information

MDPI

Article

# Addressing Misinformation in Online Social Networks: Diverse Platforms and the Potential of Multiagent Trust Modeling

**Robin Cohen [1], Karyn Moffatt [2], Amira Ghenai [2], Andy Yang [1,*], Margaret Corwin [1], Gary Lin [1], Raymond Zhao [1], Yipeng Ji [1], Alexandre Parmentier [1], Jason P'ng [1], Wil Tan [1] and Lachlan Gray [1]**

[1] Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada; rcohen@uwaterloo.ca (R.C.); mcorwin@uwaterloo.ca (M.C.); g24lin@uwaterloo.ca (G.L.); y395zhao@uwaterloo.ca (R.Z.); y43ji@uwaterloo.ca (Y.J.); aparment@uwaterloo.ca (A.P.); jsjpng@uwaterloo.ca (J.P.); wil.tan@uwaterloo.ca (W.T.); lfegray@uwaterloo.ca (L.G.)

[2] School of Information Studies, McGill University, Montreal, QC H3A 1X1, Canada; karyn.moffatt@mcgill.ca (K.M.); amira.ghenai@mail.mcgill.ca (A.G.)

* Correspondence: andy.yang@uwaterloo.ca

Received: 2 October 2020; Accepted: 12 November 2020; Published: 23 November 2020

check for updates

**Abstract:** In this paper, we explore how various social networking platforms currently support the spread of misinformation. We then examine the potential of a few specific multiagent trust modeling algorithms from artificial intelligence, towards detecting that misinformation. Our investigation reveals that specific requirements of each environment may require distinct solutions for the processing. This then leads to a higher-level proposal for the actions to be taken in order to judge trustworthiness. Our final reflection concerns what information should be provided to users, once there are suspected misleading posts. Our aim is to enlighten both the organizations that host social networking and the users of those platforms, and to promote steps forward for more pro-social behaviour in these environments. As a look to the future and the growing need to address this vital topic, we reflect as well on two related topics of possible interest: the case of older adult users and the potential to track misinformation through dedicated data science studies, of particular use for healthcare.

**Keywords:** social networks; information in society; misinformation; multiagent trust modeling

## Highlights

- broad exploration of current misinformation in diverse social networks
- novel proposals for extending multiagent trust modeling to detect fake news
- support of personalization with critical case of older adults
- attuned to future requirements of graph-based techniques and healthcare

## 1. Introduction

Online social networks are prevalent today and are populated with messages originating from many sources (including ones that are not even created by a person but are instead generated by a bot [1]). In order to address digital misinformation in this environment, it is important to first of all properly study how posts are elicited and provided; it is also very useful to take the pulse of these networks to discover real examples of misleading content (learning the unique challenges that exist in detecting and dispelling these messages for users). The first main goal of this paper is to present results of a comprehensive study of several very popular social networking environments, demonstrating that misinformation indeed exists. The second aim is to reflect on whether existing AI multiagent trust modeling algorithms can be effectively applied in order to distinguish valued content from more suspect posts which may be filtered from a user's stream. Towards this end, we examine some specific models that we have developed ourselves (a Bayesian predictive algorithm of trustworthy content [2], a data-driven classifier of low reputation messages that draws from analysis of discussions that follow each post [3], a multi-faceted personalized data analysis to weight the factors predicting trust relations between peers [4], and a process for identifying rumour spread within social networks that draws from crowdsourced labels of content, using expert labelling (originally designed for healthcare applications) [5]). The conclusion that we reach is that a combination of these methods are best to integrate into an effective procedure for addressing misinformation, and that while there are unique challenges of specific networking environments, there are in fact common concerns that can be leveraged for a concrete first step of the analysis.

It is important to note that while many researchers have indeed leveraged machine learning in order to label what constitutes a case of misinformation from existing datasets, one of the beauties of the Bayesian approach is that it reasons from first principles, progressively adjusting its beliefs about whether a post may be suspect; as such, this method is not tied to specific existing datasets and their proclivities. Typical validation of the effectiveness of these methods is done through simulated content with varying parameters (an approach endemic to the multiagent trust modeling subfield [6–9]). We return to reflect on the value of combining both data-driven and decision-theoretic reasoning into solutions for handling misinformation, when presenting our proposed algorithms for misinformation handling, and provide as well a final commentary in the ending section, Section 8. Tracking rumour spread in social media is also a topic of current research interest, with varied approaches being developed. We return to survey some of this related work as well in Section 7, as we look towards future directions.

One of the primary conclusions that we reach is that using multiagent trust methods to address misinformation is a promising way forward, one that can be designed in specific ways so that each social network benefits from its introduction. This application of trust modeling to social networking environments is a relatively novel context where these research methods can be leveraged. (Standard usage is for the challenge of enabling joint partnerships for decision making [10]. Only recently have researchers begun to explore new issues for trust modeling which arise when considered for social networks (e.g., Sen's position that dynamically changing beliefs of users must be examined [11], Falcone's observation that differing information sources require distinct trust models [12]). One distinguishing feature of our work is to suggest that a combination of decision-theoretic and data-driven methods together may hold the key for essential steps forward, due to the nature of the misinformation which arises online today.

As we discuss online misinformation and the potential of multiagent trust modeling, we also reveal two important concerns for current approaches to address, both of which we feel can be supported well within the frameworks that we have developed to date: consideration for older adult users and important attention to misleading posts about healthcare. Our concluding remarks provide a plea for even greater attention on more fragile user groups and on content such as health information which has the potential to produce dramatic end results, if misleading advice is left unchecked.

In an Impact subsection of Conclusions, we also reflect further on the contribution of our work for other researchers, identifying more clearly who will benefit the most and revealing possible steps forward, as future research. We draw out as well the value of assembling a compendium of sample posts, both as inspiration to the design of solutions and as well to reinforce the need for comprehensive approaches to address this social concern. We view our work as a specific study of current social networks, together with commentary on how certain AI methods can best be leveraged in order to assist in identifying misinformation, as a way forward in improving our online existence.

