Moderating (Mis)information

Authors:
Jacob Meyer[a]
Prithvijit Mukherjee[b]
Lucas Rentschler[c]

January 2023

[a] Cornell University Email: jm2363@cornell.edu
[b] Bryn Mawr College Email: pmukherjee@brynmawr.edu
[c] Utah State University and the Center for Growth and Opportunity Email: lucas.rentschler@usu.edu

* Generous financial support from the Center for Growth and Opportunity at Utah State University is gratefully acknowledged.

The Center for
Growth and Opportunity
at Utah State University

Abstract

This paper uses a laboratory experiment to investigate the efficacy of different content moderation policies designed to combat misinformation on social media. These policies vary the way posts are monitored and the consequence imposed when misinformation is detected. We consider three monitoring protocols: 1) individuals can fact-check information shared by other group members for a cost; 2) the social media platform randomly fact-checks each post with a fixed probability; 3) and a combination of individual and platform fact-checking is used. We consider two consequences: 1) when a post is identified as misinformation, it is flagged; and 2) when a subject is found to have posted misinformation, they are automatically fact-checked for two subsequent rounds. We compare our data to that of Pascarella et al. (2022), who study an identical environment without content moderation. We find that allowing individuals to fact-check improves group decision-making and welfare under either consequence. Platform checking alone does not improve group decisions relative to the baseline with no moderation; it can improve welfare but only in the case with persistent scrutiny. Further, there are marginal improvements when the two protocols are combined. We also find that flagging is sufficient to curb the negative effects of misinformation. Adding persistent scrutiny does not improve the quality of decision-making and instead leads to less engagement on the social media platform as fewer group members share posts.

Keywords: Partisanship, Voting, Communication, Social Media, Misinformation

JEL Classification Numbers: C92, D72, D83

1 Introduction

Social media allows individuals to connect directly with people and organizations from around the globe to share ideas and information. This decentralized approach can enhance social welfare by aggregating knowledge from many sources. At the same time, conflicting interests and objectives may give rise to false or misleading information. The prevalence and negative effects of such misinformation have received significant attention in the popular and academic press. According to a recent survey from the Pew Research Center, most Americans view social media as having "a mostly negative effect on the way things are going in the US today," especially because of misinformation.[1]

Given these concerns, policies aimed at effectively combating misinformation are a topic of popular debate. Can content moderation policies curb misinformation while retaining the broader benefits of social media? What protocols are best suited to detect misinformation, and what consequences should purveyors of misinformation face? To determine whether or not content moderation is effective, we need clean and objective measures of the extent of misinformation as well as how it affects social outcomes. Such measures are not typically available since even reaching a consensus on whether content contains misinformation is contentious. Further, measuring how misinformation affects real-world events is empirically challenging. This is particularly true when assessing the quality of electoral outcomes in the face of the increasing partisan divide.

To address these challenges, we turn to a laboratory experiment. Crucially, our controlled experiment allows us to directly measure outcomes that are unobservable in the real world. We can accurately measure misinformation, objectively determine the quality of voting outcomes, and causally assess the performance of a variety of content moderation policies.

In each period of our experiment, there is an unknown and binary state of the world. Individuals vote for one of the possible states, and each subject receives a fixed payoff if the voting outcome matches the state. In addition, each subject receives a partisan payoff that depends on their type and the voting outcome but is independent of the state. Before voting, each individual decides how much information to purchase about the state.[2] In addition, subjects choose which group members to follow on a simulated social media platform, where they can share a (potentially false) post about the information they purchased and see the posts of those they follow.

We consider six content moderation policies on the social media platform. These policies vary the way posts are monitored (a monitoring protocol) and the consequence imposed when misinformation is detected (a consequence). In all content moderation policies, fact-checks perfectly detect misinformation, and the result of each fact-check is automatically added to the corresponding post.

[1] https://www.pewresearch.org/fact-tank/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/

[2] There is a large literature that studies endogenous information acquisition in voting games. See, for example, Elbittar et al. (2016), Grosser and Seebauer (2016), Großer (2018), and Meyer and Rentschler (2022).

We also consider three monitoring protocols. The first is peer-to-peer fact-checking (P2P), in which each subject can pay a small cost to fact-check a given post. The second method is platform fact-checking (PL), where the social media platform fact-checks each post with 20 percent probability. We also consider the combination of P2P and PL fact-checking (P2P+PL). The monitoring protocol is varied on a within-subject basis. In stage one of the experiment, subjects face either P2P, PL, or P2P+PL; in the second, they face another monitoring protocol.[3]

In addition, we consider two consequences that are imposed on purveyors of misinformation and vary these on a between-subject basis. In the first, fact-checked posts are flagged and report the results of a fact-check, whether it is true or false (FL). In the second, in addition to flagging, a subject who posted misinformation is subject to persistent scrutiny (PS) for the next two periods, wherein the computer automatically checks their posts (FL+PS).

To evaluate the performance of these content moderation policies, we use data from this experiment as well as data reported in Pascarella et al. (2022).[4] The data from Pascarella et al. (2022) allow for causal comparison to baselines without any social media (NONE), with social media but without the potential for misinformation (TR), and with social media but without any content moderation (MIS). This allows us to assess whether content moderation policies can effectively address misinformation and preserve the benefits of social media.

We first consider how different content moderation policies affect the level of misinformation and then assess the relative performance of different content moderation policies as measured by the quality of voting outcomes. To account for the costs of fact-checking imposed on subjects, as well as the costs of acquiring information, we also consider the welfare of subjects. Finally, we assess the problem from the perspective of a social media platform that is interested in increasing user engagement.

Our results are striking. We find that all content moderation policies studied reduce the level of misinformation. The percentage of posts containing misinformation is lowest when individuals have the opportunity to monitor posts and when the platform monitors posts. Interestingly, adding persistent scrutiny of those caught posting misinformation does not decrease misinformation relative to the case with just flagging. We also find that both group decisions and welfare are improved when peer-to-peer monitoring is available relative to when there is no moderation. Platform fact-checking has more mixed results: it does not improve group decisions relative to the baseline with no moderation and improves welfare only in the case with persistent scrutiny.

[3] The order of monitoring protocols methods is balanced across sessions to control for order effects.

[4] Pascarella et al. (2022) study how social media affects voting outcomes in an identical experimental environment but without content moderation. The data for Pascarella et al. (2022) and the current design were collected at ExCEN Laboratory at Georgia State University between February and March of 2020. As such, our data are directly comparable, and we use the data of Pascarella et al. (2022) as benchmarks for our analysis. The data for our experiment and Pascarella et al. (2022) were collected in separate sessions; no subject participated in a session for both studies. The recruitment methods, software, and instructions were the same across the studies except for the inclusion of content moderation. The experimenter running the session was also the same.

2 Related Literature

This study is most closely related to Pascarella et al. (2022), who consider three environments that serve as baselines for our paper. In the first, there is no social media platform for subjects to share information on. In the second, there is social media but no misinformation. In the third, there is social media, misinformation is possible, and there is no content moderation. The results reported in Pascarella et al. (2022) show that the presence of a social media platform can improve voting outcomes provided that misinformation is not permitted. In their experiment, when misinformation is permitted, 41 percent of all posts contain misinformation, which decreases group welfare.

