Affordances of credibility on Social Media

Capstone Research Paper: Master of Information Management & Systems, 2020

Submitted by:

Nithya Ramgopal

Vivant Sakore

Vidya Ramamoorthy

Aneesa Chishti

Contents

Abstract

In recent years, there has been increasing interest in researching the spread of misinformation through channels of mass communication, including social media. However, there has not been conclusive research on isolating the affordances of credibility specific to social media that have contributed most significantly to dissemination of fake news. This paper reports the results of one such investigation. Using a randomized controlled trial (RCT) approach, a study was conducted to understand how social media affordances influence and build perceptions of credibility amongst its users. A game design similar to betting was used to simulate real world incentives to the research participants in order to ensure an invested decision making process about whether or not the information presented to them was real or fake. Of the affordances available, the most significant were determined through mixed methods research and the experiment helped isolate the news source as a key factor in helping people evaluate credibility. Other findings and details of analysis are further detailed in the report. The overall goal behind this research was to identify the most important markers of credibility so that platforms may use this information to safeguard key markers or intervene through their remediation, thereby preserving the integrity of civic conversation on their networks.

Literature Review

In recent years, social networks have become breeding grounds for misinformation. Platforms like Facebook, Twitter, Whatsapp, and YouTube have affordances that enable sharing information with great ease and at an unprecedented rate. The ease of publishing online has lowered the barrier of entry, allowing individuals to disseminate information on a massive scale. These platforms have not only encouraged knee jerk posting of “alternative facts” to suit an agenda, but also provided the means to manipulate crowds into buying and further sharing the agenda. Content that misinforms users about crucial matters such as political candidates, immigration policies, etc. is circulated heavily on these platforms and has drastic effects on civic processes like elections.

This research is most concerned with the impact of misinformation on democracy and civic integrity. Social networks often serve as the de facto mode of news consumption for many

people. The role of these platforms in building opinions, influencing civil discourse, and ultimately, election outcomes is undeniable. The spread of misinformation on social media constitutes an interesting topic of research for social psychology and technology. Social media platforms have gained the reputation of being the breeding grounds for misinformation. For example, there was a wave of conspiracy theories about the spread of the Ebola virus on social media in 2014 which led to widespread panic[1].

Of primary interest is the investigation of 'trust' and 'credibility' on social media platforms. The affordances on social media platforms go a long way in influencing how users perceive and interpret information. For example, the decision of showing the source of a news article vs hiding the source could probably have a drastic effect on the spread of misinformation. Similarly, displaying likes, shares, retweets could also have a potential effect on the spread of misinformation.

Some studies on the other hand[2], show that people are inherently more inclined to believe in hearsay and gossip even if the source is explicitly trustworthy. An experiment that assessed the effects of including qualifying words like "allegedly'' in news articles to prevent wrong accusations had no effect on the believability of a certain piece of news.

This asks the question what exactly causes people to trust the information they find on social media? Can we establish or diminish trust by including or changing platform affordances or is trust primarily based on preconceived notions and biases which people already espouse towards a person or topic? Based on some qualitative research, it is likely that the credibility of the source is one of the important factors that establishes trust in information. As correctly pointed out[3], online reputation in the form of ratings, reviews, likes, retweets, etc. are also crucial for establishing trust.

---

[1] Ebola Lessons: How Social Media Gets Infected. http://www.informationweek. com/software/social/-ebola-lessons-how-social-media-gets-infected/a/d-id/ 1307061, March 2014

[2] Eric W. Dolan. PsyPost, "People are strongly influenced by gossip even when it is explicitly untrustworthy"

[3] Cheshire, C. (2011) Online Trust, Trustworthiness, or Assurance?

There is no guaranteed list of initiatives to fix the problem. Real change can be brought about by the compounding effect of changes across tech (such as the use of supervised machine learning algorithms for extracting credibility signals[4]), law and policy, literacy, and user behavior.

Initial Hypothesis

The key goal of this research was to determine and manipulate key affordances in social media to understand how they influenced the perception of credibility of news on the platform. However, in order to do this effectively, it was important to understand exactly how misinformation spreads in the first place.

In order to have a perfect experiment set up, it was important to understand and mimic platform specifics so as to generate fake news articles that were as authentic as possible since participants are less likely to share the news that they themselves don't believe to be credible. Therefore, the team assembled and engineered a series of real-looking headlines based on analysis of current headlines. This was interspersed with real news articles in order to create a randomized set of news posts for research participants to sift through.

The objective of this research was also refined to examine affordances specific to the social platform Facebook that lend credibility to news. In the absence of fossilized biases, what effect do the affordances of a Facebook post have on the perceived credibility of a news headline? Given the constraints of this operation and with an inclination towards a full-factorial design, two key affordances out of the following considerations were tested through these manipulations in our pilot -

1. Image quality
2. Source
3. Engagement
4. Author name, gender, age, race
5. Post recency
6. Privacy settings

[4] Fraunhofer-Gesellschaft. 2019. Software that can automatically detect fake news. https://phys.org/news/2019-02-software-automatically-fake-news.html. (Retrieved January 7, 2020)

Pilot Experiment

The effect of **image quality** of the post and the level of **user engagement** (likes on the post) on perceived credibility of news were examined as part of this pilot study. In order to test if there was any causal effect between perceived credibility and the stated affordances, we conducted a pilot experiment (N=47) with 10 headlines: 5 real and 5 fake, each with 4 permutations as described below.

	Image Resolution	User Engagement
Treatment #1	High	Yes
Treatment#2	High	No
Treatment#3	Low	Yes
Treatment#4	Low	No

Fig 1: The 4 treatment conditions in the pilot study.

The headlines were sourced from credible journals, various fact-checking websites, and websites notorious for generating fake news. We also engineered and created our own fake news that almost looked real. For the sake of simplicity, "likes" alone was considered for engagement. To isolate the influence of affordances of interest, the source, author (display picture and name), location, and any accompanying text were blurred out.

In the real world, a person's credibility or reputation is at stake. If John shares a post on Twitter that is publicly flagged as "Fake", he loses his personal credibility in his social media circle. In this experiment, social perception was modeled in the form of monetary stakes. For each headline, participants were given a chance to bet (virtual $10) on whether it's real or fake. We built this experience as a survey on Qualtrics and the randomization of questions was handled on Qualtrics. The goal was to test:

1. What is the optimal number of headlines?
2. Which affordances would have the strongest effect?
3. How can the usability of the interface be designed to fully support the objective of the experiment?

4. What are the implications of not doing a full-factorial design? How can we pursue full-factorial within budget constraints?

After the pilot we also interviewed all participants for 1) feedback on the pilot, 2) identify key differences based on the frequency of social media use and sources of news. The key observations were:

1. 80% Respondents said image quality did not affect their decision.
2. 66% said they didn't pay too much attention to real/fake before sharing.
3. 50% said they refer to the source of the article and the person who shared it while determining how trustworthy a headline is.
4. 25% said they looked at likes while determining the plausibility of information; especially if liked by trusted 1st-degree connections.
5. Additionally, more than half admitted making decisions based on pre-existing biases.

