Managing Online Rumour Proportions During Protests

CAMERON RAYMOND*, Oxford Internet Institute, University of Oxford
P M KRAFFT, Creative Computing Institute, University of the Arts London

In protest contexts, demonstrators and observers on the ground are generally the first to know of new developments, but it is difficult whether from afar or on location to verify information in a timely manner. Social media researchers and companies have suggested a variety of design interventions to address challenges with misinformation on social media in such time-sensitive contexts. Credibility indicators, as warning tags on posts, are one recently explored mechanism. Contributing to the published research on the efficacy of credibility indicators, we develop a laboratory experiment to isolate and estimate the causal effect in a controlled context of credibility indicators on the belief in, and rate at which participants might share, rumours in protest contexts. Our experiments rely on our implementation of a Twitter-like laboratory environment in which participants observe and contribute to a synthetic social media feed about real past protest events. In this controlled laboratory context credibility indicators caused participants to share in greater alignment with what they perceived as accurate, but did not alter perceptions themselves. Thus, those with biased accuracy estimates continued to share misleading information even in the presence of credibility indicators. These results call into question the efficacy of credibility indicators as effective interventions for stemming rumours and misinformation.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in collaborative and social computing**; **Social media**.

Additional Key Words and Phrases: Misinformation, rumour, social media, experimental design

ACM Reference Format:
Cameron Raymond and P M Krafft. 2021. Managing Online Rumour Proportions During Protests. 1, 1 (January 2021), 21 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Demonstrators at protests are increasingly reliant on online communication, through social media platforms and direct messaging applications, before and during protests [6]. Social media allows demonstrators and bystanders to broadcast information in real time, increasing the speed and reach of communication. Having access to high quality information on social media during demonstrations is especially important as such information is generally too timely, local, or hard to verify to be reported elsewhere. Thus, social media represents a key piece of infrastructure for this form of political participation.

Concurrent to the rise of social media have been concerns about its role in facilitating the spread of misinformation, and misinformation's resultant harms. The susceptibility of online platforms to misinformation has received ample attention in recent years. While the bulk of this scholarly attention has been paid to political and scientific misinformation, protest misinformation has

*Both authors contributed equally to this research.

Authors' addresses: Cameron Raymond, cameron.raymond@hey.com, Oxford Internet Institute, University of Oxford; P M Krafft, Creative Computing Institute, University of the Arts London, p.krafft@arts.ac.uk.

XXXX-XXXX/2021/1-ART $15.00
https://doi.org/10.1145/1122445.1122456

the potential to sew confusion and conflict in what are, often, already charged environments [34]. Thus, concerns over the ability for misinformation to undermine protest demonstrations are well warranted. Nonetheless, while protest misinformation has invariably been spread, that only characterizes what is likely a small portion of the information ecosystem during demonstrations. Given the ambiguity that can surround large protests, much of the information shared during demonstrations is at its core unverifiable and thus is best categorized as rumour [31]. However, such ambiguity does not imply that unverified information shared online can be less plausible or pose fewer harms to those who believe and spread it. Accordingly, there are benefits for platforms to manage ambiguous and timely information in a transparent and community-driven manner.

Thus, with respect to offline activism social media presents a double-edged sword [35]. Platforms serve as crucial communication infrastructure during protest demonstrations, but also as the means through which misleading content is spread. This tension is compounded by the intrinsic ambiguity of protests themselves. While recent research has argued that labelling misleading content can reduce its spread [7, 10, 22, 24–27], it is not clear how such interventions would operate in this context, especially considering its more relatively extreme emotional and political valences. Evaluating the degree to which existing interventions improve the quality of information that activists have access to during demonstrations is the high-level objective of the present research. In particular, we contribute a preregistered controlled laboratory study using an innovative experimental design to test the causal impact of credibility indicators—a currently popular design intervention—at reducing the spread of misleading information in ambiguous contexts with participants from North America.

Our analyses leverage a minimal clone of a Twitter-like platform, implemented from scratch for the purposes of conducting behavioural experiments, increasing ecological validity. Through fine-tuned control of the experimental environment, real-world stimuli scraped from Twitter, and carefully chosen attention checks, high quality behavioural data is collected for analysis. Our results show that the treatment made participants share in greater alignment with what they perceived as accurate, but did not alter those perceptions themselves. Thus, those with biased accuracy estimates continued to share misleading information even in the presence of credibility indicators.

2 LITERATURE REVIEW

2.1 Protest Demonstrations and Social Media

Protests are a key mechanism for collective action and are increasingly being organized online [33]. Grassroots movements are able to take advantage of the broad reach that social media provides when mobilizing activists, sharing information, and discussing protest objectives. Alicia Garza, Patrisse Cullors-Brignac, and Opal Tometi helped co-found Black Lives Matter (BLM) and crystalized a movement for racial equity in the United States by using online platforms and the hashtag #BlackLivesMatter as foundational communication infrastructure [5]. While shown to be a useful agenda-setting tool [9] and important for the organization of movements and protests [36], Earl et al. illustrate how online platforms are critical *during* demonstrations [6]. For example, they found that during the 2009 G20 meetings in Pittsburgh, PA, social media allowed protestors to share the location of police — which previously only law enforcement would know — and notify others of police violence. Similar work studying the Gezi Park movement found trace data from Twitter to be predictive of offline political events [37]. Given historic power imbalances between the state and activists, social media can be an important lever for reducing information asymmetries.

Given the speed with which protest dynamics can change, the feedback loop between events and their coverage by traditional news outlets is often too long to be useful to activists during demonstrations. Social media reduces this feedback loop by allowing the real-time broadcasting of information by activists and bystanders themselves [6]. While useful in temporally constrained

situations, it also makes the platforms susceptible to misleading information. Nassauer find that a breakdown in communication is a common predecessor to the emergence of violence during protests [21].

While online platforms are heavily used by activists before and during protests, and having access to pertinent reliable information is important — little work has gone beyond descriptive observational studies. There is a lack of research focusing on how the structure and design of online platforms affect the quality of information that demonstrators have access to. This study aims to help build the evidence base for policy and design interventions that improve information systems in ambiguous contexts like protests.

2.2 Misinformation and its Debiasing

An area of work that has received ample attention in recent years surrounds misinformation, and attempts to prevent its belief and spread. Misinformation is defined as an unintentionally false utterance, while disinformation is deliberately false or misleading [13]. An important area of this literature propose and evaluate interventions at the user-level that aim to reduce misinformation's spread. Such interventions are most frequently implemented through credibility indicators. Credibility indicators are user interface (UI) elements such as icons and text that provide cues regarding the credibility of a piece of information [38].

One objective of misinformation interventions is to prevent the *belief* in misleading information. An influential study by Ecker et al. found that warning participants that a statement was false only partially eliminated participants belief in the statement [7]. This is an example of the continued influence effect (CIE), where individuals rely on false information despite knowing that it is false [14]. Similar complications were found by Pennycook et al., who observed an "implied truth effect" [24]. This is an unintended consequence of attaching credibility indicators to a *subset* of false posts where the absence of a credibility indicator was considered to imply truthfulness, even when it is not warranted. Additional work found adding stance labels, which indicate ideological affiliation, to news articles intensified selectivity bias [10]. Finally, in an overview of the cognitive biases that prevent misinformation's debiasing Lewandowsky et al. cite five relevant design considerations: to provide a contrastive, alternative account of events; to emphasize facts over the original myth; to, whenever possible, add preexposure warnings to misleading information; to provide simple and brief rebuttals to misinformation; and to foster healthy skepticism about information [17]. Such cognitive and human-computer interaction concerns are crucial in forming effective interventions.