Relatively few papers have studied the effects of different content moderation policies. The main focus has been on how individuals detect misinformation and how it influences their beliefs and self-reported decisions. Jun et al. (2017) find that individuals are less likely to fact-check statements when they believe that the fact-check may be completed by other individuals. Nieminen and Rapeli (2019) summarize studies on how individuals change their beliefs based on fact-checks as well as how the profession of fact-checking has evolved. Their review finds that misinformation affects partisan issues differently than nonpartisan ones. Barrera et al. (2020) study an election in France and find that fact-checking statements by political candidates can improve the accuracy of the voter beliefs but do not influence their decision to vote for a candidate. Nyhan et al. (2019) study the 2016 US election and find that journalistic fact-checks can reduce misperceptions but have minimal effects on candidate evaluations or vote choice.

The literature studying voting games in the laboratory finds that the quality of group outcomes improves when there is an option to communicate and information is available at no cost to the participants (Guarnaschelli et al., 2000, Goeree and Yariv, 2011, Le Quement, 2019, Pogorelskiy and Shum, 2019). This result is robust even in when partisans have an incentive to misrepresent information (Goeree and Yariv, 2011). Le Quement (2019) also finds that communication leads to improved outcomes when there are diverse preferences, though some subjects do produce misinformation.

In a related working paper, Pogorelskiy and Shum (2019) find that individuals are more likely to share signals that are in line with their partisan bias and to put more stock in private signals that lean in the direction of their partisan bias. Additionally, social media improves group outcomes only if the signals voters receive are unbiased. When voters receive biased signals, social media worsens group decisions relative to if there were no social media.

3 Experimental Design

In each experimental session, 15 subjects participate in two stages. At the beginning of each stage, the subjects are randomly assigned to groups of five. Each stage consists of a number of rounds, and groups are fixed within a s tage. In both stages, there are five rounds with certainty, with a 90 percent continuation probability for each additional round.[5] One of the treatment variables (monitoring protocols) in our experiment is varied on a within-subject basis, with subjects seeing one treatment in the first stage and another in the second stage.[6]

At the start of each stage, each subject within a group is randomly assigned an ID from the set {A,B,C,D,E} and a type. Each subject's type, $\boldsymbol{p}$, is an independent and identically distributed draw from a commonly known, discrete uniform distribution with support $\{0, 1, \ldots, 99, 100\}$. Each subject's type is private information, whereas IDs are common knowledge. Subject IDs and types are fixed within a stage.

In each round, there are two possible states of the world, brown or purple. Each state is equally likely, and the state of the world is determined randomly each round and is unknown to subjects. Groups choose either brown or purple via majority vote, and abstention is not permitted.

Each subject receives a fixed payoff of 50 experimental francs (EF) if the group votes for the true state of the world.[7] In addition, each receives a partisan payoff: if the group votes for brown, the subject's partisan payoff is $\boldsymbol{p}$ EF, and if the group votes for purple, the subject's partisan payoff is $(100 - \boldsymbol{p})$ EF. Notice that as $\boldsymbol{p}$ increases (decreases), so does a subject's partisan preference for brown (purple). These partisan preferences are independent of the state of the world and vary from moderate to extreme.[8]

Before voting, each subject can purchase up to nine units of information. If a subject purchases any units of information, they observe a binary signal that corresponds to the state of the world with probability q, and q increases in the number of units purchased. The marginal cost of units of information is increasing (see table 1). Signals are independent, conditional on the state of the world.

[5] The number of rounds is prerandomized for all groups. Groups interact in 17 rounds in the first stage and 18 rounds in the second stage. Subjects are only informed of the continuation probability.

[6] Note that having two stages in our experiment effectively doubles the number of groups in our data and allows us to balance for any potential ordering effects.

[7] The subjects' earnings, denominated in EF, are converted back to USD at $145EF = \$1$.

[8] Robbett and Matthews (2018) find that individuals are more likely to give partisan responses in a voting scenario relative to when they are the decision-maker for a policy. They also find that free access to information reduces the partisan gap in outcomes. When information is costly, individuals do not purchase information and vote according to their partisan preferences.

Table 1. Cost of Information

Units	Accuracy	Total Cost
1	55%	1
2	60%	2
3	65%	5
4	70%	8
5	75%	13
6	80%	18
7	85%	25
8	90%	32
9	95%	41

Note that for sufficiently extreme types, there is no incentive to purchase any information since a signal, regardless of how accurate, would not alter their vote.[9] The presence of such extreme partisans is a critical part of our experimental design since we are interested in misinformation. While such types may not purchase information, they do have an incentive to attempt to influence the votes of others.[10]

After making their purchase decisions, subjects can communicate with other group members via a social media platform. The social media platform is a directed network like Twitter or Instagram. Each subject decides which group members they want to "follow," and subjects can revise who they follow in each period, with the default being their choice from the preceding period.[11] A subject will observe posts made by group members they follow. After learning which group members are following each other, each group member decides whether or not to make a post. A post contains up to three components:

1. Number of units of information purchased: 0–9
2. Signal observed: brown, purple, or none
3. Subject type (optional): 0–100

[9] In particular, if a subject's type is weakly more than 75, they would vote for brown even if they knew the state of the world was purple. Analogously, a subject would vote for purple if their type is weakly less than 25, even if they know the state of the world was brown. To see this, suppose a subject knows that the state of the world is brown. They would still prefer to vote for purple if $50 + p < 100 - p$. That is, when $p < 25$.

[10] Such extreme partisans are exactly the types who are likely to post misinformation or strategically withhold information that runs counter to their partisan preference. Further, these types are particularly likely to target posts for fact-checks that run counter to their partisan preferences.

[11] In the first period, subjects follow all members of the group by default and can modify this structure before proceeding with the period.

The number of units of information purchased, the signal observed, and the type may each or all be falsely reported by a subject. The subject's type could also be omitted entirely.

All posts are subject to content moderation. We have six treatments, corresponding to six content moderation policies. Each policy is composed of a monitoring protocol, which varies how posts are monitored, and a consequence that is imposed when misinformation is detected. Subjects face one monitoring protocol in the first stage and another in the second. The consequence imposed on purveyors of misinformation is constant across stages and varied across sessions. The three monitoring protocols we consider are the following:

1. P2P: Each subject has the option to fact-check any post shared by group members they follow. Checking a post costs 5 EF.
2. PL: The computer fact-checks each post with a 20 percent probability.[12]
3. P2P+PL: The computer fact-checks each post with a 20 percent probability. Subjects can also fact-check any post for a cost of 5 EF, before observing which posts have been checked by the platform.

For all three monitoring protocols, each fact-checked post is automatically flagged as accurate only if the subject correctly states both the units of information purchased and the signal's color. If either of these are reported inaccurately, the post is automatically flagged as inaccurate. The type revealed in the post is not fact-checked, and the results of fact-checks do not report which component of the post was inaccurate. The results from all fact-checks are shared publicly with everyone who can see the post on the platform.

The two consequences we consider are the following:

1. FL: The result of each fact-check is flagged as described above, and no further actions are taken.
2. FL+PS: The result of each fact-check is flagged as described above. Additionally, if a subject's post is flagged as inaccurate, the computer will automatically fact-check any post shared by that subject during the next two rounds.

After subjects observe the results from any fact-checks that occurred, they each vote for either brown or purple. Each subject then observes the group's decision, the randomly assigned state of the world, and their own private payoff. Each player's payoff for the round is composed of the partisan and nonpartisan payoffs described above, less any expenditure on units of information and P2P fact-checks.