The results of the first experiment (run on Qualtrics) suggest image quality has no influence on news credibility, which contradicts findings from the initial literature review. Testing the impact of affordances has been tricky because participants invariably fall back on their biases/preconceived notions while guessing whether a headline is real or fake. Eg: In one of the experiments, 25% of the participants believed "*Armenia refused to evacuate citizens from Turkey" because "it seemed like the kinda thing Armenians would do*". This finding has been consistent across methods: survey results, pilot experiments, as well as interviews.

Additionally, the pilot helped identify more significant questions that would further shape the results of the experiment. These questions are listed below, and discussed in detail in the next section:

1. How do we isolate the effect of likes, comments, and shares from each other?
    a. How do we address the correlation between the 3 affordances?
    b. Should we model plausibility as a pricing function using conjoint analysis?
    c. Would a Multistage analysis - try all 27 combinations be the right approach?
2. Within subjects versus across subjects design?
3. Do we need to featurize the topic/theme of news headlines?
4. How do we determine appropriate thresholds for engagement?

    a. User surveys
    b. Averages by publications?
    c. How do we address the correlation between the 3 affordances?
5. Should we enforce a time limit to answer questions?
6. How do we design an effective screening survey? Which questions can help measure candidate relevance?

Experiment Design and Implementation

The findings from the pilot helped refine our research question to: How does the volume of engagement interact with the credibility afforded by the source?

Affordances

The findings of the pilot experiment and subsequent interviews helped 1) eliminate image quality as an affordance for this study, 2) surface the problem of participants viewing posts as templates with static affordances which were perceived to have no bearing on post credibility, 3) reveal participant discomfort with redacted information, which made the experience feel more like an experiment and less like scrolling through their newsfeed. These findings informed three significant pivots in the final design of the experiment:

1) **Better simulation of newsfeed experience:** The first pivot was focused on building a better simulation of the experience of scrolling through one's newsfeed. This included replacing static and redacted/blurred artifacts with randomized values from a fixed set to give the impression of dynamism without introducing significant bias.

2) **Replace image quality with the source:** Since extensive research has already been done on the influence of the source on credibility perceptions, we were hesitant to use it. However, we saw an opportunity to use comparable credible sources at opposite ends of the political spectrum, and investigate how the credibility afforded by the news source interacts with the credibility afforded by post engagement.

3) **Likes, Comments, Shares:** For engagement, we chose to use only likes from the reactions tray and comments and shares.

These three pivots are discussed in further detail below.

Better simulation of newsfeed experience

Affordance	Value(s)	Constant/Randomize
Race	Caucasian	Hold Constant
Gender	Male	Hold Constant
Age	35-55 years	Randomize
First name	Random sample drawn from the US Census	Randomize
Last name	Random sample drawn from the US Census	Randomize
Post recency	[12, 15, 18, 20, 22] hours	Randomize
Privacy settings	Public	Hold Constant

Fig 2a : Enumeration of Affordances that were held constant by randomly assigning similar values.

To address the problem of a post looking like a template, a set of values was identified for each affordance and randomly assigned values to each headline. We generated 10 fake portrait images of middle-aged white men using *thispersondoesnotexist.com* and assigned a number (1 through 10) to each. A CSV file of most common names for men from the US census was then pulled and split into two lists - first names and last names. According to a study by Visibli[5], a social media analytics platform

	Headlines	author_name	post_recency	face_id
0	Spain wind generation increases by 15% as coro...	(Ryan, Sharp)	15	7
1	Tobacco company Philip Morris starts life insu...	(Sam, James)	12	4
2	Trump Says Energy-Efficient Light Bulbs Make H...	(Victor, Shaw)	12	1
3	Shenzhen becomes first Chinese city to ban con...	(Justin, Stokes)	20	3
4	The U.S. sent nearly 17.8 tons of donated medi...	(John, Walker)	15	6
5	Armenia refused to evacuate citizens from Turkey	(Jack, Grant)	22	8
6	Ethiopia: Courts not to handle domestic violen...	(Jordan, Barry)	15	2
7	Xi Jinping's daughter involved in 5 million yu...	(Nathan, Gordan)	15	9
8	Mongolia bans coal exports to China amid Covid...	(Scott, Lucas)	20	10
9	Korean teenager makes the most of the lockdown...	(Michael, Bower)	18	5

[5] https://therealtimereport.com/2011/04/25/facebook-posts-receive-50-of-likes-within-1-hour-20-minutes/

that ran a study on 200 million Facebook users, 95% of the engagement with a post happens within 22 hours of posting. We used this data point to create a set of values for post recency. To each headline, we assigned a randomly selected profile picture, time of posting, first and last name. Finally, the privacy setting for all was held constant as “public”.

Replace image quality with the source

A study by Allsides[6] used blind surveys to visualize the spectrum of political bias in the media. 600 media outlets and writers were positioned on a 5 point scale (Left, lean left, neutral, right lean, right) and ranked based on a track record of reporting facts, opinions, persuasion, propaganda, and/or outright fabrication. We chose the New York Times (Lean left) and Fox News (Lean right) because of comparable popularity on Facebook, both having 17-18 Million Facebook followers. We also considered keeping one neutral source, perhaps the BBC or the Economist, but that would increase variations to be tested per headline.

Fig 3: The spectrum of political biases in media outlets, Allsides.

Likes, Comments, Shares

Facebook has 3 forms of engagement; Reactions, Comments, and Shares. The first decision was to assess whether we wanted to examine the effect of each type of engagement which would be possible to do using conjoint analysis. We very quickly ran into scaling problems, since for each type of engagement we would also have to define low and high thresholds. The second decision was whether we wanted to use all reactions or restrict to only Likes. To keep the experiment simple yet powerful, we defined engagement as the combination of Likes,

[6] https://www.allsides.com/media-bias/media-bias-ratings

Comments, and Shares collectively. The next questions were, what does high engagement look like on average? What about low engagement? What factors determine how much engagement a post gets? What does the distribution across likes, comments, and shares look like? To answer these questions we used two methods:

1) Findings from other studies
2) Eye-balling of our own Facebook feeds to develop an intuition

Based on a study conducted by Quintly[7], the interaction rate with posts by a Facebook page is 0.08% (2014), since this has declined over the years, the estimate for 2020 would be ~0.05%. For a page with > 10 Million Fans, the engagement distribution is 90% Likes, 2% Comments, and 8% Shares. The engagement numbers for popular “average” posts across publications is given below:

	Fans	Engagement	At 22 hours	Likes	Comments	Shares
NYT	17,000,000	8500	8075	7268	323	888
BBC	51,000,000	25500	24225	21803	969	2665
Fox News	18,000,000	9000	8550	7695	342	941
Economist	9,000,000	4500	4275	3848	171	470

Fig 4: Average engagement thresholds for various sources.

Extending this to posts shared by Facebook users, not pages, a cohort of users was assumed with 1500+ friends and followers to estimate ballpark high and low engagement numbers. We triangulated these numbers with observational data collected from monitoring Facebook using freshly created profiles as well as our existing profiles. The resulting threshold ranges were:

Engagement level	Likes	Comments	Shares
High	110 - 200	10 - 22	12 - 29
Low	3 - 9	1 - 3	1 - 6

Fig 5: Average engagement thresholds for various sources.