Only recently have studies begun to examine their effect on *sharing intentions*. Sharing intentions can act independent of accuracy judgements because the decision to reshare information on social media is multi-faceted, with accuracy being one factor [25]. Other considerations are whether the information is surprising, politically concordant, funny, and interesting [25]. Two studies by Pennycook et al. focusing on the COVID-19 pandemic found that individuals were better at discerning true and false content when explicitly asked about its accuracy as opposed to their intent to share the content on social media [26]. Thus, an argument for why effective credibility indicators work is that they increase the saliency of accuracy in the sharing calculus [25]. This advances a "cognitive-inattention" based account of misinformation sharing over one of ideological motivated reasoning, whereby people are biased against believing information that challenges their political ideology [11]. Pennycook et al.'s work has since been replicated [28]; however, the necessary conditions that are conducive to cognitive-inattention based interventions has not been explored.

2.3 Rumour Systems

While the relevant misinformation literature can provide insights as to possible interventions, a focus on misinformation is too narrow. This is because much of the information shared during protests is unverifiable and is therefore rumour. Indeed, an analysis by Brennen et al. found that 59% of COVID-19 "misinformation" contains true information that is spun, twisted, recontextualized, or reworked — complicating a simple true or false binary [3]. Rumour, separate from misinformation, is best categorized as "unverified information" where the proposition's veracity is unknowable [31]. Thus, it is only in retrospect that a rumour can be labelled as true, false, misinformation, or disinformation [1]. Social media amplifies the capacity for rumours to spread, increasing their scholarly relevance and, of practical importance, provides novel methods for investigation.

Rumour is an important facet of ambiguous social situations, of which protests are but one important example. Focusing on crisis situations, Huang et al. interviewed those who used social media in the aftermath of the 2013 Boston Marathon Bombings [12]. They found that the speed at which information is shared on social media and its decentralized nature are distinguishing factors — making social media especially useful in crisis situations — but also more susceptible to misinformation, intentional or otherwise. Rumour systems contain spatial and temporal elements. Spatial, temporal and emotional proximity are all associated with heightend rumouring during crises [12, 23]. Without credible sources to turn to, Huang et al.'s interviewees turned to source attributes, like if they know or trust the individual sharing the information, or content attributes to determine information's reliability.

Given the intrinsic ambiguity of rumours, it may be tempting to try and stem them all together. Yet, this would be a mistake. Either because the situation is too ambiguous, local, or temporally constrained — rumours are sometimes all that are available. Rumours introduce new discourses into the information ecosystem and many events that are later reported on as news are first reported by individuals as rumours, a process which Shibutani refers to as "improvised news" [31]. Thus, rather than simple gossip — rumours are a powerful mechanism of communication. Time series analyses of rumour systems [18] and rumour stance classifications [19, 39] provide empirical evidence that rumouring acts as a form of collective sensemaking in uncertain environments. As this takes place, Starbird et al. note a shift from speculative rumouring to presenting rumours as factual [32]. A recent study focusing on the case of India's demonetization policy, which severely disrupted existing market practices, argues that rumour is a mechanism through which disenfranchised communities can "collectively bond and make sense of their own place in society" [4]. Similarly, others have argued that informal means of communication can be an integral tool for the preservation of knowledge and safety against oppressive institutions [8]. Thus, there is a distinctive political dimension to rumour as well.

To view rumours as a problem to be solved is incomplete. While problematic if given more credence than is reasonable, rumours are an important mechanism for communication when alternatives do not exist or are not desirable. Thus, rather than eliminating rumours, interventions ought to encourage the thoughtful deliberation that allows for collective sensemaking.

3 THEORETICAL FRAMEWORK

This paper adopts the theoretical framework of rumour systems introduced by Krafft and Spiro [15]. They outline a system-level framework for managing uncertain information. Rather than stifling unverified information entirely, the framework aims to encourage healthy levels of "rumour proportions," relative to normative benchmark representations of intrinsic uncertainty [15]. This expands the study of misinformation beyond the special cases of true or false claims to include the superset of all information, verified or otherwise.

To do so they advocate for a framework that focuses on rumour proportions, in which belief in a rumour is compared with the plausibility of that rumour's proposition. They define a rumour R as a logical proposition about the world. Within the target population, say a hypothetical social media platform, N_R then is the number of people or posts discussing R, of which n_R affirm a belief in R. They define the rumour proportion of R then to be $p_R = \frac{n_R}{N_R}$. They then define three design criteria that attempt to contextualize R with respect to its evidence.

(1) If rumour proportion p_R is at a "reasonable" level, then new evidence for R should increase p_R.
(2) If rumour proportion p_R is at a "reasonable" level, then new evidence against rumour R should decrease p_R.
(3) If rumour $R1$ is more plausible than rumour $R2$ according to publicly available evidence, and both rumour proportions are at "reasonable" levels, then we ought to observe $p_{R1} > p_{R2}$ [15, p. 5].

Evidence levels E for R are quantified adopting a Bayesian approach, and related to the rumour proportion benchmark such that $f : P(R|E) \mapsto p_R$ in a monotonically increasing manner. For example, let us define $f(P(R|E)) = p_R$ and some rumour R which, given known evidence E, has a 60% likelihood of being true ($P(R|E) = 0.6$). In this scenario, the normative benchmark for the system's affirmation rate – that is, the percentage of posts in agreement with the rumour – ought to be 60% ($p_R = 0.6$). Thus, Krafft and Spiro provide a descriptive framework for how rational systems might look, and a normative framework that prescribes what ideal rumour proportions may be in a given context.

4 METHODOLOGY

As discussed in the previous section, Krafft and Spiro's conception of rumour proportions provides a concrete normative framework for the analysis of information online. A rumour proportions framework is especially useful in situations where the veracity of information cannot be assessed deterministically, as is often the case during protest demonstrations. The following experimental design assesses the efficacy of crowdsourced credibility indicators in nudging rumour systems into proportion. This is done through a simulated protest environment where participants in an online laboratory experiment are asked to share posts on a Twitter-like shared social feed discussing two rumours: that protesters are being kidnapped by federal agents ($R1$) and that law enforcement are using digital contact tracing technology to track protestors ($R2$).

All materials and plans for analysis were preregistered on the Open Science Framework (anonymized link: https://osf.io/qrksg?view_only=fb64c8c0438b4e3e96718967e8620417).

4.1 Preregistered Hypotheses

This study applies a rumour proportions framework to a simulated protest rumour system online to study the effect of the introduction of credibility indicators. However, resharing information related to a rumour is not equivalent to affirming a belief in that rumour. As such, we adopt a common qualitative coding of rumour into four categories: affirmations, denials, neutral statements, and questions [15]. Affirmations and denials are of particular interest and form the basis of the study's four confirmatory hypotheses.