Note that all of our treatments involve a content moderation policy. To assess the performance of a given content moderation policy, we compare our data against baseline environments from Pascarella et al. (2022). In Pascarella et al. (2022), subjects participated in three variations of the experimental environment

[12] There is no cost to group members if the "platform" conducts a fact-check so that the cost of fact-checks is not equal across monitoring protocols. This is an asset to our design since when real-world social media platforms fact-check user content, no costs are imposed on users of the platform, and we are interested in assessing content moderation policies as they would actually be implemented.

described above but without ex-post content moderation. In particular, subjects in the study participated in treatments that either did not include a social media platform (none), included a platform that only allowed truthful posts (TR), or included a platform that allowed for both truthful and untruthful posts (MIS). These three treatments serve as benchmarks for the content moderation policies considered in this study, as the data from Pascarella et al. (2022) are directly comparable to the data from this study.[13]

We ran a total of 12 sessions, with 6 sessions for each consequence. The 6 sessions that correspond to a consequence cover all the possible combinations and orders of monitoring protocols to control for order effects. The experiment was conducted at the ExCEN laboratory at Georgia State University in early March of 2020. A total of 180 subjects participated in the experiment across 12 sessions of 15 subjects each. The subjects were recruited using an automated system that randomly invited participants via e-mails from a pool of more than 2,400 students who signed up for participation in economic experiments. Each session was computerized using z-Tree (Fischbacher, 2007). Subjects were not allowed to communicate with each other than through the experiment's social media platform.[14]

At the start of each stage, subjects were provided with a physical copy of the instructions, which were then read aloud by the experimenter.[15] After the instructions were read, each subject individually answered a series of quiz questions to ensure comprehension. Each session lasted for approximately 2 hours and 15 minutes. The average payoff was $19.50, with a range between $9.75 and $26.40.

4 Results

4.1 (Mis)information

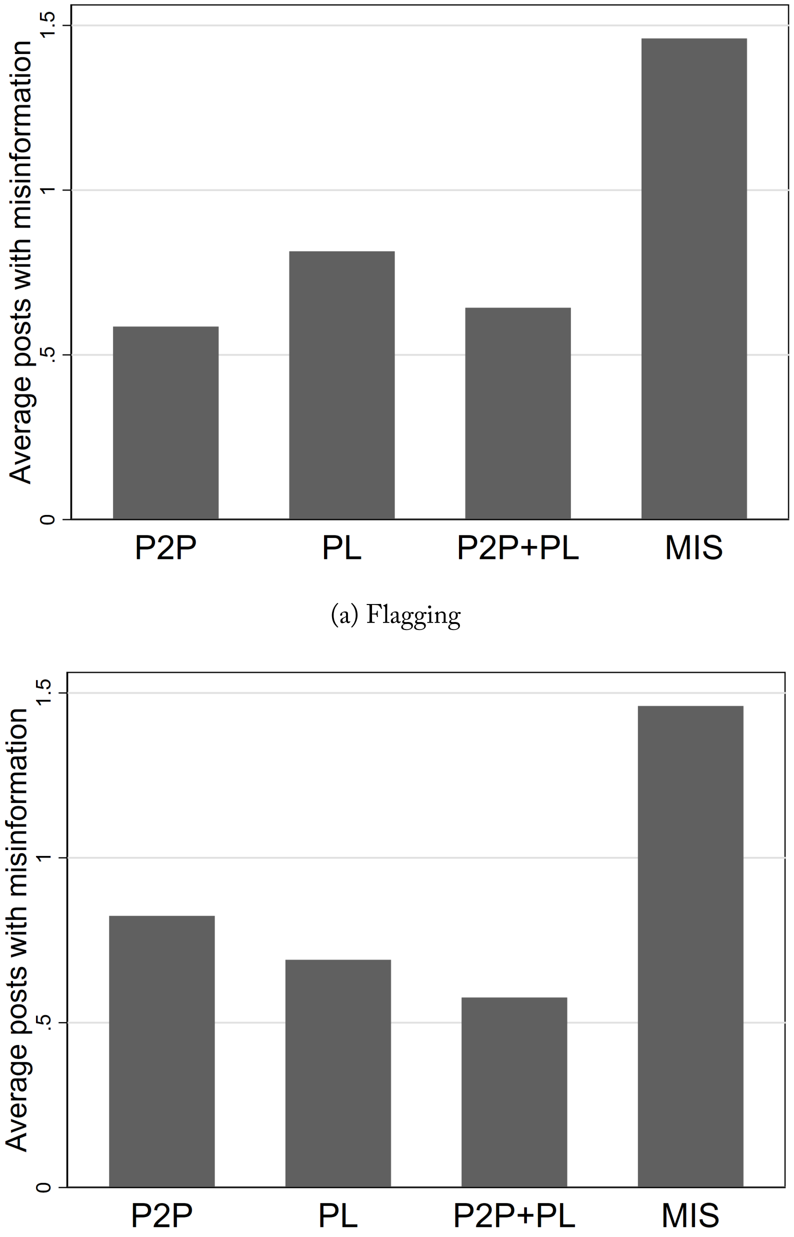
We first report how the different content moderation policies in our experiment affect the amount of misinformation shared by subjects. Figure 1 reports the mean number of posts containing misinformation across treatments. Panel (a) reports the three treatments where the consequence is flagging, while panel (b) reports the three treatments where the consequence is flagging and persistent scrutiny. Both panels contain the baseline treatment in which social media does not have any content moderation policy (MIS).

[13] See footnote 4 for more details.

[14] To ensure privacy, each computer terminal in the lab was enclosed with dividers.

[15] See Appendix A for a sample set of instructions.

Figure 1. Average Number of Posts Containing Misinformation by Treatment



(a) Flagging

(b) Flagging + Persistent Scrutiny

Of note, all content moderation policies successfully reduce the number of posts containing misinformation.[16] While this is important, content moderation may be accompanied by a reduction in the

[16] Each of the corresponding t-tests are highly significant, with $p < 0.001$. In our analysis, we focus on group decisions since the interest of our study is to learn how content moderation policies affect aggregate outcomes. Our unit of observation is a group in a given round, and we report the results of two-tailed t-tests.

number of posts themselves, so it is possible that a content moderation policy reduces the number of posts containing misinformation, while the percentage of posts with misinformation actually increases. To investigate this possibility, table 2 reports summary statistics on the number of posts in a group, the share of posts containing misinformation, and the percentage of posts that were fact-checked by each available monitoring protocol.

Table 2. Composition of Posts Shared

	Posts in a Group	Share of Group Posts with misinformation	Share of Posts Fact-Checked (P2P/PL/P2P+PL)
	(1)	(2)	(3)
Panel A: Flagging			
P2P	3.23	0.18	0.31/0/0
	(1.15)	(0.21)	(0.30/0/0)
PL	3.39	0.27	0/0.20/0
	(1.21)	(0.29)	(0/0.23/0)
P2P+PL	3.25	0.19	0.23/0.16/0.05
	(1.10)	(0.25)	(0.27/0.23/0.13)
N	210	210	210
Panel B: Flagging + Persistent Scrutiny			
P2P	2.94	0.29	0.21/0/0
	(1.26)	(0.32)	0.28/0/0
PL	3.05	0.24	0/0.28/0
	(1.24)	(0.27)	(0/0.28/0)
P2P+PL	2.92	0.19	0.17/0.15/0.02
	(1.24)	(0.23)	(0.25/0.22/0.08)
N	210	210	210
Panel C: Baseline Treatments			
TR	3.24		
	(1.39)		
MIS	3.09	0.41	
	(1.39)	(0.32)	
N	252	252	

Notes: The table contains means of per-period group outcomes, with standard deviations in parentheses.

Column 1 shows that the consequence faced for posting misinformation does affect the number of posts: persistent scrutiny reduces the number of posts relative to the case of flagging alone ($p < 0.001$).[17] However, the monitoring protocol does not affect the number of posts, holding the consequence fixed

[17] For this test, we pooled the data across monitoring protocols. The results are analogous when considering all pairwise comparisons.