[7] https://www.postplanner.com/facebook-data-on-fan-page-performance

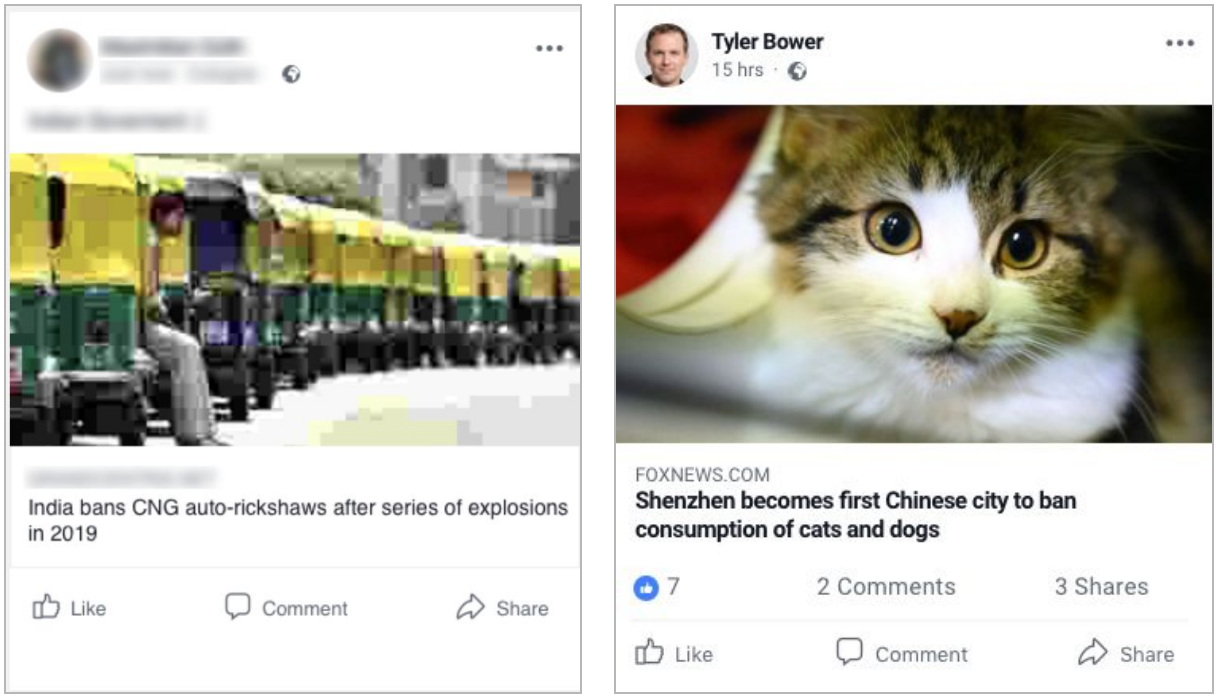


Fig 6: Blurred Facebook post mockup from pilot study (left) and revised post mockup used in the final experiment.

Headlines

The headlines were sourced from various national and international news publications, fact-checking sites such as Snopes, and fake news websites known to generate disinformation. Some of the fake headlines were also created by partially modifying real news. Initially, 50+ headlines were tested without the context of a Facebook post. Headlines were added to a Google Form and circulated among a convenient sample (N=30). Participants were asked to answer two questions for each headline,

1) Is it real or fake?

2) How confident are they about their answer (0%-100%).

From this survey, we picked 10 headlines that had balanced answers (half perceived as fake, half perceived as real) and ~50% confidence on average. The underlying assumptions were,

1) These are not commonly viewed headlines within the public domain,

2) Participants did not have strong preconceived biases about these topics.

The final step was to minimize bias due to variation in the headlines themselves. This meant rejecting any headline that did not fall under the COVID-19 theme, and specifically geopolitical. After filtering based on this theme we found our real headlines on average had a higher word count and more numbers. This was addressed to balance word and character count across all headlines.

Final Headlines	Source	
Spain wind generation increases by 15% as coronavirus lowers electricity demand	Power Technology	REAL
Japan PM pushes for controversial coronavirus cure, Avigan, known to cause birth defects	WSJ	REAL
"COVID-19 was artificially created in a lab", says Nobel Prize winner Luc Montagnier	Snopes	REAL
Shenzhen becomes first Chinese city to ban consumption of cats and dogs	BBC	REAL
The U.S. donated 17.8 tons of medical supplies to China to combat the coronavirus	US Dept. of State	REAL
Coronavirus: Armenia refused to evacuate over 10,000 citizens stranded in Turkey	NA	FAKE
Ethiopia: As coronavirus cases rise, courts dismiss thousands of domestic violence cases	NA	FAKE
Veteran Bollywood actor, Irrfan Khan, of 'Life of Pi,' succumbs to COVID-19	NA	FAKE
Mongolia temporarily bans coal exports to China to help contain COVID-19	NA	FAKE
Bulgaria: President Radev resists a unified EU action to overcome the coronavirus crisis	NA	FAKE

Fig 7: Final selection of news headlines used in the experiment.

Prototypes for Facebook Posts

For each of the 10 headlines 4 combinations were created, one for each quadrant as shown in the example on the below.

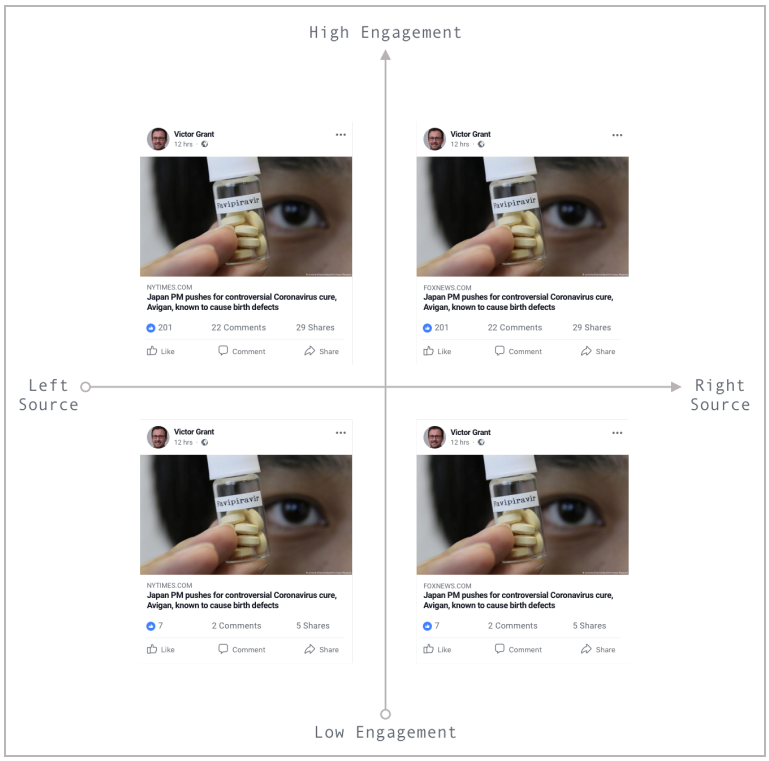


Fig 8: The 4 permutations for each headlinem using source and user engagement.