- H_1: The presence of credibility indicators in the system will decrease the affirmation rate for rumours with lower evidence levels.
- H_2: The presence of credibility indicators in the system will increase the denial rate for rumours with lower evidence levels.

- H_3: The presence of credibility indicators in the system will increase the affirmation rate for rumours with higher evidence levels.
- H_4: The presence of credibility indicators in the system will decrease the denial rate for rumours with higher evidence levels.

4.2 Experimental Design

As noted previously, one can avoid specifying evidence levels for rumours by comparing them ordinally. This study, then, will limit itself to two rumour types: high evidence rumours and low evidence rumours. For simplicity, this study will take on a 2×2 mixed factorial design, using the presence of credibility indicators as a between-subjects factor (condition assignment: credibility indicators vs. no credibility indicators) and the evidence level of various rumours as a within-subjects factor (evidence level: high vs. low evidence).

With respect to H_1–H_4, there are two dependent variables of interest: the rate at which participants affirm a rumour and the rate at which participants deny a rumour. We can define the affirmation rate (p_R^a) for a rumour (R) as the number of posts that a participant shares affirming R (n_R^a) divided by the total number of posts affirming R (N_R^a). The denial rate (p_R^d) for R can be defined similarly as $p_R^d = \frac{n_R^d}{N_R^d}$. If $P(R1|E) > P(R2|E)$ according to publicly available evidence, and both rumour proportions are at "reasonable" levels, then we ought to observe $p_{R1}^a > p_{R2}^a$ and $p_{R1}^d < p_{R2}^d$ [15].

4.3 Stimuli

The stimuli for this experiment are social media posts and videos derived from Twitter during recent anti-racism protests in the United States. Videos are used to contextualize the social media posts that participants interact with.

The first set of stimuli pertains to a rumour we label $R1$, that demonstrators at a protest were being "kidnapped" by federal agents without identifying themselves. These stimuli were based on a real rumour that originated from a video posted on Twitter showing a demonstrator at a Black Lives Matter (BLM) protest being taken and driven away in an unmarked van. According to the fact-checking website Snopes, no court or judge determined that any protesters were "kidnapped" by federal agents as this implies a crime [16]. However, it is true that protesters were being detained in unmarked vans, and in multiple cases, those who were detained claimed that this occurred without the agents' identifying themselves. As such, Snopes labelled this rumour as a mixture of true and false [16]. Broadcasting state violence is an important function of social media during demonstrations [6]. Thus, $R1$ is defined as a high evidence rumour as it has been supported by third party sources and, legal semantics aside, is highly relevant to demonstrators.

The second set of stimuli pertains to another rumour, $R2$, that demonstrators at 2020 BLM protests were being tracked by police using contact tracing technology originally intended for public health purposes. $R2$ is also based on a real rumour that originated from a video posted on Twitter where a city official used the term "contact tracing" as a metaphor for uncovering any links to organized crime that arrested demonstrators may have had. While spurring fear that law enforcement had access to critical public health data needed to stem the COVID-19 pandemic, this was not the case. No evidence was found in support of $R2$ and the rumour was later refuted by major news outlets as a poorly worded and misinterpreted metaphor [20]. Given this, and the harms that could come from a loss of faith in public health measures among demonstrators, $R2$ will serve as a low evidence rumour.

As previously noted, both rumours originated from videos posted on Twitter, with each "thread" generating a large number of user replies and discussion. Using the public Twitter application

			Rumour:	
		Overall	Kidnapping Rumour ($R1$)	Contact Tracing Rumour ($R2$)
N		167	86	81
Code	Affirms	70	23	47
	Denies	46	32	14
	Neutral	33	19	14
	Questions	18	12	6

Table 1. Summary of coded replies for $R1$, the rumour that police were "kidnapping" protesters (high evidence), and $R2$, that law enforcement are using contact tracing technology to track protesters (low evidence).

programming interface (API), all of the available replies for $R1$ ($N = 251$) and $R2$ ($N = 1293$) were collected in the 24 hours after each video was initially posted, resulting in a total of 1544 replies. After collecting those original Twitter threads on each video, all replies in each of the two threads were coded according to whether they affirm, deny, are neutral towards, or question the rumour, with a fifth category for irrelevant replies or replies that would not have sufficient context to be understood by participants. Each reply was coded twice by the authors, with a week in between coding sessions and blind to previous codes, observing a high degree of internal consistency (Cohen's $\kappa = 0.768$). Conflicting codes were resolved in a third coding session and replies coded as irrelevant ($N = 1377$) were discarded — leaving 167 remaining tweets to be used in the experimental environment. These replies serve as the stimuli that participants interact with in the mock social media environment.

4.4 Procedure

The procedure for this study has four components. After completing a brief consent page, the participants visited a preliminary questionnaire, then the social media protest scenario for $R1$, followed by the social media protest scenario for $R2$, and finally a post-study questionnaire. The experimental setup was piloted and validated for ease and clarity by nine graduate students before deploying. An anonymized version of the experimental environment can be viewed online (see, https://anonymized--pedantic-bohr-7fc231.netlify.app/).

Upon accessing the experiment's webpage, all participants are randomly assigned to one of two conditions (treatment vs. control) and asked to complete a preliminary questionnaire. This measures basic demographic information including age, gender and education — as well as left-right political affiliation, personal involvement in activism, what social movements they affiliate with and what social media platforms they use.

After completing the preliminary questionnaire, participants are asked to watch the video that catalyzed $R1$, downloaded from Twitter, of a demonstrator being detained and driven away in an unmarked van. The participants are informed: "To begin we'd like you to watch a video, originally posted on social media, which appears to show a protestor being taken and driven away in an unmarked van. Please do not watch the video and withdraw from the experiment if you anticipate that content related to abduction or state violence may cause you significant distress. This video created extensive discussion on social media, which you will later interact with. After watching the video click 'Next' at the bottom of your screen to be taken to the next page."

Given that rumours are highly contextual, and thus hard to evaluate in isolation, viewing this media provides important background information to the rumour [29]. After watching the video,

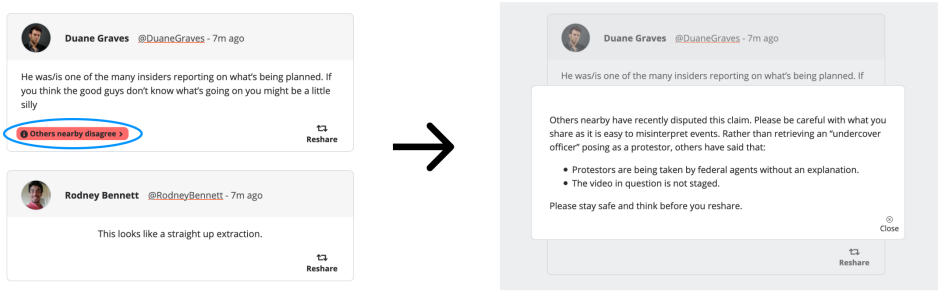

Fig. 1. Contrastive intervention which suggests alternative, local narratives to misleading content.

the participant is given a multiple-choice question asking what the video showed, and a question asking if they had seen the video before. The former question serves as an attention check to ensure that they watched the video. If the participant indicates being aware of the rumour, then it may be the case that their prior evidence level differs from what the stimuli suggests. As such, a participant indicating their awareness of either rumour, or failing an attention check will be used as further exclusion criteria.