(Kruskal–Wallis, *n.s.*, in both cases).[18] We next analyze the prevalence of misinformation as a percentage of total posts since the number of posts differs across consequences.[19] As seen in column 2, all content moderation policies reduce the proportion of posts containing misinformation relative to the full information baseline ($p < 0.001$) for all comparisons.

Comparing the monitoring protocols, the percentage of posts containing misinformation is only marginally less in P2P than in PL when pooling across consequences ($p < 0.1$). However, the combination of these protocols (P2P+PL) outperforms both P2P ($p < 0.05$) and PL ($p < 0.001$). Thus, establishing multiple channels for fact-checking is likely to be most successful in reducing the percentage of posts containing misinformation. Furthermore, adding persistent scrutiny does not affect the percentage of posts containing misinformation when pooling across monitoring protocols.[20] That is, a harsher penalty does not seem to yield any additional deterrence.

4.2 Quality of Group Decision-Making

We next report how content moderation policies affect the quality of group decision-making. We define quality of group decision-making as the frequency with which a group's vote matches the state of the world. Table 3 reports summary statistics of this measure. Notice that the P2P monitoring protocol improves the quality of decisions relative to the absence of content moderation (the MIS baseline). This is true for both consequences ($p < 0.05$ in both cases). In fact, the quality of group decisions in the P2P monitoring protocol is not significantly different than the case in which misinformation is not possible (the TR baseline). That is, P2P is an effective monitoring protocol. However, the same cannot be said for the PL monitoring protocol. In particular, the quality of group decisions is not significantly different than in the MIS baseline and is lower than the TR baseline.[21]

[18] When a test is not statistically significant at conventional levels, we denote this as *n.s*.

[19] In Appendix B, we report summary statistics where the variable of interest is the percentage of the units of information purchased, including those shared inaccurately (column 3). The results presented here extend to these data as well.

[20] Adding persistent scrutiny has no statistically significant effect within the PL or the P2P + PL monitoring protocols. Adding persistent scrutiny actually increases the percentage of posts containing misinformation when the detection method is P2P ($p < 0.001$). It is unclear what drives this increase. Even with this puzzling result, we can say that adding persistent scrutiny of purveyors of misinformation does not reduce the percentage of posts containing misinformation in our experiment.

[21] These results are all true under both consequences. The corresponding t-tests comparing the quality of group decision-making across PL and TR have $p < 0.05$. In addition, the quality of group decision-making under PL is not statistically different from the scenario without social media (the NONE baseline) under either consequence.

Table 3. Quality of Group Decision-Making

	Group Vote Is Correct
	(1)
Panel A: Flagging	
P2P	0.67
	(0.47)
PL	0.61
	(0.49)
P2P+PL	0.71
	(0.46)
N	210
Panel B: Flagging + Persistent Scrutiny	
P2P	0.67
	(0.47)
PL	0.63
	(0.48)
P2P+PL	0.68
	(0.47)
N	210
Panel C: Baseline Treatments	
NONE	0.62
	(0.49)
TR	0.71
	(0.45)
MIS	0.58
	(0.50)
N	252

Notes: The table contains means of per-period group outcomes, with standard deviations in parentheses.

The P2P+PL monitoring protocol follows the same pattern as P2P. The quality of decisions is improved relative to MIS ($p < 0.001$ in both cases) and is not statistically different from TR. That is, adding platform monitoring to the P2P monitoring protocol does not result in any gains. Similar to the results on the prevalence of misinformation, the addition of persistent scrutiny provides no benefit in terms of group decisions ($n.s$ for all comparisons).

Importantly, the results reported in this section are not driven only by the proportion of posts that are fact-checked at the group level. To see this, note that the rates of fact-checking reported in column 3 of table 2 are similar across the different content moderation policies.[22]

4.3 Welfare

P2P monitoring improves the quality of group decision-making. However, this improvement comes at a cost as subjects bear the expense of monitoring. Are the costs associated with improving group decisions offset by the increased earnings from improved group decision-making? To assess this, we now analyze group earnings (welfare).[23]

Table 4 decomposes average group payoffs into the payoffs associated with the voting outcome, information expenditures, and fact-checking expenditures. Column 4 shows that the costs associated with P2P monitoring are not sufficient to erase the gains. That is, in both P2P and P2P+PL, welfare is higher than the MIS baseline ($p < 0.05$ for both consequences) and is not significantly different than the TR baseline (again, for both consequences).

Turning to PL monitoring, we find that when there is no persistent scrutiny, the costless fact-checks provided by the platform are insufficient to improve welfare relative to the MIS baseline.[24] However, adding persistent scrutiny tips the scales; in this case PL improves welfare relative to the MIS baseline ($p < 0.05$). This is not surprising since persistent scrutiny results in additional no-cost fact-checking by the computer. However, it is important to note when comparing consequences, the addition of persistent scrutiny does not improve welfare relative to the case of only flagging. This is true for all monitoring protocols ($n.s.$ in all cases).[25]

[22] To explore other factors that may influence group decision quality, we analyze panel linear probability models in Appendix C.
[23] We define welfare as the sum of the net payoff of all members in the group.
[24] Not surprisingly, when misinformation is only flagged, PL monitoring results in lower welfare than in the TR baseline ($p < 0.05$).
[25] Regression analysis yields similar results. See Appendix C for details.

Table 4. Group Payoffs

	Voting Payoff (1)	Information Expenditures (2)	Fact-Checking Expenditures (3)	Net Payoff (4)
Panel A: Flagging				
P2P	421.30	15.86	5.95	405.45
	(124.91)	(12.66)	(6.41)	(123.20)
PL	416.23	21.26		394.97
	(128.80)	(14.54)		(126.84)
P2P+PL	426.40	17.76	6.02	408.63
	(122.88)	(13.56)	(7.47)	(121.65)
N	210	210	210	210
Panel B: Flagging + Persistent Scrutiny				
P2P	424.60	19.32	3.21	405.27
	(124.52)	(14.01)	(4.56)	(122.21)
PL	422.15	18.39		403.76
	(126.75)	(14.71)		(126.13)
P2P+PL	424.79	20.01	3.17	404.77
	(124.60)	(14.59)	(4.51)	(122.73)
N	210	210	210	210
Panel C: Baseline Treatments				
NONE	414.15	16.19		397.96
	(130.64)	(13.56)		(129.33)
TR	438.56	20.65		417.90
	(121.07)	(18.82)		(118.44)
MIS	406.99	23.72		383.27
	(132.15)	(16.71)		(130.32)
N	252	252		252

Notes: The table contains means of per-period group outcomes, with standard deviations in parentheses.