## 2. Methods: Assembling Social Network Misinformation

Our methodology for collecting samples from existing social networks was to assign to each of five different coauthors the task of learning how social networks accept posts from users to be added, unchecked, and then presenting examples discovered when browsing for instances of misinformation. This is a direct observation strategy for data gathering. (This approach has been used in artificial intelligence subfields such as computer vision: introducing examples which appear to be interesting, leading to qualitative results [13].) These examples were uncovered by our team of coauthors during the period of January 2020 to March 2020. (Quite interestingly this was as well a period of time when the coronavirus emerged as a topic of concern. Healthcare misinformation is, in general, a topic where some have argued that detecting the posts is especially important, due to the consequences that users may suffer if they misinterpret the advice [14].)

We are hypothesizing that each of several very popular social media environments today contains posts with misleading information. This hypothesis is confirmed through the presentation of a compendium of examples in Appendix A, which demonstrates that regardless of the protocols used for admitting posts from users, each environment we studied has at least some instances of questionable content (i.e., none of the primary networks today is immune to this important social problem).

Locating these instances then supports our claim that any effort to combat online misinformation must consider a wide range of possible social media. Three of our coauthors explored one of Facebook, Twitter and Reddit while one examined Instagram and Snapchat and the fifth focused on WeChat and Weibo (which supports the view that concerns with misinformation exist even in social media predominantly used in countries where English is not the main language and is thus widespread).

The goals of our research were outlined in the Introduction. In Conclusions, we return to reflect on the value of the methods we have used to achieve these goals and the community of researchers who will benefit from this work. As we begin to present our proposed algorithms for handling misinformation, we present as well in appendices some of the key methods inherent in these approaches, providing detail on the key starting points of Sardana's decision-theoretic Markov Decision process for predicting trustworthiness (Appendix C) and of Ghenai's data-driven methods for detecting rumor and the origins of its spread within social networks (Appendix E). This is prefaced by Appendix B which details some of the specific processing proposed for Twitter and for Reddit; this is also supported by Appendix D which illustrates the Reddit algorithm in more detail through application to a sample exchange within that network. Section 6 delves further into the methods for achieving our companion goals of accommodating older adults and attuning to graph-based techniques, while Section 7 includes deeper discussion of the context of healthcare misinformation and the value of reasoning about the networking properties of each social media environment.

We have presented in the Introduction key relevant literature regarding multiagent trust modeling. By the end of the paper, the reader will be able to appreciate the differences that our work offers. In later sections of the paper, we also draw out how our work contrasts with those of researchers devoted entirely to trying to mine data for patterns, in order to take steps forward. This subtopic is discussed at length in Section 7.2 when we shed more light on rumor spread and networking components which must be part of the considerations. The final recommendations provided for other researchers are raised in Section 8.1. It here that we advocate for adopting the kind of methods that we pursued to gain insights into the challenges of misinformation which need to addressed, leveraging examples to inspire algorithms for judging credibility of content, in order to create novel solutions for social networks, going forward.

The beginning of this paper includes Highlights which list the central contributions of this work. As mentioned above, the Introduction covers the primary goals, towards these key steps forward. At the end of the paper we also return to summarize the key research questions surrounding these goals, and the ways in which they are answered with our work (Section 8.2), to further enlighten the next generation of researchers.

## 3. Results: A Study of Existing Social Networks: Primary Observations

In this section we provide findings from a broad exploration of existing social networks, aimed at locating cogent current examples of digital misinformation. (As each network was explored by a different researcher on our team, we acknowledge that some of our observations may reflect subjective preferences of what constitutes interesting behaviour.)

### *3.1. Twitter*

#### 3.1.1. Interface

Twitter is a social networking platform that enables users to view the tweets of those they are following (sometimes with comments), the tweets that followers retweeted, liked, or replied to, as well as some trending tweets within a user's specific location and some recommended people to follow (based on who one or more followers choose to follow). Twitter categorizes tweets into topics to make them easier to search. When users post a tweet, they can use privacy settings in order to restrict the viewing of these tweets only to their followers. Accounts of certain Twitter users can be blocked from view, though the other person will know they are being blocked. (Another alternative is muting, where a user's tweets won't show up in the timeline and the user won't be notified of their new activity. However, their responses to other users' tweets could still show up in the timeline.) There are two possible options a user can set to try to reduce the flow of tweets to them. Turning on a quality filter will instruct Twitter to block what they perceive to be botspam or duplicated tweets; tweets the user has recently followed, however, will not be removed. Personalization allows Twitter to infer the user's interests in order to inform advertisers. Twitter supports the inclusion of text and pictures both.

#### 3.1.2. Misinformation

We inhabited Twitter in January 2020 to discover instances of Tweets which represented misinformation. (For our initial study of the social networks, we selected content that we considered suspect of being misleading, using subjective judgment. As will be seen with the explanations we provide, the platform owners themselves at times returned to label the post as false, to then remove.) We discovered that some tweets have many keywords that cause the content to become less comprehensible. While these word salads are likely not malicious, they may still count as misinformation, in the sense that they generate confusion for the user. We display some examples of these postings on the topic of 5G, as Figures A1 and A2. (Appendix A displays a compendium of examples of misinformation in online social networks.) Another case of misinformation concerns suspicious display names. This is true for example in Figure A3. Irregular punctuation may be used by adversaries to cause impulsive decisions from the reader, as another case of distraction. A popular strategy is to capitalize words in order to draw attention, for instance. Perhaps the most significant scenario for misinformation concerns misleading hashtags. This is demonstrated in Figure A4, which mentions Huawei.

Coronavirus

The topic of the novel coronavirus has arisen as one where misinformation may pervade. Examining Twitter feeds on this topic, we discovered a few examples as shown in Figure A5.