Then the participant is given a short narrative and told to reshare social media posts that they would feel are relevant for other demontrators to be aware of. The verbatim instructions were:

> Now that you've watched the video from the previous screen, we would like you to interact with a simulated social media feed as if you were in the scenario described below. Your task will be to reshare social media posts that you would feel are relevant given the video you've watched and the scenario described below. You may do this by pressing the "reshare" button on the bottom right-hand corner of a post. You will be given two minutes to reshare posts, at which time you will move on to the next scenario. Please only reshare posts that you might consider sharing yourself in the described situation. You may end the study at any time without penalty. Please do not refer to outside sources during the study.
>
> Scenario: You are on your way to a protest when you see on your social media feed the video from the previous screen, which appears to show a protestor being taken and driven away in an unmarked van. You recognize the video's background as close to the place where your protest is being held. You check the news, but it's not being covered by your local news station when you check. As you scroll further through your social media feed, you see that this video is the main topic of conversation. Please consider which of the following posts in your feed you might think that other protestors should be aware of.
>
> Press the "Start" button at the bottom of your screen to start the timer and view your mock social media feed.

The participant clicking "Start" reveals the simulated social media feed showing the coded replies, scraped from Twitter, discussing $R1$. Of the 86 replies for $R1$, a random sample of 10 affirming posts, 10 denying posts, 2 neutral posts and 2 questioning posts are shown to participants. Each post also contains a reshare button in the bottom right which the participant can choose to click on, indicating their intent to reshare that information. Posts are assigned a random profile image

(either masc or femme presenting), name, and timestamp, which they are ordered by. Relative to previous survey-based approaches, the mock environment is meant to closely mimic the look and functionality of a typical social media platform, improving ecological validity. Participants assigned to the treatment group will have three-quarters of the posts that deny $R1$ tagged with a credibility indicator. These posts misleadingly suggested that the arrest in the video was staged. Participants then complete the same procedure for $R2$.

Following the protest simulation, participants answer a brief post-study questionnaire where they are asked to estimate the likelihood of each rumour being true. This is to measure the effect of the treatment on participants' perception of the rumours' veracity. They are also asked how they decided which posts to reshare and, if they were assigned to the treatment group, how the credibility indicator factored in to their decision making process. Finally participants are asked if they had any issues completing the study, if they consent to their answers being released in an anonymized dataset, and if they had any final comments.

Credibility Indicator Intervention. In accordance with Lewandowsky et al. [17], misleading posts will be tagged with a credibility indicator that cites contrasting, alternative and local narratives, while noting that it is hard to verify information and easy to misinterpret events during protests. Upon clicking the credibility indicator, a modal window will appear linking to contrastive narratives surrounding the rumour. The modal window approach to credibility indicators has been widely implemented by social media platforms and allows for an unintrusive tag that, when clicked on, provides greater context to the credibility indicator's rationale (Figure 1).

4.5 Participants

Participant recruitment was done via Prolific. American and Canadian participants who use social media were recruited and paid £8.20 per hour (roughly $11/hour) for an estimated 10 minute experiment. After completing the experiment, the exclusion criteria are participants who: do not use social media, were previously aware of either rumour, fail either of the attention checks or who are statistical outliers in completion time. These criteria were preregistered. Attending protests was originally a criterion for inclusion; however, the importance of allowing participants to not reveal such sensitive information led to dropping this criterion. After verifying each participant's results and compensating them, participants' Prolific IDs were encrypted and the original IDs removed from the dataset.

Power Analysis. To ensure that the study has sufficient power, a Monte-Carlo simulation was run using synthetic data that affirms H_1 and H_3. This was done by training a linear mixed-effects model with p_R^a as the dependent variable, and the condition assignment, evidence level and their interaction as fixed effects. The participant ID was incorporated as a random effect to account for repeated measures for high and low evidence rumours [2, 38]. The coefficient of interest, and effect size, is the interaction term and was set to be -0.2 — indicating a one fifth reduction in the affirmation of low-evidence rumours. In order to achieve a power of 0.80 ($\alpha = 0.05$) the study must recruit at least 100 participants. To account for exclusion criteria and participant drop out, and in accordance with budgetary constraints, 189 participants were recruited.

5 RESULTS

After removing participants based on the exclusion criteria, the final dataset contains 107 participants, each exposed to 56 posts, totaling 5,992 "resharing decisions." An exclusion criterion that was omitted in the preregistration, but deemed relevant, is to remove participants who did not reshare any posts. These participants indicated that they act as observers on social media and do not share content. Thus, they are unlikely to contribute to the spread of misleading information.

				Condition:		
		Missing	Overall	Control	Treatment	p value
N			107	54	53	
Age		0	31.6 (12.2)	30.5 (11.7)	32.7 (12.6)	0.33
Education	High school	0	23	11	12	0.57
	Community college		15	6	9	
	Undergraduate degree		51	30	21	
	Graduate degree		15	6	9	
	Doctorate degree		2	1	1	
	None		1	0	1	
Political affiliation	Left	0	30	16	14	0.17
	Centre-left		35	23	12	
	Centre		14	4	10	
	Centre-right		11	5	6	
	Right		5	2	3	
	None		12	4	8	
Attends protests?	No	7	82	41	41	0.67
	Yes		18	8	10	
Gender	Woman	0	56	29	27	0.82
	Man		48	24	24	
	Non-binary		3	1	2	
Affiliated movements	Racial equity	0	91	48	43	0.99
	Climate change		82	44	38	
	Gender equity		81	41	40	
	Free speech		70	36	34	
	LGBTQA+ rights		65	35	30	
	Indigenous rights		59	33	26	
	Labour rights		57	27	30	
	Religious rights		43	23	20	
	Small government		23	13	10	
	Other		3	2	1	
	None		2	0	2	
Social medias	Instagram	0	87	47	40	0.72
	Facebook		82	41	41	
	Reddit		76	39	37	
	Twitter		59	29	30	
	Tiktok		42	26	16	
	Snapchat		40	22	18	
	WhatsApp		1	0	1	
	Clubhouse		1	1	0	
	Telegram		1	0	1	
	Tumblr		1	1	0	

Table 2. Descriptive statistics for participants, segmented by condition assignment. Mean age is reported along with standard deviation in parentheses. p values are calculated using a two sample T-test for continuous variables and chi-square test for categorical variables.

The study ran from June 3, 2021 until June 7, 2021. As indicated in Table 2, the results from Prolific indicate a diverse range of participants across age, gender, political orientation, and affiliated social movements — and are balanced on all measured variables.

Of the 107 participants, only 18 indicated that they had attended a protest in the past two years. As a result, to avoid reducing statistical power non-protestors were included in the analyses. The anonymized data used in these analyses, less those who did not want their responses shared, is available online (see, https://osf.io/6ug4m/?view_only=e7832e62e57f4b438ea92ce2a26aa73f).

First, this section will present the confirmatory analyses, which test hypotheses H_1–H_4 based on Krafft and Spiro's [15] rumour proportions framework. After conducting these *system-level* analyses, exploratory analyses are conducted at the *post-level*. These exploratory analyses investigate how findings from previous studies map to rumour systems and provide the basis for future research.