4.4 Social Connectedness

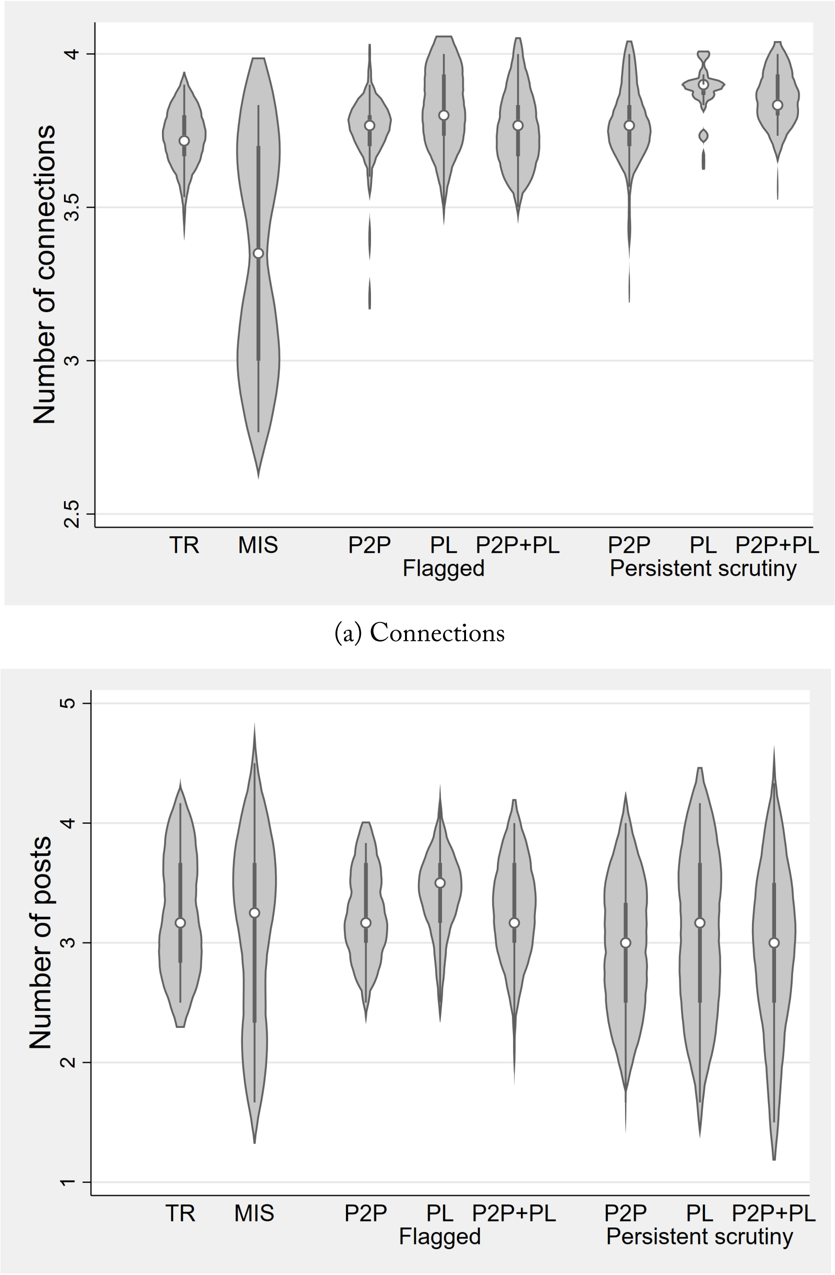
Pascarella et al. (2022) show that in the absence of content moderation policies, misinformation degrades group decision-making, in part due to social connections decaying over time. Despite this, they find that activity on the social network, as measured by the number of posts, is not reduced by misinformation. Thus, a key question for this study is whether content moderation policies can improve the number of connections relative to the MIS baseline.

Panel (a) of figure 2 shows violin plots of the distributions of the number of connections in groups in the TR and MIS baselines as well as in the six content moderation policies we study. All six content moderation policies result in more connections than in the MIS baseline ($p < 0.001$ in all cases); that is, the introduction of content moderation increases engagement on the social media platform. Indeed, some content moderation policies increase the number of connections relative to the TR baseline.[26] Panel (b) shows the violin plots for the number of posts. Note that there are no economically significant differences in the number of posts made within a group.[27]

[26] For the PL monitoring protocols, this is significant for both consequences ($p < 0.001$, in both cases). For both P2P and P2P+PL with persistent scrutiny, this difference is significant ($p < 0.001$, in both cases). However, for both P2P and P2P+PL with only flagging, there is no significant difference.

[27] While some of the pairwise comparisons are statistically significant, the magnitudes of the differences are economically trivial.

Figure 2. Violin Plots of the Number of Connections and Posts



(a) Connections

(b) Posts

5 Discussion

Pascarella et al. (2022) find that introducing a social media platform can improve group outcomes, provided that misinformation is not permitted. When misinformation is permissible, all the gains from introducing a platform are erased, and group outcomes are worse than the scenario in which there is no social media platform at all.

We study whether content moderation policies can preserve the benefits of social media even when they may not be able to perfectly screen out all misinformation. In our experiments, a content moderation policy consists of a method of monitoring social media posts as well as a consequence that is imposed on identified purveyors of misinformation. We consider monitoring protocols in which social media users can pay a cost to initiate a fact-check, the platform fact-checks each post with a known probability, and a combination of these two. We consider two consequences for posting misinformation: flagging a post's accuracy whenever a fact-check is performed and (in addition) persistently scrutinizing identified purveyors of misinformation.

All the content moderation policies we study reduce the level of misinformation on the social media platform relative to the case with no moderation. Further, they each reduce the share of posts that contain misinformation. The combined monitoring by peers and the platform performs the best, reducing the prevalence of misinformation by 22 percentage points (a 54 percent reduction relative to baseline). Notably, the addition of persistent scrutiny as a consequence of posting misinformation has no effect on the share of posts containing misinformation. This suggests, at a minimum, some diminishing returns to the punishments imposed on purveyors of misinformation. However, all policies considered do reach the primary goal of reducing the presence of misinformation on the platform.

Importantly, our design also allows us to speak to how content moderation policies affect group decisions and welfare, which would be unobservable outside of an experimental setting. We find that both group decisions and welfare are improved when peer-to-peer monitoring is available relative to when there is no moderation.[28] That is, peer-to-peer monitoring, though costly for subjects to use, improves both decisions and welfare. Platform fact-checking has more mixed results. When used alone, it does not improve group decisions relative to the baseline with no moderation. It can improve welfare but only in the case with persistent scrutiny, where the computer is fact-checking close to 30 percent of posts.

Our research provides the first rigorous evaluation of content moderation policies and provides valuable insights in the fight against misinformation. Our results also highlight the value of laboratory studies: policy proposals can be tested in a controlled setting before being rolled out in the real world. This experiment can serve as a foundation for future studies on (mis)information-sharing and voting outcomes, and we believe there is much to be done in this promising area.

[28] In fact, when peer monitoring is allowed, group decisions and welfare are not statistically different from the baseline with costless, perfect content moderation (the TR treatment).

One important avenue for future research is to investigate whether our results are robust to changing the parameters implemented in our experiment. That is, it is important to conduct additional experiments that test the edge of the validity of our results, in the spirit of Smith (1982). In particular, a promising avenue would be to study content moderation policies with a higher cost of initiating a peer-to-peer fact-check as well as policies with more efficient platform fact-checks. Other mechanisms to curb misinformation could also be explored.

References

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, 182:104123.

Elbittar, A., Gomberg, A., Martinelli, C., and Palfrey, T. R. (2016). Ignorance and bias in collective decisions. *Journal of Economic Behavior & Organization*.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Goeree, J. K. and Yariv, L. (2011). An experimental study of collective deliberation. *Econometrica*, 79(3):893–921.

Großer, J. (2018). Voting game experiments with incomplete information: A survey. *Available at SSRN 3279218*.

Grosser, J. and Seebauer, M. (2016). The curse of uninformed voting: An experimental study. *Games and Economic Behavior*, 97:205–226.

Guarnaschelli, S., McKelvey, R. D., and Palfrey, T. R. (2000). An experimental study of jury decision rules. *American Political Science Review*, 94(2):407–423.

Jun, Y., Meng, R., and Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, 114(23):5976–5981.

Le Quement, Mark Tand Marcin, I. (2019). Communication and voting in heterogeneous committees: An experimental study. *Journal of Economic Behavior & Organization*.

Meyer, J. and Rentschler, L. (2022). Abstention and informedness in nonpartisan elections. *Working paper*.

Nieminen, S. and Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3):296–309.