In the first case, the claimed text cannot be found in the article that is referenced. Many replies arose which pointed out this flaw. Mean-spirited responses eventually pervaded. For the second case, readers were advised of a possible conspiracy theory. Part of the problem is that journalists may look to Twitter to see what is worth covering [15]; this article points out that the fact that users can create accounts without much checking enables groups of trolls or bots to swamp the conversation. An article from the Daily Mail also points out the shortcomings of the current quality filter supported by Twitter [16]. The fact that this solution does not scale well when there are overnight bot attacks is also mentioned. A final mention of using Twitter feed misinformation to divert attention arises in [17]:

the article goes on to reveal how fake news promoted money being sent to a a charity, which suggested that eating exotic animals in China was the primary cause of concern.

### *3.2. Reddit*

#### 3.2.1. Interface

Reddit is an online forum where anonymous users can submit posts containing texts, links and media. Other users may rate posts by upvoting or downvoting them, and comment on posts to continue discussion. Subreddits exist for discussions on a specific subject matter. The main page of a subreddit displays a listing of posts with just their basic information such as title, number of comments and rating. The list can be sorted by various criteria such as new or top. Posts across various subreddits can also be viewed on a single page. Each post can be viewed in more detail. The comment section is organized as a forest. Reddit users also have options to create personalized feeds, save posts or customize settings. Registered users have a score tied to their account, called karma, based on the ratings of their posts and comments. Figures A6 and A7 show the main page of the r/science subreddit and a sample individual page for a post in Reddit, respectively.

#### 3.2.2. Misinformation

While searching for examples of misinformation on Reddit, the following distinguishing criteria seemed to be most relevant. Two preliminary observations are that satire may at times be mistaken for misinformation and certain subreddits tend to be more satirical; and that moderators in certain subreddits inject additional constraints which may assist in limiting misinformation (for example, the r/worldnews and r/science subreddits require a link to any source of information). Moderators in these contexts also have the option of removing suspicious content.

One approach for locating misinformation in Reddit is to search for responses which claim that the poster is lying. We were able to check out these cases by subsequently checking the link provided to see whether the source in question actually supported the claim or not. We also found that it was easier to find relevant content by sorting lists by controversial or top, showing posts receiving many downvotes or many upvotes.

Observations that may alert someone to a post being misinformation, which we discovered during our manual search for such occurrences, includes: (a) source material: (i) whether there is supporting evidence or not (ii) whether the authors' intent is to be self-gratifying, to misguide or to be malicious (iii) whether the purpose is intended to be personal or informative (b) community: (i) whether there is moderator action or not (ii) whether the community response is dramatic or challenged (c) user: (i) whether there is a good user score (ii) whether a user has repeat occurrences

Specific examples located on r/worldnews, r/science, r/AskReddit, r/teenagers and r/OldSchool included ones where: a user claimed a system he created had functionality which it actually did not have; a user stole and reposted a story made by another user, claiming it as their own; a user claimed a picture was taken in the 1970s but it was clearly modern; an article that was referred to that came from a questionable source. Figure A8 displays some sample pages with misinformation. Using our characterization categories of content purpose, author intention, whether user provided supporting evidence, what the community response was yields the following analysis: Figure A8a: personal, self-gratifying, no, dramatic effect; Figure A8b: personal, self-gratifying, no, dramatic effect; Figure A8c: informative, self-gratifying, no, challenged; Figure A8d: informative, misguided, yes, challenged; Figure A8e: informative, misguided, yes, challenged.

## *3.3. Facebook*

### 3.3.1. Interface

Users in Facebook set up a user profile. For viewing, users can examine the Facebook News Feed, a conglomeration of content one's Facebook friends have interacted with, including posts made by Groups one is a part of or on Pages one chooses to follow. Figures A9 and A10 present an example of Page and Group.

### 3.3.2. Misinformation

It is important to note that there is no real checking of profile information created, so users can create fake identities. (Facebook does have a system in place which may reduce fake accounts, one that looks for suspicious activity such as mass account creation from a single IP address [18]. Facebook also sometimes verifies the names of people by requesting they submit photos of their IDs for review [19].) Some posts contain misinformation that is not challenged; this includes posts that receive a lot of shares or a lot of discussion as well as posts that do not receive much attention. Some posts contain misinformation that is challenged: some are challenged almost immediately, others provoke a lot of discussion and then some continue to have relatively little attention.

In searching manually for instances of misinformation we discovered the following kinds of posts: a misattributed quote, with several comments, none of which pointed out the falseness and a deliberately inflammatory post that offers no evidence for its claim, which generated a lot of comments all of which supported the post. Figures A11 and A12 display examples of these kinds of posts respectively.

There were also posts that had suspect content, but did not receive a lot of responses, coming from a Page. This is displayed in Figure A13. In addition, one Page source had obvious misinformation that was treated as a joke. This is presented in Figure A14.

Some observations are that misinformation was often spread due to a user having a particular agenda. Within groups, people seemed dedicated to their own ideology rather than calling out falsehoods. This is shown in Figure A15.

#### Healthcare Examples

Two primary topics for misinformation on health have crept into Facebook recently. The first is the discussion of vaccines. The second is the concern over the novel coronavirus.

With respect to vaccines, it is interesting to note that the search term "anti vaccine" gives no suggested results whereas "vaccine" does, which suggests an effort to discourage the location of anti-vaccine rhetoric. Groups that discuss vaccines get flagged with a blurb that tries to point the users to important credible sources. Figure A16 shows some examples on this topic. We also noted that other kinds of hoaxes were more freely shown, for instance with respect to climate change. Figure A17 illustrates this. This may suggest a particular dedicated effort at Facebook with respect to the vaccines topic. It could therefore be the case that Facebook is being required to address certain kinds of suspected misinformation, due to pressure from various governments.

Facebook also appears to be doing some fact checking, with efforts to present additional information to users when viewing some questionable sources. An example of additional information being shown is in Figure A18. Note that on this critical topic of coronavirus, it is possible to see somewhat more oversight. Users can click on notices of suspected False Information and then be pointed to additional sources, to examine.

## *3.4. Snapchat and Instagram*

We briefly explored two social media environments that are more attuned to images and photographs: Snapchat and Instagram.