5.1 Preregistered Analyses

Of interest for the preregistered hypotheses (H_1–H_4) are the affirmation rate (p_R^a) and denial rate (p_R^d) for the high evidence rumour ($R1$) and low evidence rumour ($R2$). For all participants, p_{R1}^a, p_{R2}^a, p_{R1}^d and p_{R2}^d are aggregated and visualized in Figure 2. The left plot visualizes p_R^a for the high

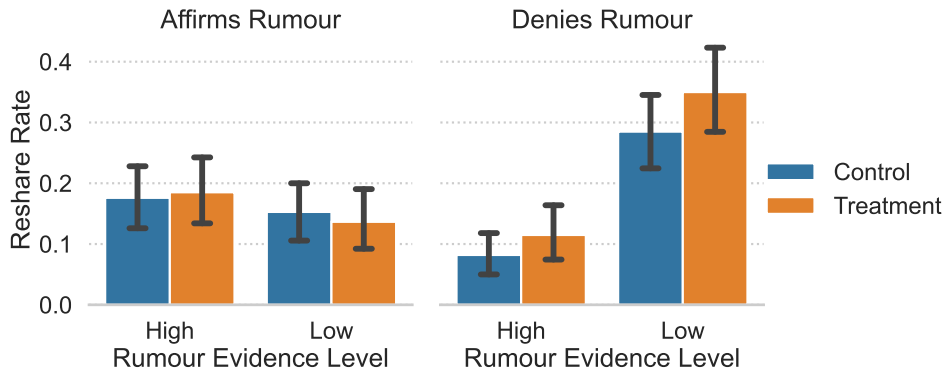

Fig. 2. Percentage of posts shared affirming and denying, high and low evidence rumours for each participant. Error bars show 95% confidence interval.

($M_{(c)ontrol}$=0.18, SD_c=0.19; $M_{(t)reatment}$=0.19, SD_t=0.20) and low (M_c=0.15, SD_c=0.19; M_t=0.14, SD_t=0.18) evidence rumours, and the right plot visualizes p_R^d for the high (M_c=0.08, SD_c=0.12; M_t=0.11, SD_t=0.15) and low (M_c=0.28, SD_c=0.22; M_t=0.35, SD_t=0.16) evidence rumours. These statistics are segmented by condition assignment.

Contrary to what was observed in the coded Twitter discussions, this indicates a system in proportion according to Krafft and Spiro, as $p_{R1}^a > p_{R2}^a$ and $p_{R1}^d < p_{R2}^d$. However, many participants exclusively reshared posts affirming $R2$, or denying $R1$, indicating that interventions may still be useful.

Model Estimation. To formally evaluate H_1 and H_3, a linear mixed-effects model was trained with p_R^a as the outcome variable, and the condition assignment, evidence level and their interaction as fixed effects. As mentioned earlier, there are two measures of p_R^a for each participant, p_{R1}^a and p_{R2}^a. To account for repeated measures, the participant ID was included as a random effect [2, 38]. H_2 and H_4 were similarly assessed using p_R^d as the outcome variable. While the preregistration indicated that demographic controls may be included as parameters, none significantly altered the treatment effect estimate and thus were omitted for clarity in the final regressions. Both models were fit using restricted maximum likelihood (REML) and all p values were estimated via t-tests using the Satterthwaite approximations to degrees of freedom. To account for positive skew in both outcome variables, both p_R^a and p_R^d were square-root transformed. This ensured that the distribution of residual terms followed a normal distribution. As with the dataset used, the scripts used to estimate these models were preregistered.

Regression summary. H_1 and H_3 predicted that the treatment would decrease p_R^a for low evidence rumours and increase p_R^a for high evidence rumours respectively. By investigating the coefficient estimating the interaction between the treatment condition and the rumour evidence level, it is clear that this was not supported ($\beta = -0.06, SE = 0.06, p = 0.33$).

Similarly, H_2 and H_4 predicted that the treatment would increase p_R^d for low evidence rumours and decrease p_R^d for high evidence rumours respectively. This also was not supported ($\beta = 0.01, SE = 0.06, p = 0.89$). Thus, it is clear that the intervention did not have a detectable effect on the sharing of misleading information in the mock-protest scenario. The full table presenting the regression results can be seen in Table 3.

	Dependent Variable:	
	Affirmation Rate (p_R^a)	Denial Rate (p_R^d)
	(1)	(2)
Constant	0.324^{***}	0.185^{***}
	(0.037)	(0.032)
(A) Low Evidence	−0.034	0.301^{***}
	(0.038)	(0.044)
(B) Treatment	0.022	0.052
	(0.054)	(0.047)
(A) X (B)	−0.056	0.008
	(0.055)	(0.062)
Observations	202	202
Log Likelihood	-15.080	2.994
Akaike Inf. Crit.	42.160	6.012
Bayesian Inf. Crit.	62.010	25.862

*p<0.1; **p<0.05; ***p<0.01

Table 3. Linear mixed-effects model demonstrating the effect of treatment on the affirmation and denial rate of rumours. Participant ID was included as a random effect. Parameter standard errors are reported in parentheses.

Attention is all you need? As previously noted, cognitive-inattention based interventions suggest that increasing the salience of veracity — relative to other factors like novelty or political concordance — is a key mechanism by which credibility indicators are effective. Thus, one explanation for why this study's intervention was not effective is that the treatment failed to increase the salience of veracity in the resharing calculus. Using the participant's veracity estimates from the post-study questionnaire one can investigate this possibility. This is measured, for each rumour, on a scale from 1–10, where a value of 1 indicates the participant estimating the rumour to be "completely false" and 10 indicates estimating the rumour to be "completely true." Figure 3 visualizes the relationship between the estimated veracity and p_R^a for each participant and rumour, segmented by condition assignment. As can be seen visually, there is a much stronger association between perceived veracity and sharing behaviour in the treatment group.

To formalize this relationship, a similar linear mixed-effects model was trained, using the square-root transformed p_R^a as the outcome variable; the condition assignment, the perceived rumour veracity, their interaction, and the rumour evidence level as fixed effects; and the participant ID as a random effect. As can be seen in the regression summary in Table 4, there is a significant positive interaction between the condition assignment and the perceived rumour veracity on p_R^a ($\beta = 0.03, SE = 0.01, p < 0.01$). These results indicate that the intervention made participants more likely to share in accordance with their perceived veracity judgements. Thus, in line with Pennycook et al. the intervention increased the salience of veracity in the sharing calculus [25].

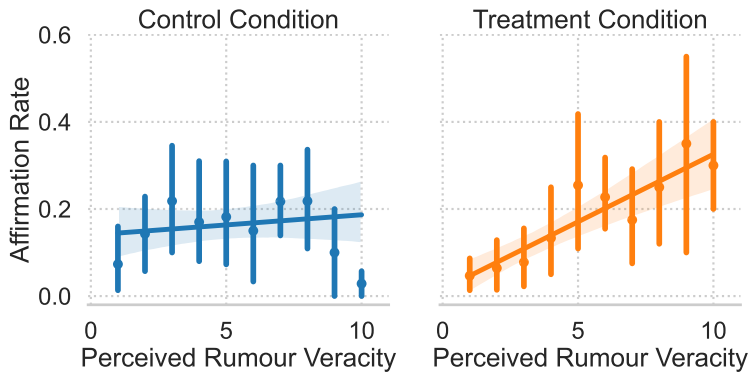

Fig. 3. Scatterplot (with best-fitting regression line) showing the association between perceived rumour veracity and p_R^a. The error band on the regression represents the 95% confidence interval. For clarity, each scatter point represents the mean affirmation rate for that discrete value along the x-axis, with a 95% confidence interval.