Nyhan, B., Porter, E., Reifler, J., and Wood, T. J. (2019). Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, pages 1–22.

Pascarella, J., Mukherjee, P., Rentschler, L., and Simmons, R. (2022). Social media, (mis)information, and voting decisions. *Working paper*.

Pogorelskiy, K. and Shum, M. (2019). News we like to share: How news sharing on social networks influences voting outcomes. *Available at SSRN 2972231*.

Robbett, A. and Matthews, P. H. (2018). Partisan bias and expressive voting. *Journal of Public Economics*, 157:107–120.

Smith, V. (1982). Microeconomic systems as an experimental science. *American Economic Review*, 72(5):923–955.

A Sample Instructions

Welcome

No Talking Allowed

Once the experiment begins, we request that you not talk until the end of the experiment. If you have any questions, please raise your hand, and an experimenter will come to you.

Payment

For today's experiment, you will receive a show-up fee of \$5. All other amounts will be denominated in Experimental Francs (*EF*). These Experimental Francs will be traded in for Dollars at a rate of $145EF = \$1$.

Stages

This experiment will be conducted in two stages. At the beginning of each stage:

1. Five individuals from the room would be randomly matched to form a group.
2. Each group member will be randomly assigned a Subject ID – A, B, C, D, or E.
3. Each group member will be assigned a whole number, p, which is randomly chosen between 0 *EF* and 100 *EF*. Any number between 0 *EF* and 100 *EF* has the same chance of being selected. It is independently drawn for each group member. Therefore, the draw of p for one group member is not affected by those of the other members of the group. Each group member knows their own p, but not those of the other group members.

In each stage, everyone in the group will make decisions for at least 5 rounds. There is a 90% chance there will be a sixth round. If there is a sixth round, there is a 90% chance there will be a seventh round, and so on. Thus, at the end of each round (after the fifth round) there is a 90% chance that there will be at least one more round.

You can think of this as the computer rolling a 10 sided dice at the end of each round after the fifth round. If the number is 1 through 9 there is at least one more round. If the number is 10 then the stage end.

Round overview

Each round will consist of the following sequence:

1. **Jar Assignment:** Each group is randomly assigned either a Brown Jar or a Purple Jar. The color of the jar will not be known by any group member.
2. **Buying Information:** Each group member has the option to buy information about the color of the Jar that has been randomly assigned.
3. **Connections:** Each group has a message board. Each group member has the option to follow other members of the group on this message board.
4. **Post Choice:** After making connection decisions, each group member will have a choice regarding whether or not to post information.
5. **Post Creation:** Each group member who has opted to post information determines the content of their post for the message board.
6. **Viewing and fact-checking posts on the message board:** Each group member will observe the posts of those group members they are following, provided they have decided to make a post.

   Each group member has the option to Fact-Check the accuracy of any Post they observe on the message board at a cost (*Peer to Peer protocol)*. In addition, there is a 20% chance that the computer automatically Fact-Check each group member's Post (*Platform and Peer to Peer and Platform protocols)*.
7. **Results from Fact-Checking:** If a group member's Post is Fact-Checked the results of the Fact-Check will be observed by all group members who observe the Post.
8. **Vote:** Each group member casts a vote for either Brown or Purple. The color that gets three or more votes is the group decision.

Decision Environment & Choices

All Treatments

Jar Assignment

At the beginning of every round, the computer will randomly assign one of two options as the correct Jar for each group: the Brown Jar or the Purple Jar. In each round, there is a 50% probability that the Jar assigned to a group is the Brown Jar and a 50% probability that the Jar assigned to a group is the Purple Jar. The computer will choose the Jar randomly for each group and separately for each round. Therefore, the chance that your group is assigned the Brown Jar or the Purple Jar shall not be affected by what happened in previous rounds or by what is assigned to other groups. The choice shall always be completely random in each round, with a probability of 50% for the Brown Jar and 50% for the Purple Jar.

Voting Task

Each group member will decide between one of the two colors: Brown or the Purple. Specifically, at the end of each round, the group will simultaneously vote for either Brown or Purple. The group decision will be determined by the color which gets three or more votes.

Payoff from voting

The payoff from voting that each group member earns in the round depends on the outcome of the vote. There are two parts to this payoff.

Part A

Remember that at the start of each stage, each group member is assigned a whole number, p, which is randomly chosen between 0*EF* and 100*EF*. Any number between 0*EF* and 100*EF* has the same chance of being selected. It is independently drawn for each group member. Therefore, the draw of p for one group member is not affected by the draw of p of the other members of their group. Each group member knows their own p, but not those of their other group members. Remember that the value of p assigned to each group member remains fixed within each stage but is randomly assigned in each stage.

Each group member gets **p *EF*** if the group votes for Brown, and **$100 - p$ *EF*** if the group votes for Purple. This payoff does not depend on the color of the Jar,which is randomly assigned at the start of each round. Notice that this means that each group member is likely to get a different payoff if Brown wins the vote because each group member is likely to have a different p. Similarly, each group member is likely to get a different payoff if Purple wins the vote.

Part B

If the color chosen by the vote matches the color of the Jar that was randomly assigned at the start of the round, each group member gets a payoff a payoff of ***50 EF***.

Example 1: Suppose you are assigned ***p* = *70***, Purple is the color chosen by the vote, and the color of the Jar that is randomly assigned to your group is Purple.
Since you're assigned ***p* = *70***, your payoff from **<u>Part A</u>** is: **100 – 70 = 30EF.**
Since the color of the Jar assigned to your group is the same as the color chosen by your group in the vote, your payoff from **<u>Part B</u>** is: **50EF**.
Thus, your total payoff from the voting task is: **80EF**

Example 2: Suppose you are assigned p = 70, Brown is the color chosen by the vote, and the color of the Jar that is randomly assigned to your group is Purple.
You earn 70 EF from the voting task.
Since you're assigned p = 70, your payoff from **<u>Part A</u>** is: **70 EF**.
Since the color of the Jar assigned to your group is NOT the same as the color chosen by your group in the vote, your payoff from **<u>Part B</u>** is: **0EF**.
Thus, your total payoff from the voting task is: **70EF**.

Buying Information

Remember that none of the group members know the color of the Jar that has been randomly assigned to a group prior to voting on a color. Each group member has an option to buy multiple units of information.

If a group member purchases information, he/she will observe a Report, which is either Brown or Purple. The probability that the color of this Report is the same as the color of the Jar depends on how many units of information the group member purchases.

Each group member can buy any number of units between 0 and 9 (in increments of one unit). If a group member buys a single unit of information, their Report is correct 55% of the time. If a group member purchases two units of information, their Report is correct 60% of the time. Each additional unit of information that a group member purchases increases the probability that their report is correct by 5

You can think of this process as the computer starts with a box with 50 Brown and 50 Purple balls. Each unit of information a group member purchases the computer adds 5 balls with the color of the Jar randomly assigned and removes 5 balls of the other color. The computer then mixes the balls and selects one randomly. The color of this selected ball is the color in the Report. So, for example, if a group member buys four units of information, the box from which the computer randomly selects a ball contains 70 balls with the color of the Jar which is randomly assigned and 30 balls of the other color. The cost of units of information is detailed in the table below.

Units	Probability Correct	Total Cost
1	55%	1
2	60%	2
3	65%	5
4	70%	8
5	75%	13
6	80%	18
7	85%	25
8	90%	32
9	95%	41

If a group member chooses not to buy any units of information, they do not get a Report.

Example:

Suppose you choose to buy 5 units of information and your Report is Brown.The cost of the 5 units of information is 13 *EF*,and there is a 75% chance the information is correct.