### 3.4.1. Interfaces

Snapchat is an instant messaging app that is only for mobile devices. There is an official app that can be downloaded for iOS and Android. In the app users can sign up and choose a unique user name; once signed up, they can add friends by searching for usernames or scanning a QR code in their friends' apps. The primary method of communication between users is by sending "snaps", which are messages that disappear immediately after being viewed. There are three types of snaps: photo, video and text. The primary type of snap is the photo snap, with often includes an amusing caption. Videos can have captions as well. Text snaps can be pinned to a friend's chat history by either the sender or the receiver. Users can also add snaps to their "story", a sequential collection of a user's snaps that can be viewed repeatedly in a 24 h period. They are viewable from a user's friend list in the "Chat" page as well as at the top of the "Discover" page. See Figure A19 for an example of a "Discover" page. Organizations (e.g news outlets) and celebrities can also post broadly using the Discover page. Users can arrange for their favourite stories to be at the top of the page or can scroll through the Discover page to find new discussion.

Instagram is a social network predominantly focusing on sharing photos and videos with other followers and following other users. The platform can be accessed through a desktop website and mobile apps. Users can sign up with their Facebook account or with their email, phone number. Users can upload one or more photos and/or videos to a post with a comment, which may include content tags called hashtags. Other users who see the post can then like, comment on, or share the post. (Currently, users can share posts by sending them to other users in a direct message or adding it to their own Instagram story.) Each user has a page where all of their past posts can be found in reverse chronological order. Instagram's interface is depicted in Figure A20. A user's page also displays summary statistics about their account as well as a list of the user's followers and whom they follow. Users can also create stories, similar to Snapchat. These stories can be viewed by others when clicking on the user's profile picture. Users can also create longer-lasting stories based on a theme that will appear at the top of their page. In the home page of "Instagram feed", users can see posts from people and hashtags they follow as well as suggested posts.

### 3.4.2. Misinformation

For our manual inspection of Snapchat, we felt that almost all user content was created for entertainment purposes; the ephemeral nature of snaps and stories made it difficult for any potential misinformation to persist. We believe there is some control over who can possibly masquerade as a news source, so that the news stories posted at least were affiliated with a specific well known organization (e.g., The Washington Post).

On Instagram, we discovered misinformation both in posts and in comments. One example was a posting by an anti-vaxxer group called @pharmthesheep. Another example was posted by a fan account @mokka_commentary claiming that a new Hawkeye show was cancelled. This was strongly refuted by Disney. These 2 examples are displayed as Figures A21 and A22.

From these cases, we suggest that a classification system for misinformation may be useful, in order to direct any approaches for coping with the suspected content. For example, we could examine

- intent (1–5) where 1 is unintended and 5 is deliberate
- persistence (1–5) where 1 is an isolated incident and 5 is frequently
- acceptance (1–5) where 1 is universally rejected and 5 is widely accepted
- popularity (1–5) where 1 is very few likes and 5 is a large number of likes (using a log scale where 1 means zero or more likes, 2 means 100 or more likes, 3 means 10,000 or more likes, 4 means 1,000,000 or more likes, and 5 means 100,000,000 or more likes)

Our anti-vaxxer example could be labelled as: intent 5, persistence 5, acceptance 5, and popularity 1 whereas our Hawkeye example might have intent 4, persistence 2, acceptance 4, and popularity 2.

### *3.5. WeChat*

#### 3.5.1. Interface

WeChat is the most popular message app in China with more than one billion active users every month, just behind Facebook's WhatsApp and Messenger. It started out as a messaging app but transformed into a platform with multiple functionalities including games and online purchases. The primary use of WeChat is messaging; chat groups with friends or like-minded people can be created. You can send links, articles and videos. Two unique formats for WeChat conversations are mini-programs and official-account articles, both of which can be sent as links in chats. A very closed social circle is also supported by a feature in WeChat called Moments: this content can be liked but not reposted. One can browse through multiple posts in this feed.

#### 3.5.2. Misinformation

The danger of mini-programs is the fact that they do not need to be a registered or a published application. At times when users are invited to participate by these programs, they freely provide personal information. The content here is provided entirely by the creator, so that misinformation can potentially be supported. Official-accounts are similar to Pages on Facebook, gathering followers, sending notifications and posting articles. When these articles are reposted as links in chats and on Moments, users can then decide whether to like, comment, interact with commenters or even donate to the creators.

While Moments tends to reach a limited number of people, official-account articles can be liked and reposted by millions of people in chats. A viral spread in a short space of time is quite possible. With speedy delivery, the ability to fact check before wide reach has been attained is a concern.

##### Health-Related Rumors on Coronavirus

We examined WeChat feeds and posts surrounding the coronavirus outbreak at its outset and observed the following. The kinds of information being distributed included: information on the virus and the spread of the virus, government policy, virus prevention and protection, possible medicines and cures. Older adults in China in particular may be particularly susceptible to false claims in this social network as WeChat is one of the only sources of information that they receive.

Official-accounts may be owned by media companies, aiming to receive a high number of likes and reposts. As an example, one specific Chinese medicine was touted as preventing coronavirus, to be used as a possible treatment for it. The proposed course of action for patients was basically to consume an easily purchased over-the-counter drug. No scientific backing for the claim was provided. The rumour was that a lab had proven its effectiveness, through testing. Within hours of this rumour coming out and spreading, this medicine was sold out at every major drug store in China, with people lining up for hours to purchase it. Manufacturers had to triple their production to keep up with demand. Only later, did scientists debunk this recommendation as ill-advised. All of this happened simply with an initial post with false claims about scientific results.

With Facebook posts, people may try to verify the validity of content because it comes from unknown users. With WeChat Moments, everything users see is posted from someone they know. So they may skip the steps of verifying the course and choose to repost their comments. This snowball effect may be quite troubling for the case of health-related misinformation.