However, it also implies that a cognitive-inattention based theory is insufficient to explain the null results.

Initially, it may appear incongruent that the introduction of credibility indicators would make participants share in greater accordance with their veracity estimates, but fail to alter sharing behaviour. However, these findings can be reconciled *if the intervention does not alter the participants' veracity estimates.* Indeed, when training a final linear mixed-effects model using the participant veracity estimate as the outcome variable; the evidence level, condition assignment, and their interaction as fixed effects; and the participant ID as a random effect, we see that there is no effect of the treatment on how plausible the participants perceived each rumour ($\beta = -0.68, SE = 0.68, p = 0.32$).

Summary. This section presented findings from a preregistered experiment exploring the impact of credibility indicators on sharing behaviour during a mock-protest scenario. Of interest were system-level metrics outlined by Krafft and Spiro [15], which capture the rate at which propositions are affirmed or denied relative to their plausibility. In the present study, there is not significant evidence to suggest that the introduction of credibility indicators had a detectable effect on the system-level metrics of interest. Importantly, this is not because participants failed to consider what was accurate when resharing, as the treatment increased the relationship between perceived veracity and rumour affirmation. Instead, a more plausible explanation is that participants were sharing *what they considered to be accurate*, but their veracity estimates remained biased despite the treatment. This raises theoretical and practical implications for the prospect of cognitive-inattention based interventions, which are considered in the discussion section.

5.2 Exploratory Analyses

Finally, we present exploratory analyses, at the post-level, to investigate two research questions. Research question one (RQ_1) asks how temporal proximity affects sharing behaviour during protest demonstrations. The second research question (RQ_2) investigates any evidence of an implied truth effect, whereby attaching credibility indicators to a *subset* of false posts is considered to imply truthfulness in the posts without a credibility indicator.

	Dependent Variable:
	Affirmation Rate (p_R^a)
Constant	0.287^{***}
	(0.053)
Low Evidence	-0.070^{***}
	(0.026)
(B) Treatment	-0.151^{**}
	(0.074)
(C) Perceived Rumour Veracity	0.011
	(0.008)
(B) X (C)	0.032^{***}
	(0.012)
Observations	199
Log Likelihood	-8.799
Akaike Inf. Crit.	31.598
Bayesian Inf. Crit.	54.651

*p<0.1; **p<0.05; ***p<0.01

Table 4. Linear mixed-effects model demonstrating the effect of treatment and perceived rumour veracity on a rumour's affirmation rate. Participant ID was included as a random effect. Parameter standard errors are reported in parentheses.

Given that credibility indicators are attached to individual posts, such analyses are especially pertinent. Of note, only 22.22% of participants in the treatment condition clicked on a credibility indicator to view the alternative, contrasting narrative to misleading content. In total, credibility indicators were only clicked on 3.30% of the time.

Temporal Proximity. Huang et al. [12] found that temporal proximity was associated with rumouring. This aligns with Earl et al.'s work [6], which found that demonstrators relied on temporally relevant information, broadcast on social media in real time. We may expect, then, that posts randomly assigned to be "more recent" via their timestamps to be perceived as more relevant, and thus more likely to be shared. This is especially relevant given the short life cycle with which information is needed during protests. As currently implemented on social media platforms, the labelling of misinformation can take 10-20 hours [30]. Thus, content may proliferate before it can be tagged. This relationship is informally demonstrated in Figure 4, which visualizes the reshare rate for posts as a function of their assigned timestamp.

Implied truth effect. Finally, this exploratory analysis investigates the possibility of an implied truth effect, whereby attaching credibility indicators to a subset of posts is considered to imply truthfulness in the posts without a credibility indicator. As noted by Pennycook et al. [24], and

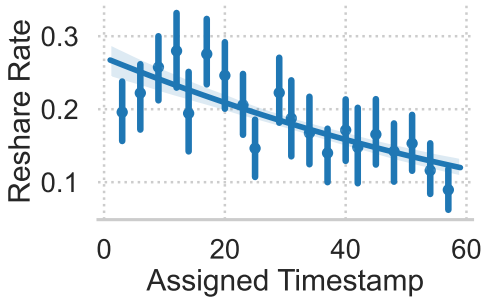

Fig. 4. Scatterplot showing the association between temporal proximity and reshare rate. The error band on the regression represents the 95% confidence interval. For clarity, each scatter point represents the mean reshare rate for each 3 minute increment along the x-axis, with a 95% confidence interval.

commonly known as the "bullshit asymmetry principle", it is much easier for misleading information to be propagated than verified. Therefore, while a community-driven approach to rumour management, where individuals close to the rumour's origin provide evidence for and against propositions, is preferable — the nature of rumours imply a level of ambiguity in attaching credibility indicators to social media posts. Thus, even with a robust, community-driven approach to rumour management, it is still likely that harmful content will go untagged. If there is an implied truth effect then the benefits of credibility indicators could be negated, as individuals may simply shift to sharing misleading posts that fail to be tagged. Given that only three quarters of misleading posts (*i.e.*, affirming a low evidence rumour or denying a high evidence rumour) were tagged with credibility indicators, this study is able to test whether the untagged posts fell prey to the implied truth effect.

Model estimation. The outcome variable in these analyses is the binary choice to reshare a given post or not. We employed a binomial logistic regression (BLR), including the participant ID and post ID as random effects to account for repeated measures for varying items (56 observations per participant). Whether the post is misleading, defined as affirming the low evidence rumour or denying the high evidence rumour, the source profile gender flavour (masc vs. femme-presenting profile image), and timestamp randomly assigned to each post are included as fixed effects. Similar to Pennycook et al. [24], the control condition is used as the baseline and tests for a warning effect with a "warned" dummy that indicates the post being in the credibility indicator treatment and having a warning. Finally, we test for an implied truth effect with an "untagged" dummy that indicates a post being in the credibility indicator treatment and not having a warning. An interaction term between the untagged dummy and whether the post is misleading is also included. As with the dataset, the script used to train the BLR is available on the study's OSF page.

Regression summary. RQ_1 asks how temporal proximity, as operationalized through timestamps randomly assigned to posts, affects resharing behaviour. Confirming what was visualized in Figure 4, there was a statistically significant effect of temporal proximity on resharing odds, with less recent posts being reshared at a lower rate ($\beta = -0.02, SE < 0.01, p < 0.01$).