Headline	Char count	Randomized Constants	Source	Engagement
The U.S. donated 17.8 tons of medical supplies to China to combat the coronavirus	81	Max Rice, 20 hrs, face#6	NYT	Low
			NYT	High
			FOX	Low
TRUE			FOX	High
Coronavirus: Armenia refused to evacuate over 10,000 citizens stranded in Turkey	80	Jim Barry, 15 hrs, face#8	NYT	Low
			NYT	High
			FOX	Low
FALSE			FOX	High

Fig 9: Comparison of the components of a one real, and one fake headline.

Survey Design

Consent

In adherence with IRB requirements, we restricted our study to only US residents who were at least 18 years of age. The consent page of the study offered an overview of the motivation without offering details that could prime participants and potentially compromise the reliability of findings. Participants were informed that the study was interested in examining the spread of fake news on social media platforms like Facebook, and aimed to assess their ability to identify fake news. Additional information about average time, wages, voluntary nature of participation and confidentiality was also presented. As a cautionary measure, direct contact of the principal investigator was also shared.

Setting context

After recording consent, participants were given instructions for the survey. They were informed that they would be shown a mix of real and fake news posts shared by some users on Facebook, and then asked questions about each. To help familiarize them with the format, sample questions along with instructions about timers were also given.

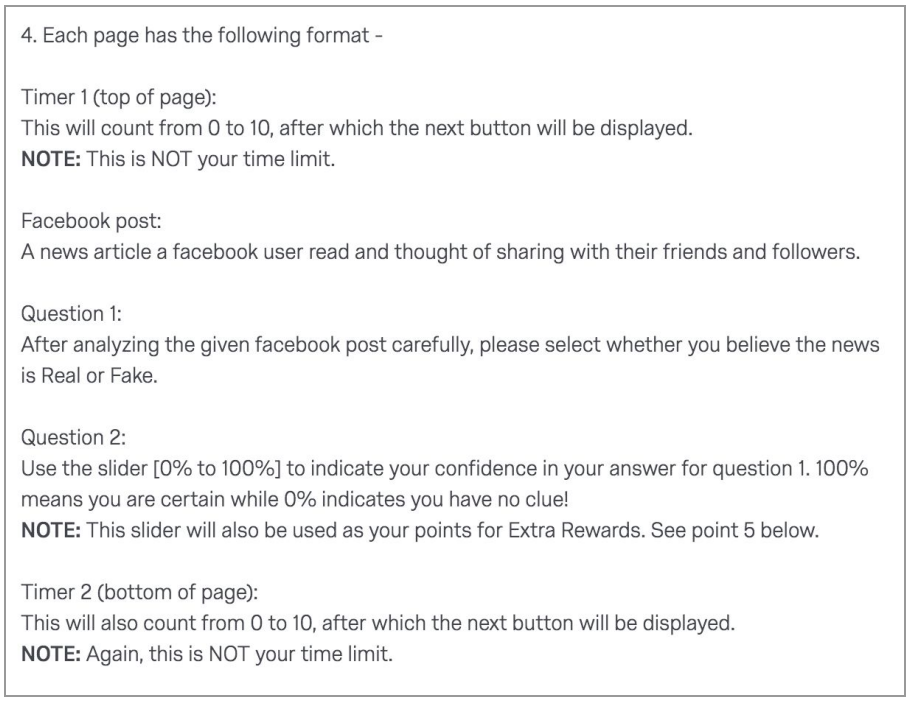


Fig 10: Screen capture of the instructions on the Qualtrics Survey.

The Facebook Posts

Each question on the survey had a variant of one of the 10 headlines rendered as a post in a user's Facebook newsfeed, followed by a question to assess whether the news was real or fake. As part of our experiment design we did not want a simple binary outcome variable, and thus decided to go a step further and ask respondents to also tell us how confident they were of their assessment. This additional information was recorded using a probability slider from 0 to 100, with increments of 10. The decision to not to use increments of 1 or 5 came from :

1) The assumption that the increased granularity would not offer significant gains in terms of data quality,

2) Good usability standards of reducing choices presented to users.

The interactivity of posts (hyperlinks on source and "share" icon) was also considered as a way of measuring sharing intent as well as a subject's tendency to fact-check before sharing. Despite the rich insights post interactivity would offer, the constraints of the study compelled us to push interactivity to a follow-up study, and keep the current scope narrow.

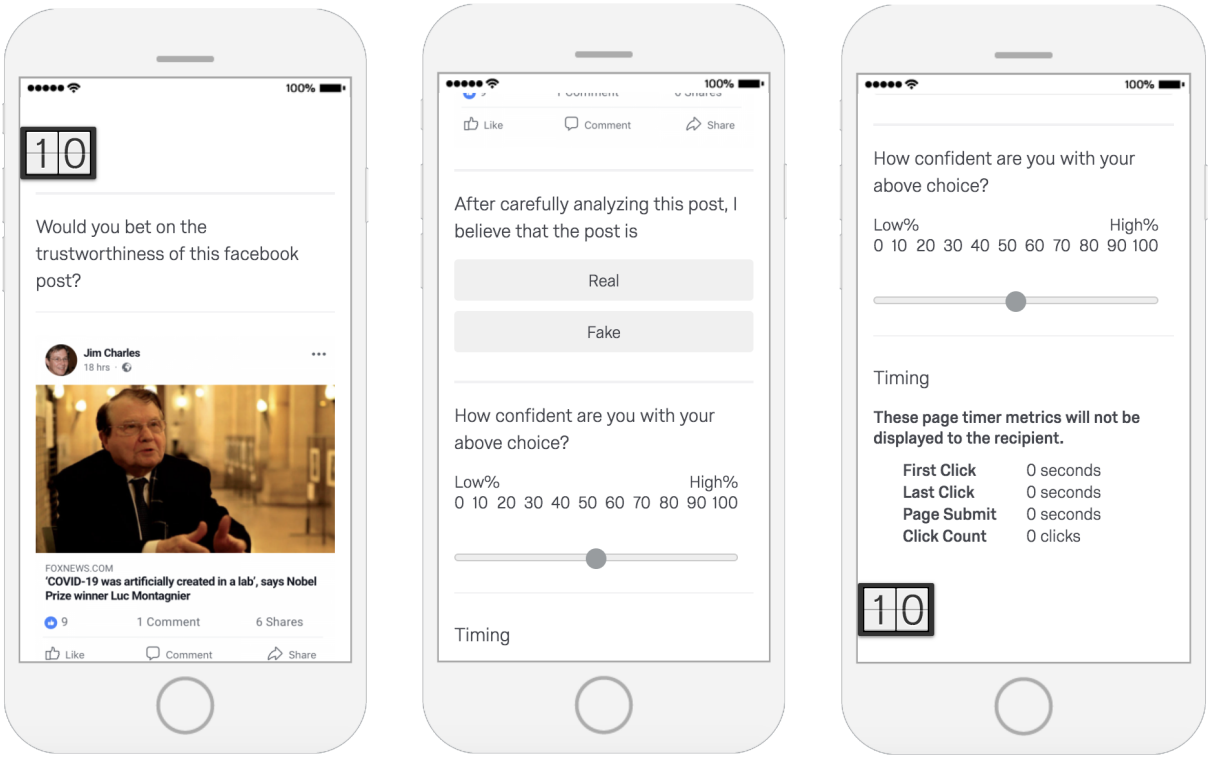


Fig 11: Screen captures of the Qualtrics Survey used in the final experiment.