### *3.6. Weibo*

Sina Weibo is an online news, social networking and microblogging platform that has over 300 million monthly active users as of 2017 [20]. Users can follow hot topics as well as popular hashtags, and can share their own life experiences. One special feature is a fan headline (most vital news of the past 24 h), a fan tunnel (exchanging posts on a specific topic) and red pockets (enabling exchange of money). Users outside of China can examine Weibo in order to learn what is trending within China. Sina Weibo profits mainly from in-app advertisements. Businesses compete to be showcased on headlines or tunnels, as well as on banner ads.

Misinformation

The need to attract customers through paid advertising ends up making it possible as well to have individual accounts put into headlines, with fake news and rumours being spread. If Weibo were to monitor the ads and the practice of receiving special status within the system, it may be possible to dissuade the posting and spread of misinformation. This particular social network is less popular with older adults, so that there is a different demographic to consider when trying to address false content being posted.

## 4. Older Adult Users

In this section, we comment more specifically about the user base of older adults, discussing why there may need to be greater attention on this class of users, when trying to detect and address online misinformation.

One lens of current importance when it comes to digital misinformation is understanding the needs and preferences of older adult users. Towards this end, we first investigated current literature on some of the specific vulnerabilities of this demographic. We also brought in some of our ongoing user studies with this user base, conducted specifically for the context of Facebook. It is very important to recognize that there is no universal solution and that individual users may have considerable differences. Some interesting findings have emerged from our studies, however, which we summarize briefly in this section.

The first point is that studies have shown that older adult users may be disproportionately responsible for spreading fake news (among users who distribute due to being ill-advised, rather than with malicious intent) [21,22]. On the other hand, older adults may be more prone to receiving misinformation as well [23]. Research also suggests that older adults may be using Facebook groups in order to learn more about managing their health [24]; if this is the case, it is all the more important to provide strategies for informing users about potential misinformation. The fact that older adults have been particularly vulnerable to online financial scams [25], suggests a certain level of fragility here, to address.

There are positive directions which can be leveraged when investigating misleading messages in online social networks, for this demographic. Of interest is research showing the strong attraction to connecting with family, often facilitated by grandchildren [26,27]. One suggestion therefore is to enable streams intended for older adults to be viewed and filtered by trusted, reliable family members.

When we present proposed AI solutions to intelligently predict and reason about possible misinformation in Section 5, we also step back to indicate how the algorithms can be tweaked in the presence of older adult users. This serves to illustrate how the solutions we are devising not only support personalization but also have the potential to assist a user base that is of increasing importance today.

We have been conducting user studies as well, with the goal of improving accessibility of social media for older adults [28]. For instance, the Facebook interface could be redesigned in order to offer better support for these users, including a rearrangement of layout and added help features. Preferences of users with respect to concern for privacy, intended use of the social network and interface presentation, were all part of our study with participants; personality differences and background were gauged through additional survey questions. All of the observations below should be considered preliminary. Our findings to date suggest that people who perceived their health to be moderately good may be more likely to prefer to benefit from the social bonding features that Facebook provides, compared with those in excellent or good health. Older adults with higher sociability personalities [29] also appeared to have more interest in these social bonding features. This research at least highlights certain conditions under which concern for the spreading of misinformation may be heightened.

## 5. Results: Artificial Intelligence Trust Modeling to Detect Misinformation

In this section, we explore some solutions for reasoning intelligently about how to identify misinformation in online social networks. (We use Appendix B to display some of the more detailed elements of our proposals.) We first developed solutions independently for Twitter, Reddit and Facebook. We then step back to suggest a comprehensive approach, indicating where platform-specific elements may need to be modeled and integrated.

There are four primary starting points that inspired the solutions developed for these social networks. The first is a Bayesian approach to reasoning about whether messages in social networks should be recommended to users or rejected, based on expected utility of those messages to users. Anchored by a model of partially observable Markov decision processes (POMDPs) and intended for environments where ratings of peers are known, the framework progressively reasons about three primary factors: similarity of the user to the rater (based on commonly rated items, in the past), credibility of the rater and the actual ratings that are provided [2]. We refer to this as the Sardana model (displayed in Appendix C). The second model was developed in order to analyze reactions to posts in discussion-oriented networks [3] as a bellweather of the reputation of the user who initiates the discussion. The third model suggests that examining a collection of trust indicators from data analysis is the ideal approach to predicting trustworthiness, in a model where personalized solutions for clusters of users can also be supported [4]. The final model examines rumour spread in social networks such as Twitter, using crowdsourcing to help to identify false information for applications such as health discussion boards [5,30,31].

### 5.1. Twitter

The first social network we explored in more detail was Twitter. This was the network examined in the work of Ghenai [5,30,31] which employed crowdsourcing to assist in obtaining labelled data (used when training learning algorithms to detect misinformation). That study was restricted to the healthcare domain, where expert opinion was also available. Ghenai's work also tracked rumour spread, identifying rumours on the basis of certain features. To slow the spread of new rumours on Twitter, queries are constructed with the help of general-domain and medical-domain search engines, searching for tweets on fake cancer cures. Filtering then takes place using crowdsourcing to reduce the set to those relating to the problem, clustering into subtopics. Users who are rumour spreaders are determined on the basis of differences between their behaviour and those who are not providing false information. Interesting elements of the analysis included the use of geolocation (with some rumours confined to restricted areas) [5,31]. We observe that more natural language processing training might assist in automating the query generation and crowdsourcing steps.

Determining whether a tweet contains misinformation is a classification problem, so the final output should be a binary (yes or no) or a probability that describes how likely a tweet contains misinformation (on a scale from 0 to 1). We propose an algorithm that contains the strengths of previous models, by conditionally using those models depending on the tweet given, and then using parameters estimated by those models as observations for Sardana's POMDP model [2]. Our algorithm first selects a subset of users not including the current user. It then models the users, to establish which ones are more likely to tweet or retweet misinformation, using Ghenai's approach [30]. This metric then gets saved as the user's credibility. Taking this step assists considerably in acquiring and representing the credibility factor included in Sardana's work.