Finally, RQ_2 posited the possibility of an implied truth effect. By investigating the coefficient for the "untagged" dummy one can see that there is no main effect ($\beta = 0.13, SE = 0.22, p = 0.53$). However, there is a positive interaction effect between a post being untagged and that post being

	Dependent Variable:
	Reshared?
Constant	-1.131^{***}
	(0.195)
Warned	-0.141
	(0.255)
(D) Misleading	-0.784^{***}
	(0.188)
(E) Untagged	0.134
	(0.215)
(D) X (E)	0.453^{*}
	(0.249)
Masc-gender Profile Image	0.011
	(0.076)
Timestamp	-0.023^{***}
	(0.002)
Observations	5,992
Log Likelihood	-2,416.299
Akaike Inf. Crit.	4,850.599
Bayesian Inf. Crit.	4,910.882

*p<0.1; **p<0.05; ***p<0.01

Table 5. BLR mixed-effects model demonstrating the effect of post attributes and the implied truth effect on resharing odds. Participant ID and post ID are included as random effects. Parameter standard errors are reported in parentheses.

misleading ($\beta = 0.45, SE = 0.25, p = 0.07$). Misleading posts that failed to be tagged were reshared at a higher rate than those that were — though both were reshared at a lower rate than posts that were not misleading. Thus, there is weak evidence to support an implied truth effect.

6 DISCUSSION

This preregistered study set out to investigate how the addition of credibility indicators alters sharing behaviour in rumour systems. The case of interest for this study being protest demonstrations in a North American context. In doing so, we examined how credibility indicator interventions affect key system-level metrics. These analyses leverage a minimal clone of a Twitter-like platform, implemented from scratch for the purposes of conducting behavioural experiments, increasing ecological validity. Through fine-tuned control of the experimental environment, real-world stimuli scraped from Twitter, and carefully chosen attention checks we were able to collect high quality

behavioural data for analysis. The results present a nuanced picture with respect to the efficacy of credibility indicator interventions in rumour systems.

The preregistered analyses focused on the rate at which participants affirm or deny rumours in a protest scenario. Our results show that credibility indicator interventions strengthened the relationship between perceived veracity and sharing behaviour. Thus, participants in the treatment were more likely to take into consideration how accurate they thought each post was when deciding what to reshare. Simultaneously, the credibility indicators failed to meaningfully alter sharing behaviour. It was then argued that this is likely a result of the intervention failing to alter participants' veracity estimates.

The exploratory analyses then moved beyond aggregate, system-level metrics to analyze the effect of credibility indicators at the level of the post. These preliminary findings uncovered a significant association between temporal proximity and resharing. Finally, these exploratory analyses found preliminary evidence for an implied truth effect in rumour systems.

6.1 Contributions

This paper developed four key contributions to the study of rumour, misinformation and human-computer interaction — each with methodological, theoretical, and practical implications. First, this study leveraged prior theoretical work to move beyond a strict true or false binary for rumour classification in empirical analysis that is likely problematic. Second, a minimal clone of a Twitter-like platform was implemented for the purposes of conducting behavioural experiments, increasing ecological validity. Third, this study's null results help delineate the settings conducive to cognitive inattention-based interventions. And finally, preliminary empirical evidence is found demonstrating the importance of temporal proximity in rumouring.

Contribution one: Moving beyond a true/false binary. This study adopts a Bayesian framework for information management, first proposed by Krafft and Spiro [15]. This approach contrasts the dominant paradigm of misinformation research, which focuses on verifiably true or false information (*e.g.*, "Pope Francis shocks world, endorses Donald Trump for president"). This differs from unverified information but the two are often conflated, or unverified information is ignored entirely. Rather than stifling all unverified information, this study's rumour proportions approach adds much-needed nuance to the study of misinformation to include the all unverified information, of which proven true or false claims are a special case. Thus, a key contribution of this study is operationalizing a rumour proportions framework in an experimental setting.

Contribution two: A more ecologically valid experimental environment. For the purposes of this study, a minimal clone of a Twitter-like platform was implemented, increasing ecological validity. Given the immense amount of time and money that social media firms spend fine tuning the design of their platforms for engagement, it is important that HCI research mimics these settings as close as is feasible. Thus the experimental environment, which is open-sourced on Github allowing for further refinements, represents a necessary first step towards experimental laboratory studies on rumours and misinformation with greater ecological validity.

Contribution three: The limits of increased attention. The main finding of this study is that credibility indicators made participants share in greater alignment with what they perceived as veracious but did not alter those perceptions themselves. Thus, those with biased veracity estimates of the two rumours still shared misleading information. This resulted in null findings, where the treatment did not alter the system-level metrics of interest. In another light, the degree to which resharing habits reflect known evidence can be viewed as a function of the accuracy of one's veracity judgements and the degree to which veracity factors into one's decision to reshare information, relative to

other factors like novelty or political concordance. It appears the interventions from this study had a meaningful effect on the latter — while having weak evidence, or possibly a much smaller effect, on the former.

This finding poses a complication to Pennycook et al.'s cognitive-inattention based interventions — which presuppose that individuals have access to reasonable inferences based on evidence, but are not taking them into account. However, for a variety of reasons such inferences may be biased. In these cases, increasing the saliency of veracity is unlikely to be effective. This is likely to be the case in rumour systems, where events can unfold, and narratives can be distorted, very quickly — but also in cases where specialized knowledge is required to evaluate information, as with scientific or medical misinformation. In these cases, the intervention must also encourage deliberation and a healthy skepticism as to why some piece of information may be misleading, rather than simply increasing the significance of prior veracity judgements. Thus, a key contribution of this study is to delineate the limits of addressing cognitive-inattention in stemming the spread of misleading information.

Contribution four: The importance of temporal proximity. A final key contribution that this study makes to the field of human-computer interaction and misinformation relates to the role of temporal proximity and rumouring online. Similar to Huang et al., this study outlined empirical evidence that temporal proximity is associated with resharing behaviour. This is of practical import in rumour systems as information is often needed within a short timeframe to be useful [6]. This is especially relevant given the possibility of an implied truth effect. Thus, the ability to quickly evaluate which posts are misleading and which are not is of utmost importance. Given the lag that exists between post creation and tagging, attaching credibility indicators to posts directly may be infeasible [30]. If this is the case then system-wide interventions (*e.g.*, a banner at the top of the social media page asking individuals to post deliberately) may be more practically feasible and relevant.

6.2 Limitations

As for limitations of our work, there were three most notable for discussion. First, as with any laboratory experiment, the stimuli form a limited set. This study used two rumours for its analysis. To reduce the possibility of rumour-specific variation threatening external validity, further work ought to expand the set of rumours analyzed.

In addition, this study only considers information posted in good faith. As with all aspects of social media, coordinated attacks are likely and could undermine public trust in such interventions. This poses practical challenges to the implementation of credibility indicators, as malicious actors could have unverified information flagged unduly. Similarly, the event of incorrectly flagging information as misleading, even if unintentionally, is a real possibility. Inappropriate credibility indicators may undermine their efficacy more broadly. The effect of incorrectly tagging information as misleading on subsequent sharing behaviour was not observed in this study, but is relevant as this is a likely outcome of implementing credibility indicators in rumour systems.

Finally, in environments where official sources are unavailable, individuals will often turn to the user posting the information as an indicator of credibility. For example, people are likely to trust information shared by those they know personally [12]. A future extension of this work could connect the experimental environment with the participant's social media account and assign profiles based on their real-world connections. Of course, making clear that any the participants' responses are private and that posts "shared" in the experimental environment are not actually shared to their friend network would be important.