Ensuring balanced treatment distribution

The survey was hosted on Qualtrics because of its customizable interface and strong randomization controls. The design of the experiment required each test subject to receive a randomized but balanced distribution of the 4 variants across headlines. This design is illustrated in the table below:

treatment_ID	Source	Engagement
1	Foxnews	High
2	Foxnews	Low
3	NYT	High
4	NYT	Low

Fig 12: The 4 treatment conditions, combining source and user engagement.

user_ID	avigan	bulgaria	catmeat	domviolence	donation	evacuation	khan	mongolia	nobel	spain
xxxxxxx	1	4	3	2	1	3	4	1	4	2
xxxxxxx	4	3	2	4	1	4	3	3	2	1
xxxxxxx	2	4	4	1	1	3	1	2	2	3
xxxxxxx	1	3	2	3	4	4	3	2	1	4
xxxxxxx	3	1	1	2	3	1	4	4	3	2

Fig 13: Intended treatment distribution across questions for respondents.

An examination of the survey data from 1,500 respondents confirmed the randomization was also balanced across each question, which means each question was shown to roughly equal number of respondents.

question_name	treatment_id	ProlificID	treatment_code	variable_a	variable_b
avigan	1	354	354	354	354
	2	357	357	357	357
	3	352	352	352	352
	4	354	354	354	354
bulgaria	1	355	355	355	355
	2	351	351	351	351
	3	355	355	355	355
	4	356	356	356	356
catmeat	1	358	358	358	358
	2	353	353	353	353
	3	358	358	358	358
	4	348	348	348	348

Fig 14: Post experiment pivot table showing distribution of participants across each treatment condition for a subset of questions.

Timing

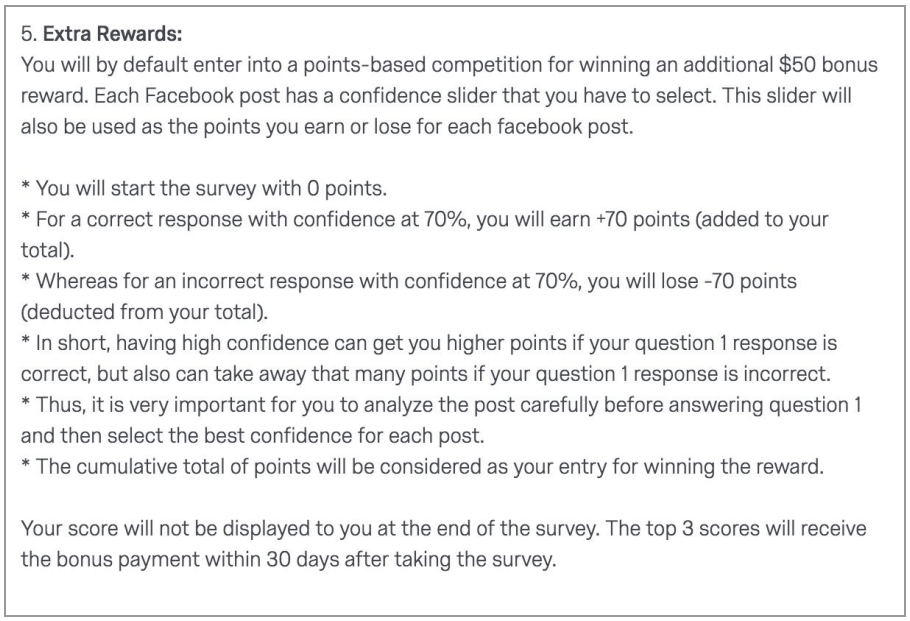
No upper-time limit was imposed at the question or survey level, but respondents were required to spend a minimum of 10 seconds on each page before the "Next" button was displayed. On average, participants spent about 20 seconds analyzing each Facebook post, so the 10 seconds minimum was introduced to ensure participants weren't randomly clicking options, and were taking the time to actually consume the information presented. The obvious concern of not adding an upper-limit was that it would raise the risk of cheating. This might allow someone to do a Google search and validate the correctness of all headlines.

Disincentivizing Cheating

Since there was no upside to guessing correctly, the best way to disincentivize cheating without imposing a time limit was by priming respondents upfront by assuring them that we were not looking for right or wrong answers. Respondents were encouraged to thoroughly analyze the displayed Facebook posts to reach any conclusion while answering the questions. As an extra precaution, the threat of disqualification was also used.

Extra Rewards

All respondents were by default entered into a points-based competition for winning an additional $50 bonus reward.



5. **Extra Rewards:**
You will by default enter into a points-based competition for winning an additional $50 bonus reward. Each Facebook post has a confidence slider that you have to select. This slider will also be used as the points you earn or lose for each facebook post.

* You will start the survey with 0 points.
* For a correct response with confidence at 70%, you will earn +70 points (added to your total).
* Whereas for an incorrect response with confidence at 70%, you will lose -70 points (deducted from your total).
* In short, having high confidence can get you higher points if your question 1 response is correct, but also can take away that many points if your question 1 response is incorrect.
* Thus, it is very important for you to analyze the post carefully before answering question 1 and then select the best confidence for each post.
* The cumulative total of points will be considered as your entry for winning the reward.

Your score will not be displayed to you at the end of the survey. The top 3 scores will receive the bonus payment within 30 days after taking the survey.

Fig 15: Terms of the experiment highlighting incentives/rewards for participants

Sampling and Data Collection

Various crowd-sourcing platforms were considered for hosting this survey. After consulting with other researchers, the two main options were MTurk and Prolific. The following factors were considered while deciding between the two platforms.

1. **Platform-specific features**

   Prolific is a platform that is specific to running studies. The interface, design and pricing has been optimized to ensure that research studies get high quality data and more motivated participants. MTurk on the other hand has a wide variety of tasks such as data collection, transcribing, annotations etc.

2. **Pre-screening participants**

   Prolific has pre-screened workers who have been vetted in terms of quality of work and qualifications. MTurk on the other hand does not use any pre-screening criteria. Alternatively, MTurk has a "Master" category which includes workers who have consistently performed well in prior tasks and have good ratings. Both platforms also offer various options for pre-screening the participants in terms of demographics etc. This was an important consideration since it was not possible to incorporate relevant attention checks into the survey itself to reject invalid responses. Based on the online reviews on various forums like Reddit.com it was determined that the quality of workers was better on Prolific, particularly for online studies.

3. **Wages and Costs**

   Prolific has a higher average hourly wage and also has a higher minimum hourly wage that needs to be paid to participants. MTurk does not set any restrictions on wages. Both platforms impose a service charge based on the reward that is paid to the participants. Prolific charges approximately 20% of the reward. MTurk charges 20% of the reward as service charge. In addition to this, MTurk levies an additional 20% charge for studies with more than 10 participants, and an additional 5% for employing the participants from

the ‘Masters’ category. Prolific also offered a $150 joining bonus for studies of $250 or more.