From here, we employ a metric used in Parmentier's multi-faceted trust modeling work [4], the Pearson Correlation Coefficient (PCC), to generate the similarity between the current user and that subset of users. Parmentier had noted that "friendship links between users is only slightly correlated with a similar reviewing behaviour" so that to make predictions of what will be trusted by a user, PCC improves. This therefore advances from Sardana's original vision of similarity as a certain distance from previous rating habits. Our metric makes use of these modifications: (1) the review given by a user on a tweet is 1 if the user likes the tweet and 0.5 if the user retweets the tweet (2) the review given by a user on a tweet is $-1$ if the user responds to the tweet and the sentiment calculated falls below the neutral threshold.

Parmentier considered Natural Language Processing (NLP)-related features as well as important non-NLP-related ones, such as a disagreement index, when examining discussions within social networks. This latter factor is a measure of how much a single branch of responses alternates between negative and positive ratings (arguments along a branch) [3]. In a Twitter-based environment, one would be focused on a tweet with a certain number of likes and the behaviour that then ensues. Our stand-in for a Tweet's rating in our algorithm would be motivated by Parmentier's model for tweets that have enough responses, assuming that a tweet reaches that level when there is a tree of depth two with at least two leaves. If a minimum level of attention on a tweet has not transpired, the rating would be derived using Ghenai's model instead.

Basically, we have been able to obtain a stand-in now for the three parameters modeled by Sardana: rating, credibility and similarity. The core algorithm is presented in Figure A24.

However, the algorithm only flags tweets given a set of users. If interested in warning of possible misinformation, the following steps would be taken. The first $T$ tweets can be examined for an estimate of the number with misinformation, and then a warning generated. The algorithm to carry out this is as presented in Figure A25.

One final consideration for future extensions is the personalization advocated in [4]. That work demonstrates that it is more valuable to analyze on the basis of clusters of users, when processing data from social networks to predict trust links between peers than to assume that all users can be treated as members of a homogeneous set. The selection of tweets to show can be coloured by differing profiles and preferences. Caution should be exercised, however, as showing tweets in this kind of restricted manner may create echo chambers that will end up reinforcing misinformation. Some clever algorithmic decisions about what to display (e.g., rearranging the order) may end up proving helpful to address the echo chamber behaviour. The ability to recommend content to dedicated groups of users is a topic we return to discuss in Section 6.3, when we draw out how the solutions we are proposing here may assist with the user base of older adults.

### *5.2. Reddit*

The second environment we examined in more detail was that of Reddit. This was the social network forming the basis for Parmentier's work on establishing user reputation on the basis of reactions to a user's posts, through a hierarchy of discussion [3].

A first observation is that credibility of authors is more difficult to track in this context, with multiple accounts being possible and anonymity of users often a norm. There are still some possible avenues for modeling a user's behaviour over time. Each user on Reddit has a karma score, not quite identical to the net vote (as some votes count less, not dissimilar from the idea of weighting less heavily the reports from less valued advisors, within peer-based trust modeling approaches such as that of [32]). Karma still reflects where the user's message has received postitve reactions, so it could be used to estimate credibility. In addition, some history of a particular user may still be viewable publically, and this may also provide some important insights into likelihood of spreading misinformation.

As for reasoning about ratings in Reddit, it may be difficult for a user to know the true number of upvotes/downvotes due to a habit of "fuzzing" (to prevent spammers from learning their popularity). The displayed net score is still a good representation of how a post is received by the community, even if there is some uncertainty about how it is actually calculated.

In examining Reddit, we decided that several factors should influence the algorithm that reasons about potential misinformation, beyond the core considerations of ratings, credibility and similarity that are in focus for the model of Sardana [2]. We begin by observing that to determine whether some message is misinformation or not, the direct way would be to verify whether the message is true or not. Although this would be the most accurate, this requires knowledge of ground truth, which is often difficult or even impossible to verify. So, the solution outlined below forgoes the integration of ground truth. Instead, we examine features that are related to the presence of misinformation.

Author credibility is certainly one important consideration. But we observe that it is often features of the message itself which assist in determining this factor. For instance, poor language being used could be sign of lack of credibility, or incorrect terminology could indicate that the author is not well informed on the topic of the message. Claims without supporting links may also be more suspect. Another feature worth examining is the influence on other users. For example, a post could initially receive many positive reactions but after something untruthful is discovered by a user, subsequent reactions could be more negative and could assist in detecting misinformation. So it may be useful to distinguish the kinds of impact that misinformation have within the network. Our proposal is to integrate a consideration of all of these factors into an overall reasoning process. Tables A1 and A2 (in Appendix B) show tables displaying possible values of features we suggest modeling and their descriptions, respectively.

An analysis taking these features into consideration would yield a prediction of the likelihood of misunderstanding. We then propose that two additional factors be integrated in order to decide whether a message should be shown to a user. One interesting consideration is severity of topic (how consequential would it be if a user were misinformed); another is the user's tolerance of misinformation. There may well be important differences between users' reactions to exposure to untruths, and differences as well simply due to the topic in question. A general strategy could be to avoid showing posts with high likelihood of misinformation, high severity, and low user tolerance.

It is also important to consider a few more decisions with respect to reasoning about what to show to users, once misinformation is suspected. Three crucial axes are: when, where and what. "What" is to acknowledge that there are different options for the display of possible misinformation (such as attaching some kind of red flag or requiring some kind of explicit consent before the message is revealed). "Where" allows for the option of maintaining a completely separate feed. "When" suggests that less valuable messages may simply be shown less frequently (so that it is not simply the most popular messages but the most credible ones which end up with elevated status).

Our final reflection is devoted to some of the seminal trust models that are being examined in this paper. We discuss some specific considerations for Reddit which suggest extensions to these models. Part of our perspective is that we should consistently search for new factors to integrate into the reasoning about misinformation; that stance in particular is not dissimilar from the view of [4], advocating for detailed reasoning about a multitude of trust indicators.