Connections

Only in Communication Treatments

At the start of each stage all group members are following each other on the message board. After each group member has decided how many units of information they wish to purchase, and viewed their Reports (when applicable), each group member decides who they would like to follow on the group's message board.

Each group member can only see Posts made by group members they are following. Each group member is identified by the Subject ID assigned to them at the beginning of each stage. Remember that the Subject ID assigned to each group member remains the same within a stage.

Note that if you follow a particular group member, but they do not follow you, then they do not see your posts.

Post Choice

After connection decisions have been made, each group member is shown:

1. Group members they are following.
2. Group members who are following them.

Each group member then chooses whether or not to make a Post.

Post Creation

Each group member who has opted to make a post determines the following contents of their post:

1. p randomly assigned to the group member.
   The group member can input a number between 0 *EF* and 100 *EF*. Their Post will state that the inputted number is the value of p assigned to them. Note that the number imputed does not have to be equal to their p. He/She can also opt not to input a number.

2. Units of information the group member purchased.
   Each group member states the number of units of information they purchased. He/She can input any whole number between 0 and 9. Note that the number they state does not have to be equal to the actual number of units of information they purchased.

3. The group member states the color of their Report, if they state they have purchased one or more units of information. Notice that the color they state does not have to equal the actual color of their Report.

Viewing posts on the message board

Each group member will observe the Posts of those they are following, provided they decided to make a Post.

Each group member then has the option to Fact-Check the accuracy of any of these Posts. It costs 5EF to fact-check a Post. (*Peer to Peer protocol*)

In addition, there is a 20% chance that the computer automatically Fact-Check each group member's Post. (*Platform protocol*)

If a Post is Fact-Checked, the computer will check the accuracy of the following information stated in the Post:

1. The number of units of information purchased.
2. The color of the Report, if the Post states that any units of information were purchased.

Note: A Fact-Check does not verify ⊠.

Results from Fact-Checking

The results from any Fact-Check will be displayed with the Post, and will be observed by anyone who is able to observe the Post. These results will be displayed before the group votes.

If a group member's Post is Fact-Checked, the results of the Fact-Check will be displayed with the Post. The result will state whether the Post is Accurate or Inaccurate. (*Flagging*)

In addition, if a Post is Inaccurate the next two subsequent Posts shared by this group member will be automatically Fact-Checked by the computer. (*Persistent Scrutiny*)

Example 1: Suppose your assigned Subject ID is C and you choose to follow Subjects B, D, and E, and not follow Subject A. Further suppose only Subject A and Subject B chose to create Posts. Since you are following Subject B, you will see their Post. Since you are not following Subject A, you will not see their Post. Suppose you decide to Fact-Check the accuracy of the information in the Post created by Subject B. You and everyone following Subject B will be able to see the results from the Fact-Check before the group decides to vote.

Example 2: Suppose your assigned Subject ID is C, and Subjects A and E choose to follow you, while Subjects B and D choose not to follow you. Further suppose that you decide to make a Post. Subjects A and E will see your Post, while Subjects B and D will not see your Post. Suppose neither A nor E choose to Fact-Check your post, but the computer randomly Fact-Check the accuracy of your post. The results from the Fact-Check will be available to both Subjects A and E before the group decides to vote.

Vote

After all group members observe the Fact-Checking results (if any) on the group message board, each group member casts a vote for either: Brown or Purple.

Each group cast their vote without knowing the votes of the other members of their group.

The computer sums up the number of votes for Brown and for Purple. The color which receives three or more votes is the group's decision.

Final Payoff

Each group member's final payoff for the round is given by: **Peer to Peer protocol and Peer to Peer and Platform Protocol**

Final Payoff = Payoff from Voting – Total cost of buying information - Cost of Fact-Checking

Platform protocol

Final Payoff = Payoff from Voting – Total cost of buying information

Remember that the payoff of each group member in the round depends on the outcome of the vote, and the number of units of information they purchased.
Remember that there are two parts of the payoff from Voting:

Part A: Each group member gets ***p EF*** if the group votes for Brown, and ***100 – p EF*** if the group votes for Purple. This payoff does not dependent on the color of the Jar, which is randomly assigned at the start of each round.

Part B: If the group's decision matches the color of the Jar that was randomly assigned at the start of the round, each group member get a payoff a payoff of ***50 EF***.
The cost of Fact-Checking each post is ***5 EF*** (*Only in Peer to Peer*). The cost of information depends on the number of units of information purchased. The table below contains these costs.

Units	Probability Correct	Total Cost
1	55%	1
2	60%	2
3	65%	5
4	70%	8
5	75%	13
6	80%	18
7	85%	25
8	90%	32
9	95%	41

Example: Suppose you are assigned ***p = 70***, the group's decision is Purple, and the color of the randomly assigned jar is Purple.Further suppose that you purchased 5 units of information and got a Brown Report. Suppose in addition, you chose to Fact-Check one Post.

Your final payoff for the round is 62 *EF*. You get 30*EF* from the group's decision being Purple (Part A) plus you 50 *EF* since the group's decision matched the color of the randomly selected jar (Part B) *minus* 13 *EF* for buying 5 units of information *minus* 5 *EF* for Fact-checking one Post.

B Composition of Information Shared

Table 5. Group Information Purchasing and Sharing

	Units of Information Purchased (1)	Shared Truthfully (2)	Shared Misinformation (3)	Strategically Withheld (4)	Nonstrategically Withheld (5)
Panel A: Flagging					
P2P	8.89	0.72	0.09	0.13	0.06
	(4.29)	(0.28)	(0.16)	(0.21)	(0.14)
PL	9.86	0.71	0.08	0.14	0.08
	(4.04)	(0.31)	(0.14)	(0.23)	(0.19)
P2P&PL	8.84	0.80	0.06	0.08	0.05
	(3.94)	(0.25)	(0.13)	(0.18)	(0.11)
N	210	210	210	210	210
Panel B: Flagging + Persistent Scrutiny					
P2P	9.45	0.61	0.12	0.16	0.12
	(4.77)	(0.33)	(0.23)	(0.23)	(0.21)
PL	8.64	0.74	0.12	0.07	0.08
	(5.12)	(0.31)	(0.21)	(0.16)	(0.22)
P2P&PL	9.15	0.75	0.06	0.12	0.07
	(4.79)	(0.28)	(0.13)	(0.19)	(0.17)
N	210	210	210	210	210
Panel C: Baseline Treatments					
TR	9.50	0.87	0.00	0.10	0.04
	(5.74)	(0.24)	(0.00)	(0.22)	(0.12)
MIS	10.03	0.49	0.30	0.11	0.11
	(4.63)	(0.33)	(0.33)	(0.20)	(0.21)
N	252	252	252	252	252

Notes: Strategically withheld indicates information withheld if the color of the report did not match their partisan bias. Nonstrategically withheld indicates information not shared even when the color of the report matched their partisan bias.

C Regression

C.1 Quality of Group Decisions

We analyze factors that influence group decision quality in a panel linear probability regression analysis with the standard error clustered at the group level.[29] Table 6 presents the results for the five models we estimate. The dependent variable is whether the group's vote matched the correct policy. In columns 1 and

[29] Logistic and probit regression models yield comparable results.

2, we include data from the fact-checking and the information-sharing protocols. The independent variables include a dummy for all the protocols; the excluded category is the peer-to-peer fact-checking when posts are flagged. We also control for the order in which the protocol was introduced and learning across periods.[30] In column 2, we include a variable to account for the total information purchased by groups.