4. **Time limits**

   Prolific automatically sets an upper limit for a participant to spend on a survey. MTurk also has a time limit for each participant to complete a task.This prevents the possibility of a participant spending excessive time on the survey.

5. **ID matching**

   Both platforms have ways to uniquely identify workers and confirm the completion of a task. Prolific redirects all the participants to a link on their website and logs the participants’ ID. MTurk requires a unique completion code that is generated by Qualtrics for submitting the survey. MTurk has a slightly more tedious process which involves generating random IDs and logging the MTurk WorkerID. After the survey has been set up according to the platform guidelines. The participant ID on Prolific/MTurk can be matched to identify the responses for each participant.

The survey was hosted on Prolific on May 7th 2020, at 7:43 am (PST). The survey was hosted for 1,500 participants in 3 phases. The compensation rate was $10/hr. This compensation rate was determined to be ‘good’ by the compensation scale on Prolific. Respondents who spent more than a specified amount of time were automatically rejected and new participants were recruited for the study. It took approximately 9 hours to get 1,500 responses.

Some participants were in the UK and therefore they could not participate in the study since the IRB approval was only for US residents. As a result. 87 respondents in the first phase did not consent to take the survey as they were not eligible. This problem was rectified in the second phase by pre-screening participants to only include US residents.

The survey had 1500 respondents and the following preliminary checks were performed before analysing the data.

Check	Description	Number of Respondents
Consent	Respondent did not meet the eligibility criteria	79
Duplicate Response	A respondent completed the prolific with the same ProlificID more than once. Only the first response was kept	3
Incomplete Response	The respondent did not complete survey	0
Invalid Prolific IDs	The respondents Prolific ID was incorrect	1

Fig 16: Tabulated rejected survey responses.

After the preliminary checks, there were 1417 valid responses in total.

Results and Discussion

Quantitative Analysis

1. **Participant Score Distribution**

   Using the data from Qualtrics, the performance of participants was evaluated. All participants started with a score of zero. For every correct answer, they were rewarded points and for every wrong answer they lost points. The exact magnitude of points that they won was determined by the confidence level. For example, for a correct answer with 70% confidence, they were awarded 70 points and for a wrong answer with 70% confidence, they lost 70 points.

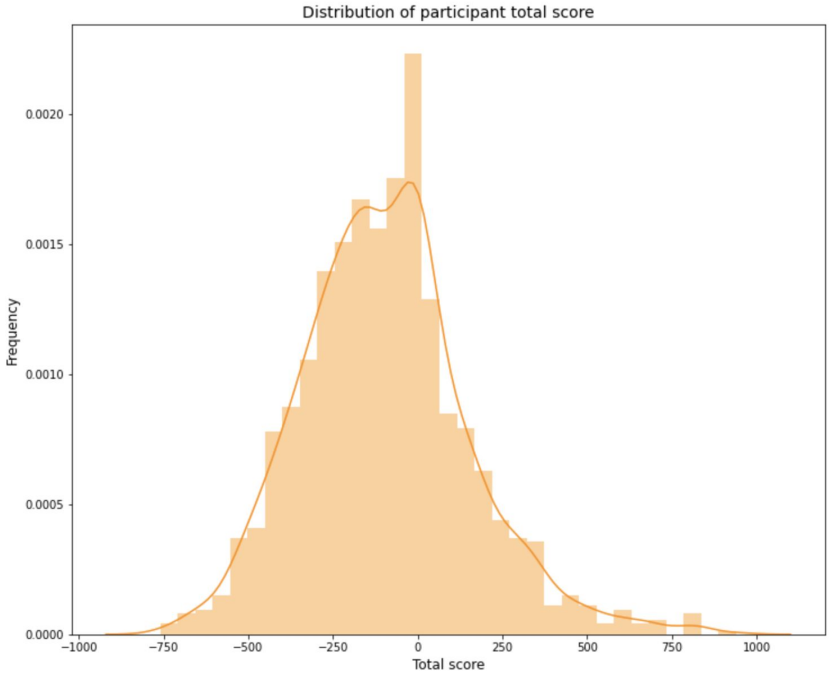


Fig 17 : Graph showing the distribution of participant scores on the survey. A higher score indicates The participant got more answers correct with a high confidence.

The figure above shows the distribution of the final scores. Based on the distribution, it is evident that very few participants were able to identify fake news with high confidence. It is also important to note that more than half the participants got a score of less than 0. This indicated that many participants got more answers wrong than right.

2. Confidence Score Distribution

The figure below shows the distribution of the scores assigned by the participants to the posts that they were shown. The X-axis extends from a score of -100 to +100. A score of -100 indicates that a participant had indicated that a post was fake with a high confidence and a score of +100 indicates that a participant had indicated that a post was real with a high confidence. A score of 0 indicates that the participant made a random guess.

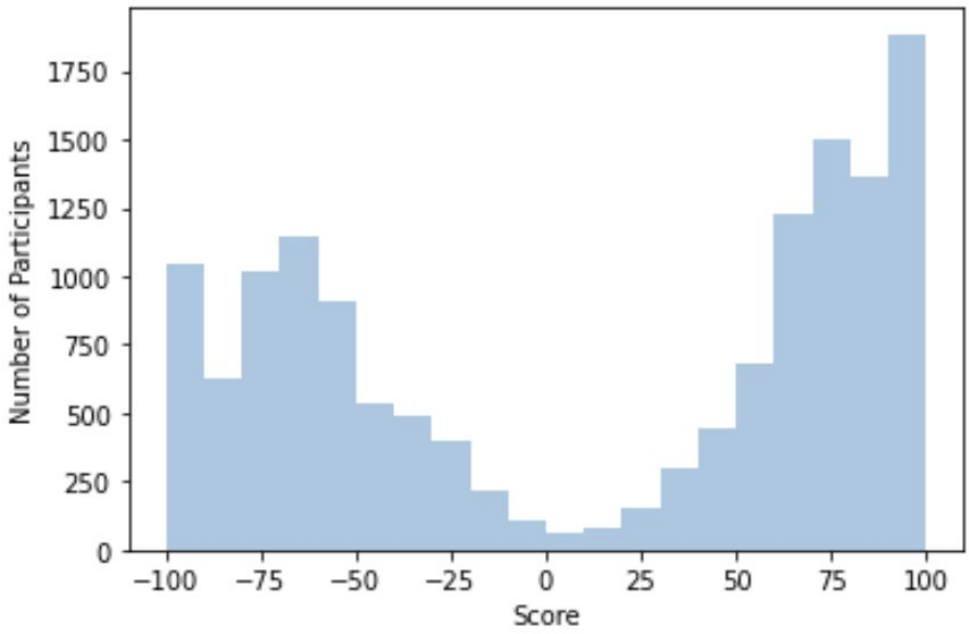


*Fig 18: Graph indicating participant confidence about responses (+100 is high confidence in post being true, -100 is high confidence in post being false, 0 is low confidence in response_*

The distribution in the figure above indicates that most participants had a high level of confidence about their responses. Very few participants made random guesses.