While the Sardana model holds considerable promise with its focus on reasoning about similarity, ratings and credibility, there are possible improvements for environments such as Reddit. Instead of simply equating ratings with up and down votes, we note that this captures more a sense of positivity or negativity than it does a notion of truthfulness. Delving into the meaning of comments may be required in order to translate from votes into trustworthiness more effectively. It is also the case the commenting occurs more often than voting, which could lead to a shortage of valuable advisors. Perhaps what will be most useful to consider is evidence of distrust, revealed by reactions to post. This is not unlike the approach of [3], which leverages responses to messages as part of the reasoning about potential misinformation.

The distinction between negativity and distrust also suggests some extensions to the Parmentier framework itself. We also note that the kind of comment trees examined by this model would then be the place to search for evidence of distrust (e.g., the post author makes a poor justification for a claim and then the community registers their reaction to this).

### 5.3. Facebook

A third social network which we explored in some detail was Facebook. We examined some of the unique qualities of this network in proposing a particular algorithm for detecting misinformation. Our solution returns once more the approach of Sardana [2] but now explicitly attempts to integrate a reasoning about the central concern of author credibility, as well as developing a stand-in for post credibility, based on the kinds of group reactions that may be revealed in a network like Facebook. A final concern that we integrate, which seems a prevalent factor in this environment, is the use of links within posts and then the credibility of these links.

We begin by focusing on posts made in Pages/Groups. This puts front and centre the concerns of spreading misinformation (far more than, say, simply making a questionable post on a News Feed). Including a reasoning about links that are posted continues to draw out this theme of paying attention to what might be shared, which exaggerate the undesirable effects.

One of our first challenges is to develop a stand-in for the rating of a post. In Facebook there are a variety of emoticon reactions (Like, Heart, Laugh, Wow, Cry, Angry). We chose to disregard the rather ambiguous Sad, Angry, Laugh and Wow. For instance, individuals might be angry about the author's intentions in the post itself (e.g., the post says something misguided and inflammatory) or they might be angry on the author's behalf (e.g., the post is about the author's experience with racism). Both situations can lead to Angry reactions on the post, but with different reasoning. Heart and Like are considered and treated as positive endorsements of the post.

From the original POMDP model of Sardana we retain much of its central reasoning about whether to recommend a post to a user, but first of all extend the possible actions to not only recommend, reject or poll advisors but also an action referred to as Assess. The assessment phase is intended to determine a credibility label for any given state, judging the credibility of author, of the page or group and of the link source (used if the post was made by sharing a link from another website). These credibility measures are progressively updated as part of the Bayesian reasoning about trustworthiness and message recommendation. The kind of belief update that is central to Sardana's process for deciding whether to recommend a message, then also becomes a stage in the overall reasoning of our model.

While Sardana advocated for a principled reasoning for trust modeling, as an initial placeholder we discuss some heuristics which may be useful for combining the influence of different credibility elements (as we propose within our model). Some of the central calculations which are required, translated into the operations of the Facebook environment are as follows. The number of positive endorsements on a post is derived from Heart and Like reactions, but this should also be focused on those coming from credible users. How much we value a community response can be derived from this. We also advocate allowing for updates of the opinions of each user. For example, if someone posted a lot of misinformation in the past but has become more careful with fact checking, then a gradual change in their credibility assessment should be possible. We suggest, for example, examining the last 100 posts from a user as a bellweather of their behaviour. We also note that credibility scores should technically be continuous but these can be converted to a value of 1 (if 0.6 or higher) or 0 (if 0.4 or lower). For the middle range values, one can simply opt to be cautious and round down to 0 or be more liberal and round up to 1; this could be based for instance on the criticality of the topic or other domain-specific considerations.

The kind of reasoning that results in the end turns out to be similar to that of Sardana, with the additional considerations we have introduced. At a high level: if a post is made by an Author in a Page and sharing a Link, we would be influenced by how many non-credible users Liked or reacted with a Heart to that post and some weight could determine how much the community reaction should influence our overall decision. Once observations and decisions are made, credibility values can be updated for use in future rounds. An author's credibility over time will progress based on the number of valued posts. Note as well that we may need to factor out Likes of the author's own posts, when performing these calculations.

### *5.4. Returning to Sample Cases of Misinformation*

The algorithms presented in this Section are each intended to be of assistance in coping with the kinds of misinformation that we presented in Section 3. In Appendix D, we select a specific example and then illustrate how the processing presented here could change the course of actions within that social networking environment. This serves to bring our discussion of existing cases and solutions for coping with such cases together in unison. Figure A26 displays a specific Reddit post that misinformed, while Figure A27 shows the thread that followed this post on that social media platform. Through a discussion of this particular instance, we are able to provide more detail on how Bayesian multiagent trust modeling can be leveraged when specific cases arise.

### 5.5. Towards a Comprehensive Approach for Detecting Social Network Misinformation

With our dedicated study of several social networking environments: the central arenas of Twitter, Reddit and Facebook, as well as other relevant networks such as Instagram and WeChat, we first of all step back to collate some of our key observations. We then draw out some key suggestions from the algorithms presented in this Section, to suggest where common ground lies and where solutions that are developed will need to have network-specific considerations.

A few common themes have emerged from what began for us as completely independent studies of differing networking environments: (1) The use of links within posts and the power of these links to convince or misinform. (2) The nature of the language used within posts where in some cases totally false claims may merely be an attempt at satire so that language analysis becomes a concern. (3) The impact of any spread of misinformation, which may be moderate or dramatic, based for instance on the number of followers or the reach of the spread. (4) The effort or lack of effort to introduce supporting evidence when making claims. (5) The ability for users to shield their identity or even to masquerade as authorities. (6) The consequences of misinformation where in some cases the concern is rather minor and in other cases the side-effects of unwarranted belief may be truly significant. (7) The tendency of the platform owners to attempt some quality control. We note that we saw observations of these key factors across differing social networks. For instance, the comment that if there are very few likes, the impact may not be so disconcerting was raised in Section 3 both with respect to Instagram and to Facebook. The use of misleading identities came up as a concern both for Twitter and for Facebook. And the worrisome consequences of dramatic reach arose both when discussing Reddit and for the case of WeChat.

