951	p = 0.815	p = 0.948
ItemTypeFalse NA:toxic_score	-0.020	-0.016	-0.020
	(0.070)	(0.050)	(0.070)
	t = -0.292	t = -0.317	t = -0.283
	p = 0.771	p = 0.751	p = 0.778
ItemTypeTrue Discord:toxic_score	0.076	0.005	0.063
	(0.089)	(0.042)	(0.090)
	t = 0.849	t = 0.121	t = 0.706
	p = 0.397	p = 0.904	p = 0.481
ItemTypeTrue Concord:toxic_score	0.069	-0.008	0.064
	(0.173)	(0.046)	(0.176)
	t = 0.398	t = -0.165	t = 0.362

	p = 0.691	p = 0.869	p = 0.718
ItemTypeTrue NA:toxic_score	-0.020	-0.016	-0.020
	(0.070)	(0.050)	(0.070)
	t = -0.292	t = -0.317	t = -0.283
	p = 0.771	p = 0.751	p = 0.778
Constant	-0.241*	0.850***	-0.200
	(0.118)	(0.103)	(0.145)
	t = -2.047	t = 8.220	t = -1.387
	p = 0.041	p = 0.000	p = 0.166

Note: *p **p ***p<0.001

Regression predicting item-type note count by item type, political extremity, partisanship, and controls (each interacted with item type); with clustered SEs by participant

Table S29. *Flag count at tweet type level predicted by tweet type, political extremity, and standard controls. Tweet categories with 'NA' indicate instances where user partisanship could not be determined (e.g., users did not follow political accounts). Political extremity is associated with greater flagging of false discordant tweets relative to false concordant and true concordant tweets. Additively (OLS model), political extremity is associated with a greater number of flags on false discordant tweets than true discordant tweets. Multiplicatively (QP model), political extremity is associated with a similar increase in the number of flags on false discordant tweets and true discordant tweets.*

d. Supplemental Figures

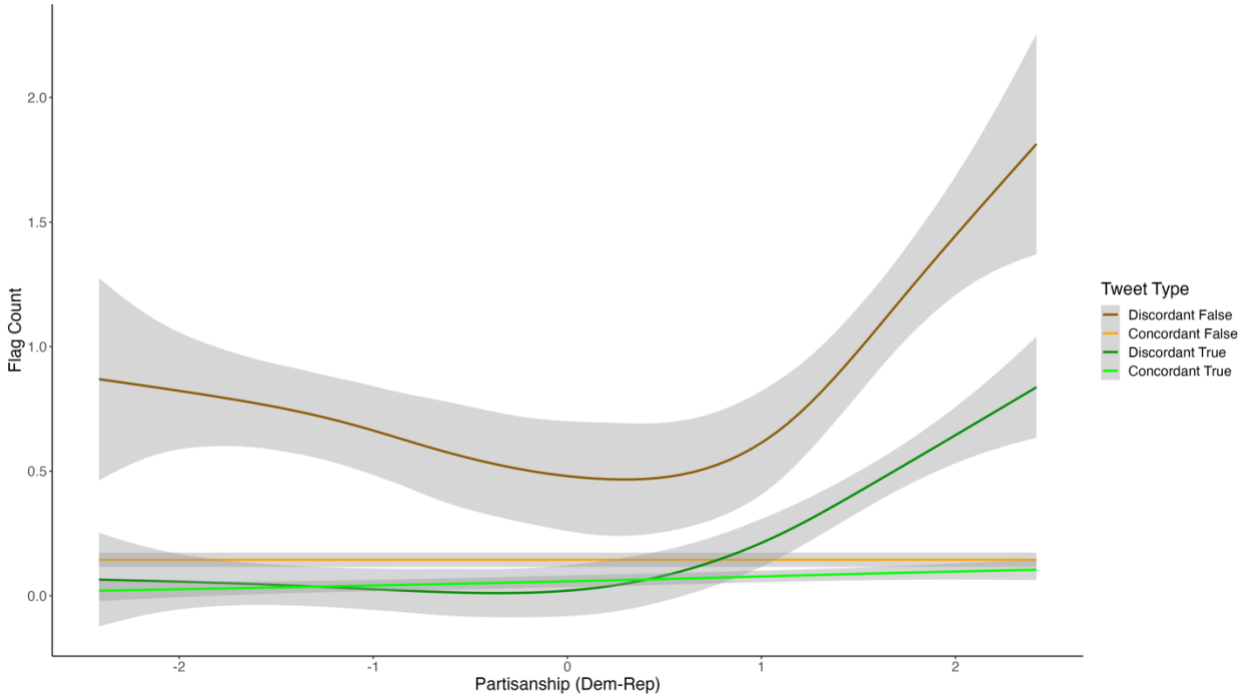

Figure S3. Republican-leaning Community Notes users flag a greater proportion of discordant notes and exhibit lower agreement with fact-checkers. Shown are the predicted number of flags by post type and political motivation. Republican users flag a greater number of both false discordant *and* true discordant tweets, and the number of flags increases with political extremity. Democratic users flag primarily false discordant tweets. Error bars reflect 95% confidence intervals.