7 CONCLUSION

Each year, millions flood the streets and attend protests, exercising a core promise of democracy. To do so, they use social media and direct messaging as foundational infrastructure to organize, broadcast updates, and alert others to potential dangers. However, the harms that can result from the spread of misleading information in what are often precarious situations can be equally grave. This tension is compounded by the ambiguity of protests, as information is often impossible to verify. Despite this, little experimental research has been done that asks how the structure and design of online platforms affect the information that demonstrators spread and have access to in protest situations. Interventions that improve the quality of information spread during protests could have tangible benefits for demonstrators and community members.

This preregistered study examined the impact of credibility indicators on the spread of two protest rumours. In doing so, it sought to evaluate how such design interventions affect the rate at which rumours are affirmed and denied, relative to the rumours' evidence levels. This study then presented exploratory analyses, at the post-level, to analyze how source attributes impacted resharing behaviour, and the possibility of an implied truth effect. To answer these, 189 participants from North America were recruited to complete a behavioural experiment simulating a mock-protest environment, which was built from scratch. Participants were segmented into a control group and a treatment group, in which misleading posts were tagged with credibility indicators, and then presented with a high and low evidence rumour. The participant's task was to share information that they would feel is relevant given the protest scenarios.

Results evaluating the preregistered hypotheses suggest that the intervention was successful in focusing participants' attention on the veracity of information. However, the intervention failed to shift participants' perception of what was true. As a result, those in the treatment condition continued to share misleading information.

This study's exploratory analyses then demonstrated that, regardless of the actual content of a post, temporal proximity has an effect on resharing. This combined with evidence to support an implied truth effect, where the absence of a credibility indicator is assumed to imply truthfulness, means that misleading information may spread before it has a chance to be tagged, posing practical concerns.

From a policy perspective, this research suggests that encouraging the thoughtful deliberation of rumours requires more than focusing attention on accuracy. In situations where the veracity of information may be unknowable processes must be put in place that amplify the voices of those close to the information's source, inspire a healthy skepticism of information shared online, and encourage demonstrators to consider the impact of misleading information. Social media platforms are often designed to reduce friction between what one thinks and what one posts. However, in protest demonstrations, the speed of sharing must be balanced with the reflexivity needed for collective sensemaking.

Ultimately, further research is needed to understand how design factors affect knowledge integration and the spread of information. This ought to be done within specific contexts since, as evidenced by this study, different use cases dictate different practical and theoretical constraints. Through an interdisciplinary approach — melding community members, domain experts, researchers, and practitioners — a better information ecosystem can be cultivated.

REFERENCES

[1] Prashant Bordia and Nicholas DiFonzo. 2004. Problem solving in social interactions on the Internet: Rumor as social cognition. *Social Psychology Quarterly* 67, 1 (2004), 33–49.

[2] Markus Brauer and John J Curtin. 2018. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or

within-items. *Psychological Methods* 23, 3 (2018), 389.
[3] J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute* 7 (2020), 3–1.
[4] Priyank Chandra and Joyojeet Pal. 2019. Rumors and Collective Sensemaking: Managing Ambiguity in an Informal Marketplace. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[5] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social media participation in an activist movement for racial equality. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10.
[6] Jennifer Earl, Heather McKee Hurwitz, Analicia Mejia Mesinas, Margaret Tolan, and Ashley Arlotti. 2013. This protest will be tweeted: Twitter and protest policing during the Pittsburgh G20. *Information, communication & society* 16, 4 (2013), 459–478.
[7] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition* 38, 8 (2010), 1087–1100.
[8] Gary Alan Fine and Patricia A Turner. 2001. *Whispers on the color line: Rumor and race in America*. Univ of California Press.
[9] Deen Freelon, Charlton McIlwain, and Meredith Clark. 2018. Quantifying the power and consequences of social media protest. *New Media & Society* 20, 3 (2018), 990–1011.
[10] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–16.
[11] R Kelly Garrett and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1047–1058.
[12] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.
[13] Caroline Jack. 2017. Lexicon of lies: Terms for problematic information. *Data & Society* 3 (2017), 22.
[14] Hollyn M Johnson and Colleen M Seifert. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of experimental psychology: Learning, memory, and cognition* 20, 6 (1994), 1420.
[15] P M Krafft and Emma S Spiro. 2019. Keeping rumors in proportion: managing uncertainty in rumor systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[16] Jessica Lee. 2020. Were Portland Protesters 'Kidnapped' by Federal Officers in Unmarked Vans? *Snopes* (2020). https://www.snopes.com/fact-check/feds-unmarked-vans-portland/
[17] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.
[18] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1751–1754.
[19] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2015. Crowdsourced rumour identification during emergencies. In *Proceedings of the 24th International Conference on World Wide Web*. 965–970.
[20] Sara Morrison. 2020. Minnesota law enforcement isn't "contact tracing" protesters, despite an official's comment. *Vox* (2020). https://www.vox.com/recode/2020/6/1/21277393/minnesota-protesters-contact-tracing-covid-19
[21] Anne Nassauer. 2018. Situational dynamics and the emergence of violence in protests. *Psychology of violence* 8, 3 (2018), 293.
[22] Elmie Nekmat. 2020. Nudge effect of fact-check alerts: source influence and media skepticism on sharing of news misinformation in social media. *Social Media+ Society* 6, 1 (2020), 2056305119897322.
[23] Onook Oh, Manish Agrawal, and H Raghav Rao. 2013. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS quarterly* (2013), 407–426.
[24] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.
[25] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
[26] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

[27] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[28] Jon Roozenbeek, Alexandra L. J. Freeman, and Sander van der Linden. 2021. How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science* 0, 0 (2021), 09567976211024535. https://doi.org/10.1177/09567976211024535 arXiv:https://doi.org/10.1177/09567976211024535

[29] Ralph L Rosnow. 1988. Rumor as communication: A contextualist approach. *Journal of Communication* 38, 1 (1988), 12–28.

[30] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*. 745–750.

[31] Tamotsu Shibutani. 1966. *Improvised news: A sociological study of rumor*. Ardent Media.

[32] Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. Could this be true? I think so! Expressed uncertainty in online rumoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 360–371.

[33] Zachary C Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. 2015. Online social networks and offline protest. *EPJ Data Science* 4, 1 (2015), 1–9.

[34] Thi Trana, Rohit Valechaa, and H Raghav Raoa. 2020. False Claims Hurt: Examining Perceptions of Misinformation Harms during Black Lives Matter Movement. *Research Gate* (2020). https://www.researchgate.net/profile/Thi-Tran-57/publication/352295604_False_Claims_Hurt_Examining_Perceptions_of_Misinformation_Harms_during_Black_Lives_Matter_Movement/links/60c247ad299bf1949f495c42/False-Claims-Hurt-Examining-Perceptions-of-Misinformation-Harms-during-Black-Lives-Matter-Movement.pdf

[35] Zeynep Tufekci. 2017. *Twitter and tear gas*. Yale University Press.

[36] Jeroen Van Laer and Peter Van Aelst. 2010. Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society* 13, 8 (2010), 1146–1171.

[37] Onur Varol, Emilio Ferrara, Christine L Ogan, Filippo Menczer, and Alessandro Flammini. 2014. Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM conference on Web science*. 81–90.

[38] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[39] Li Zeng, Kate Starbird, and Emma Spiro. 2016. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10.