3. Linear Regression

A linear regression model was trained on the data. The results of the linear regression can be found in the table below.

OLS Regression Results

Dep. Variable:	final_score	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	18.99
Date:	Thu, 14 May 2020	Prob (F-statistic):	2.77e-12
Time:	01:30:16	Log-Likelihood:	-80681.
No. Observations:	14170	AIC:	1.614e+05
Df Residuals:	14166	BIC:	1.614e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8668	1.206	2.377	0.017	0.502	5.231
fox_low	-1.5393	1.708	-0.901	0.367	-4.887	1.809
nytimes_high	9.0452	1.706	5.301	0.000	5.700	12.390
nytimes_low	7.3592	1.707	4.311	0.000	4.013	10.706

Omnibus:	65743.887	Durbin-Watson:	1.899
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1664.448
Skew:	-0.170	Prob(JB):	0.00
Kurtosis:	1.356	Cond. No.	4.79

Fig 19: Regression table indicating coefficient for all four permutations of source and engagement

This regression compares the confidence scores assigned by participants across the 4 conditions. When compared to a post with Fox News as a source and high user engagement, a post with New York Times as a source has a 9-point increase in confidence.

Similarly for posts with low user engagement, posts with New York Times as a source have 8.8 more points than posts with Fox News as the source.

There was no statistically significant effect of altering the user engagement, given the source was constant. Fox News posts with high likes, comments, shares did not have more credibility that Fox News posts with low likes, comments and shares.

4. Heterogeneous Treatment Effects

In order to estimate how the treatment effect varies across the different headlines, the regression model was altered to include heterogeneous treatment effects.

OLS Regression Results

Dep. Variable:	final_score	R-squared:	0.034
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	24.84
Date:	Thu, 14 May 2020	Prob (F-statistic):	8.48e-47
Time:	01:38:05	Log-Likelihood:	-40296.
No. Observations:	7097	AIC:	8.061e+04
Df Residuals:	7086	BIC:	8.069e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8668	1.188	2.413	0.016	0.538	5.196
treatment_3	8.1446	1.528	5.331	0.000	5.150	11.140
q_avigan_3	-30.2727	3.582	-8.452	0.000	-37.294	-23.252
q_bulgaria_3	13.2703	3.568	3.719	0.000	6.276	20.264
q_catmeat_3	17.0333	3.554	4.792	0.000	10.066	24.001
q_domviolence_3	13.6214	3.572	3.813	0.000	6.619	20.624
q_donation_3	-19.1983	3.577	-5.367	0.000	-26.210	-12.186
q_evacuation_3	11.4112	3.568	3.198	0.001	4.417	18.405
q_khan_3	1.5665	3.577	0.438	0.661	-5.445	8.578
q_mongolia_3	17.7875	3.554	5.005	0.000	10.820	24.755
q_nobel_3	-25.6420	3.582	-7.159	0.000	-32.663	-18.621
q_spain_3	8.5673	3.563	2.404	0.016	1.582	15.552

Omnibus:	39297.147	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	736.250
Skew:	-0.163	Prob(JB):	1.33e-160
Kurtosis:	1.456	Cond. No.	1.85e+15

Fig 20: Regression results comparing Treatment 1 (Fox News, High user engagement) and Treatment 3 (New York Times, Low user engagement) with Heterogeneous Treatment Effects

The regression results in Fig 20. show that there is a statistically significant difference in credibility due to source. This effect varies to some extent with the content in the headlines. While most headlines have shown a higher credibility score for posts with New York Times as the source, some of them have had the opposite effect. This is perhaps due to the fact that certain headlines seem to inherently contradict the political ideologies that the new source typically conveys.

OLS Regression Results

Dep. Variable:	final_score	R-squared:	0.023
Model:	OLS	Adj. R-squared:	0.022
Method:	Least Squares	F-statistic:	16.89
Date:	Wed, 13 May 2020	Prob (F-statistic):	1.15e-30
Time:	23:51:33	Log-Likelihood:	-40207.
No. Observations:	7073	AIC:	8.044e+04
Df Residuals:	7062	BIC:	8.051e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3275	1.199	1.107	0.268	-1.023	3.678
treatment_4	8.1471	1.541	5.288	0.000	5.127	11.167
q_avigan_4	-23.6836	3.597	-6.585	0.000	-30.734	-16.633
q_bulgaria_4	4.3738	3.587	1.219	0.223	-2.659	11.406
q_catmeat_4	18.9450	3.625	5.227	0.000	11.840	26.050
q_domviolence_4	15.4550	3.592	4.303	0.000	8.414	22.496
q_donation_4	-13.5195	3.587	-3.769	0.000	-20.552	-6.487
q_evacuation_4	9.3424	3.592	2.601	0.009	2.301	16.384
q_khan_4	-1.2774	3.592	-0.356	0.722	-8.319	5.764
q_mongolia_4	14.5255	3.615	4.018	0.000	7.439	21.612
q_nobel_4	-19.1384	3.583	-5.342	0.000	-26.162	-12.115
q_spain_4	3.1243	3.597	0.869	0.385	-3.926	10.175

Omnibus:	36376.607	Durbin-Watson:	1.920
Prob(Omnibus):	0.000	Jarque-Bera (JB):	760.074
Skew:	-0.137	Prob(JB):	8.95e-166
Kurtosis:	1.417	Cond. No.	1.15e+15

Fig 21: Regression results comparing Treatment 2 (Fox News, low user engagement) and Treatment 4 (New York Times, Low user engagement) with Heterogeneous Treatment Effects

The results show in Fig 21. are very similar to the results for the regression model comparing treatment 1 and treatment 3. There is an overall positive change in credibility when the source is changed from Fox News to New York Times, but there is some variation in the effect across headlines.

OLS Regression Results

Dep. Variable:	final_score	R-squared:	0.024
Model:	OLS	Adj. R-squared:	0.022
Method:	Least Squares	F-statistic:	17.26
Date:	Wed, 13 May 2020	Prob (F-statistic):	2.03e-31
Time:	23:54:22	Log-Likelihood:	-40223.
No. Observations:	7084	AIC:	8.047e+04
Df Residuals:	7073	BIC:	8.054e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8668	1.188	2.413	0.016	0.538	5.196
treatment_2	-1.3399	1.530	-0.876	0.381	-4.338	1.658
q_avigan_2	-31.0507	3.560	-8.722	0.000	-38.029	-24.072
q_bulgaria_2	11.8064	3.587	3.291	0.001	4.774	18.839
q_catmeat_2	9.0963	3.578	2.542	0.011	2.082	16.111
q_domviolence_2	16.4218	3.587	4.578	0.000	9.389	23.454
q_donation_2	-13.6336	3.564	-3.825	0.000	-20.621	-6.646
q_evacuation_2	2.4846	3.597	0.691	0.490	-4.566	9.535
q_khan_2	6.0293	3.564	1.691	0.091	-0.958	13.017
q_mongolia_2	18.0470	3.583	5.037	0.000	11.024	25.070
q_nobel_2	-20.5071	3.578	-5.731	0.000	-27.521	-13.493
q_spain_2	-0.0339	3.569	-0.010	0.992	-7.030	6.962