In columns 3 and 4, we include data from all protocols where a social media platform was present to separate the platform's effect on group decision quality. In column 3, in addition to the total information purchased, we control for the average number of connections on the platform. In column 4, we include two variables that account for the nature of information shared on the platform. First, units of information that are wasted due to the corresponding post containing misinformation. Second, we include a variable that accounts for the units of information strategically withheld. In column 5, we only include data from the fact-checking protocol and include a variable to account for the total number of fact-checked posts.

The overall quality of group decisions is not significantly different across the three fact-checking protocols for both the consequences but is significantly lower when there is no option to fact-check and when participants can share posts containing misinformation. Once we account for the total information purchased, we find that when only the platform is fact-checking and flagging posts, it leads to a lower quality of decision-making among the fact-checking protocols. Recall that, apart from platform fact-checking under persistent scrutiny, this leads to an increase in information purchased in the presence of a social media platform. The increase in the information purchased does not translate to an increase in the quality of the group decision when only the platform is fact-checking and when there is no option to fact-check in the MIS baseline.

In columns 3 and 4, we consider only the protocols where a social media platform is present. Recall that fact-checking protocol leads to an increase in the average number of connections relative to TR and MIS baselines. The increase in the average number of connections leads to an improvement in the quality of group decisions. Not surprisingly, we find that an increase in the information wasted due to misinformation lowers group decision quality. In column 5, we only consider the three fact-checking protocols and find that platform fact-checking when posts are flagged lowers the quality of decisions. Although the platform fact-checking, when there is persistent scrutiny of posts, leads to a lower quality of group decisions, these differences are not significant once we account for the lower information purchase.

[30] The variable used to pick up order effects is a dummy variable for whether or not the protocol was the first protocol a participant participated in. Learning across periods is controlled for via $log(t-1)$.

Table 6. Quality of Group Decision-Making

	All Treatments (1)	All Treatments (2)	Social Media Platform (3)	Social Media Platform (4)	Content Moderation (5)
PL-FL	-0.0611	-0.0782*	-0.0825*	-0.0842**	-0.0847*
	(0.050)	(0.042)	(0.044)	(0.043)	(0.044)
P2P+PL-FL	0.0253	0.0263	0.0263	0.0156	0.00710
	(0.054)	(0.054)	(0.053)	(0.053)	(0.055)
P2P-FL+PS	-0.0107	-0.0177	-0.0174	-0.0182	-0.0157
	(0.055)	(0.044)	(0.044)	(0.042)	(0.042)
PL-FL+PS	-0.0552	-0.0498	-0.0573	-0.0588	-0.0617
	(0.046)	(0.043)	(0.044)	(0.045)	(0.046)
P2P+PL-FL+PS	0.0107	0.00789	0.00208	-0.00580	-0.0122
	(0.056)	(0.046)	(0.047)	(0.048)	(0.050)
NONE	-0.0573	-0.0513			
	(0.046)	(0.040)			
TR	0.0470	0.0152	0.0170	-0.00777	
	(0.048)	(0.038)	(0.039)	(0.041)	
MIS	-0.101**	-0.142***	-0.117***	-0.0863**	
	(0.046)	(0.034)	(0.037)	(0.040)	
Information Purchased		0.0211***	0.0217***	0.0243***	0.0241***
		(0.002)	(0.002)	(0.002)	(0.003)
Number of Connections			0.0627***	0.0660***	0.0864*
			(0.022)	(0.022)	(0.045)
Information Withheld				-0.0102	-0.0113
				(0.007)	(0.008)
Information Wasted (Misinfo)				-0.0194***	-0.0297***
				(0.006)	(0.009)
Posts Fact-Checked					0.00672
					(0.012)
Order	0.0384	0.0743***	0.0834***	0.0847***	0.0762***
	(0.027)	(0.022)	(0.023)	(0.023)	(0.027)
Learning across Period	-0.00926	0.00981	0.00977	0.0102	0.00623
	(0.013)	(0.013)	(0.014)	(0.014)	(0.017)
N	1,926	1,926	1,680	1,680	1,188

Notes: Standard errors are in parentheses.

$*p < 0.10, **p < 0.05, ***p < 0.01$

C.2 Welfare

We also analyze the results in a linear panel regression analysis with the standard errors clustered at the group level. Table 7 presents the results for the models we estimate. The dependent variable of interest is the total group payoff. In columns 1 and 2, we include data from both the fact-checking and information-sharing protocols. The independent variables include a dummy for all the protocols; the excluded category is the peer-to-peer fact-checking when posts are flagged. Also, we control for the order in which the protocol was introduced and learning across periods. In column 2, we include a variable to account for the total information purchased by groups.

In columns 3 and 4, we include data from all protocols where a social media platform was present to separate the platform's effect on group decision quality. In column 3, in addition to the total information purchased, we control for the average number of connections on the platform. In column 4, we include two variables that account for the nature of information shared on the platform. First, units of information that are wasted due to misinformation (where the corresponding post contains misinformation). Second, we include a variable that accounts for the units of information strategically withheld. In column 5, we only include data from the fact-checking protocol and include a variable to account for the total number of fact-checked posts.

The fact-checking protocols under the two consequences leads to higher total welfare relative to the MIS baseline where sharing misinformation is possible but there is no option to fact-check. These results are robust even after we account for the increase in the total information purchased in the presence of a social media platform. Not surprisingly, when we analyze the data in the presence of a social media platform, we find that having more connections on the platform increases total welfare, whereas units of information purchased wasted to misinformation lower the total welfare. Comparing the two consequences of fact-checking, we do not find any significant difference in welfare.

Table 7. Total Welfare

	All Treatments (1)	All Treatments (2)	Social Media Platform (3)	Social Media Platform (4)	Content Moderation (5)
PL-FL	-10.89	-12.95	-13.86	-14.32	-15.09
	(9.851)	(8.921)	(9.773)	(10.009)	(10.654)
P2P&PL-FL	0.933	1.056	1.037	-1.854	-1.195
	(14.080)	(14.116)	(13.972)	(13.722)	(13.739)
P2P-FL+PS	-1.553	-2.403	-2.237	-2.356	-3.345
	(12.760)	(11.499)	(11.222)	(10.689)	(10.532)
PL-FL+PS	-4.645	-3.986	-5.953	-6.701	-6.392
	(9.758)	(9.915)	(10.202)	(10.375)	(10.768)
P2P&PL-FL+PS	-0.530	-0.825	-2.290	-4.343	-4.962
	(11.351)	(10.197)	(10.478)	(11.016)	(11.263)
NONE	-13.50	-12.75			
	(12.460)	(11.442)			
TR	7.233	3.702	3.756	-2.651	
	(9.605)	(8.836)	(9.450)	(9.932)	
MIS	-28.96***	-33.95***	-27.77***	-20.92**	
	(8.121)	(7.987)	(8.400)	(10.589)	
Information Purchased		2.510***	2.461***	3.110***	3.604***
		(0.555)	(0.601)	(0.610)	(0.799)
Number of Connections			15.97**	16.39**	23.39**
			(6.697)	(6.698)	(11.417)
Information Withheld				-3.018	-3.440*
				(1.878)	(1.956)
Information Wasted (Misinfo)				-4.583***	-6.702***
				(1.572)	(2.139)
Number of Posts Fact-Checked					-1.738
					(3.285)
Order	26.89***	31.20***	33.47***	33.74***	32.57***
	(6.289)	(5.887)	(6.010)	(5.966)	(7.002)
Learning across Period	4.063	6.157*	6.734*	6.906*	7.373*
	(3.444)	(3.496)	(3.591)	(3.649)	(4.248)
N	1,926	1,926	1,680	1,680	1,188

Notes: Standard errors are in parentheses.

$*p < 0.10, **p < 0.05, ***p < 0.01$