Omnibus:	35777.715	Durbin-Watson:	1.946
Prob(Omnibus):	0.000	Jarque-Bera (JB):	751.848
Skew:	-0.050	Prob(JB):	5.47e-164
Kurtosis:	1.407	Cond. No.	1.06e+15

Fig 22: Regression results comparing Treatment 1 (Fox News, High user engagement) and Treatment 2 (Fox News, Low user engagement) with Heterogeneous Treatment Effects

OLS Regression Results

Dep. Variable:	final_score	R-squared:	0.020
Model:	OLS	Adj. R-squared:	0.018
Method:	Least Squares	F-statistic:	14.14
Date:	Wed, 13 May 2020	Prob (F-statistic):	3.98e-25
Time:	23:53:47	Log-Likelihood:	-40303.
No. Observations:	7086	AIC:	8.063e+04
Df Residuals:	7075	BIC:	8.070e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.9120	1.200	9.923	0.000	9.559	14.265
treatment_4	-1.4752	1.544	-0.955	0.339	-4.502	1.552
q_avigan_4	-24.6458	3.608	-6.831	0.000	-31.718	-17.573
q_bulgaria_4	3.4115	3.599	0.948	0.343	-3.643	10.466
q_catmeat_4	17.9828	3.636	4.946	0.000	10.855	25.110
q_domviolence_4	14.4928	3.603	4.022	0.000	7.429	21.556
q_donation_4	-14.4817	3.599	-4.024	0.000	-21.536	-7.427
q_evacuation_4	8.3801	3.603	2.326	0.020	1.317	15.444
q_khan_4	-2.2396	3.603	-0.622	0.534	-9.303	4.824
q_mongolia_4	13.5632	3.627	3.740	0.000	6.454	20.672
q_nobel_4	-20.1006	3.594	-5.593	0.000	-27.146	-13.055
q_spain_4	2.1621	3.608	0.599	0.549	-4.910	9.235

Omnibus:	39284.009	Durbin-Watson:	1.936
Prob(Omnibus):	0.000	Jarque-Bera (JB):	782.577
Skew:	-0.258	Prob(JB):	1.16e-170
Kurtosis:	1.456	Cond. No.	1.59e+15

Fig 23: Regression results comparing Treatment 3 (New York Times, High user engagement) and Treatment 4 (New York Times, Low user engagement) with Heterogeneous Treatment Effects

The results from Fig 22. and Fig 23. indicate that there is no overall effect of changing the user engagement level. However, for certain questions there is a statistically significant change in the perceived credibility of the post based on the number of likes, comments and shares.

Qualitative Analysis

At the end of the survey, respondents were given the opportunity to provide feedback about any part of the experience. Although this was optional, most respondents chose to give feedback that we were able to categorize into 5 broad themes:

1. **Request for more information**

This was the most recurring piece of feedback. Respondents felt they did not have enough information to assess the trustworthiness of the news headlines presented to them. This was interesting because we received similar feedback during the pilot study, where participants felt the most important information was blurred and they did not have enough to go by. At this stage there was a demand for Source, post author information and the number of likes, comments and shares. During the final experiment, all of these blanks were filled, but the concern remained the same - not enough information. Just the nature of requested information changed to a new set of affordances. These included article excerpts, dates the articles were published, the first paragraph of the articles, hyperlinks to the actual source, different reactions (laugh, angry etc), and expanded view of the comments threads.

2. **Diversity of sources**

This was the second most popular request. Respondents felt the selection of New York Times and Fox News was confusing and misleading. Some felt these choices were looking for particular responses based on partisanship. Others reported that they first look for the source when they encounter a sensational headline, and acknowledge there are plenty of obscure, but legitimate-sounding websites that post 'fake news' that is rapidly disseminated on social media because many people don't check sources. Additionally the views on the sources varied from, “all news was by the 2 most biased news sources in america” to “Fox is usually slanted or only analyzes part of a story”.

3. **Survey usability**

Respondents reported the instructions were too detailed and could have been contextualized better. The question-level timers were polarizing, some felt they helped stay on track while others felt they were distracting and interfered with their ability to focus on the problem at hand. Some key words were not clearly defined at the outset, so participants weren’t always sure of their interpretations. For instance, the word “fake news” itself was perceived as 1) outright fabrication that was actually published and widely circulated, 2) fabricated news that was created for the purpose of this experiment.

For instance, the fake news about Russia releasing 500 lions to terrorize citizens to stay at home during COVID was actually published and widely believed, while news about Xi Jinping's daughter crashing a Ferrari was an event we fabricated. Similarly, participants were not always sure about the terms "trustworthy", "credible" and "misinformation". Is it misinformation by the news source or fabricated information by the person submitting the post?

Limitations

The real headlines we picked were taken from various news sources and therefore it is likely that participants had previously seen them. Since this was not a supervised lab experiment, participants could have cheated. Judging by the feedback offered by our respondents, it is reasonable to consider that the sample may not be representative of the general US population since the participants sourced by a platform such as Prolific might have higher technical proficiency as well as higher media literacy than the average American.Our experiment did not recreate a high-stakes context where believing false news would have any real consequences.
The simulation offered an additional reward of $50 to the participants. This does not simulate a high stakes situation where people have a lot to lose if they are not cautious about sharing misinformation.

Future Work

Examining Other Key Affordances

Misinformation is a vast space and there is a plethora of research that is being conducted in this field to help combat this problem. The findings from the study provide some useful insights about the affordances that influence credibility on social media platforms such as user engagement metrics, image quality and news sources. However, there are still many affordances that remain unexplored. Exploring these additional affordances will give more information about how misinformation spread on social media and enable platforms to make design changes to combat this issue.

Examining Other Platforms

This study exclusively focussed on misinformation on Facebook. It would be useful to explore the same phenomenon on other platforms like Twitter and Whatsapp to come up with a pan-platform understanding of credibility markers.

Experimenting with credit/reputation systems

The primary insight from the pilot as well as the final study was that source plays a huge role determining how people perceive content on social media. The source refers to both the news agency as well as the person who shared the information. Based on this, it would be useful to investigate the effect of a credit system that penalizes fake news dissemination. If people were assigned low credibility scores for sharing misinformation, it is unlikely that their peers would believe the content they share in the future. This would also encourage more cautious sharing of information on these platforms.

References

1. Ebola Lessons: How Social Media Gets Infected. http://www.informationweek.com/software/social/-ebola-lessons-how-social-media-gets-infected/a/d-id/ 1307061, March 2014
2. Eric W. Dolan. PsyPost, “People are strongly influenced by gossip even when it is explicitly untrustworthy”
3. Cheshire, C. (2011) Online Trust, Trustworthiness, or Assurance?
4. Fraunhofer-Gesellschaft. 2019. Software that can automatically detect fake news. MI.
5. https://therealtimereport.com/2011/04/25/facebook-posts-receive-50-of-likes-within-1-hour-20-minutes/

6. https://www.allsides.com/media-bias/media-bias-ratings

7. https://www.postplanner.com/facebook-data-on-fan-page-performance