Our first suggestion therefore is that any proposed solution for detecting and addressing misinformation in any social network should pause to reason about each of the central concerns outlined above, which can then become distinguishing factors with respect to actions to take, to assist users (e.g., weighing in on links which are revealed, or noticing a lack of effort to support claims in key arguments).

If we examine more closely the specific suggested algorithms, we can begin to tease apart the core reasoning processes and the network-specific elements. Both the approach developed for Twitter and the one for Facebook saw value in using the the Sardana model as a starting point [2]. That framework is ultimately reasoning about whether a message should be recommended to a user or not, based on key concerns regarding credibility, similarity and voting behaviour. For the case of Facebook, the environment enables additional information to be understood, such as whether the information is being pushed to a specific subgroup of the overall network. This then enables a richer evaluation of the rater's credibility, moving more towards an assessment of the reliability of a set of raters. The prevalence of links within posts also suggested a way to initiate analysis of the credibility of a message as well, while the author credibility had a possible starting point the number of endorsements received. The primary extensions here may then be viewed as an expansion of the consideration of credibility as a factor in the overall reasoning.

For the case of Twitter, credibility of the raters continues to be important but the algorithm sees some promise in deriving an initial estimate of this independently through a process similar to the one used by Ghenai [30,31] to label certain users as more suspect. For this particular social network, there is a wealth of labelled data and an ability to crowdsource some experiments. The Twitter solution also pays more attention to the contributing factor of similarity, suggesting a richer interpretation, also available from analysis of the Twitter data.

Two proposals in common to the solution presented for Twitter and that of Reddit are focusing more on analysis of language and content, as is an element of the work of Parmentier [3]. The Reddit proposal specifically leverages a combination of features in its ultimate classification of messages as trustworthy or not. Some of this is motivated by Reddit-specific concerns such as anonymity of authors. The Twitter proposal delves much further as well in modeling the rating behaviour (beyond the simple yes/no responses imagined by Sardana [2]). The entire structure of discussions and depth of reactions (as promoted in [3]) now becomes relevant to leverage as well.

All of this reflection on the models presented in this Section leads to our second primary recommendation for coping with misinformation, in any social networking environment: try to progressively learn and predict whether messages will be worth recommending to users or not and continue to draw out the key elements of credibility, ratings and similarity but then: (i) leverage any information freely available from the network in order to expand the modeling beyond the level of individual raters through to groups of raters (where each network will have different elements which provide this extra information) (ii) delve further into message content through more detailed analysis of content (making use of any publicly available datasets from the social networking environment as well) (iii) pay attention to the content of posts with respect to their use of links and how this introduces a secondary element to scrutinize.

## 6. Materials and Methods: Addressing Detected Misinformation

In this section, we discuss a few directions for turning our proposed solutions into active adjustments within social networks.

### *6.1. Holding Anti-Social Posts to Further Inspection*

We begin by describing a companion topic, addressing hate speech in social networks, and briefly present some work we have done which is of particular value in discussion-oriented environments such as Reddit.

The high level idea is to imagine that the value of a particular post may be determined (at least in part) by the reactions received to that post, and whether the respondents were generally positive or not, and what the sentiment of that community appeared to be. This was the challenge examined by [3]. We aimed to learn which features of the text analysis might be the most useful indicators of the reputability of the original post. We discovered certain correlations, through big data analysis: for example, we learned that the disagreement index of scores in descendants (whether all those who reacted had the same positive or negative vote afterwards) turned out to be a very good predictor of the score of a parent comment. From this initial project, we then decided that looking more carefully at the collection of reactions as a graph, and leveraging the graph properties that would be of use. To achieve this, we implemented a solution using graph attention networks (GAT) [33] as the basis for learning and prediction. (Our preliminary ideas for integrating graph attention networks have recently been accepted for publication [34].) A companion consideration was to carefully curate the large Reddit dataset beforehand in order to focus on the most valuable posts.

The heart of this proposal is to bring posts on social networks such as Reddit under greater scrutiny by examining the natural language and metadata features of the response to those posts, under the assumption that reactions to genuinely upsetting behaviour will provoke a detectable pattern of disruption that merits attention.

The new insight on this particular project was to understand that a certain novel deep learning approach, namely that of graph attention networks, may have particular value in speeding the learning process and in freeing the algorithms that we had developed [3] from simply operating with a collection of features that were selected by hand. We learned that GAT can reach performance about on par with the more heuristic approach of hand crafted features but feel that this direction is still promising:

it suggests that textual analysis of posts, not just sentiment but factors such as agreement among peers, may shed light on user reputation. And this in turn may help to draw out the primary sources of misinformation. The fact is that these particular networks excel in scenarios where relationships among peers that are connected through links (graph-based) can be emphasized and drawn into the reasoning process. The kind of discussion-oriented hierarchy that we had been examining (parents and descendants, with children playing a significant role) were ideally suited for this method of machine learning and prediction. To date, our results of accuracy, precision and recall when identifying posts that promote hateful sentiment are modest but we are continuing to look into ways of integrating more intelligent processing of natural language (e.g., word embeddings [35]).

### *6.2. Options to Provide to Users*

Once misinformation has been detected, there are various options for improving the experiences of users in these online social networks. We first observe that it is difficult to decide whether an authority should debunk the suspected messages or refuse to give these posts any credibility by acknowledging them. We have noticed some middle ground steps taken by certain platforms (such as presenting users with alternate sources of information, as was done with Facebook (Section 3.3.2)). But as we will point out in our examination of healthcare and coronavirus in Section 7, certain authorities who control the flow of information may be themselves suspect.

Per our presentation of approaches to analyze the discussion following a post as a stand-in for judging user reputation (Section 6.1), there may also be an executive decision to make about whether to focus a user's attention on an initial post (ones deemed to be of questionable value) or to in fact draw their attention to the reactions which have followed, instead. Exploring the significance of other reactions, especially in concert with looking at the sharing behaviour of the users may provide important insights. Deeper data analysis of existing streams of responses to posts may thus be of particular assistance. Some platforms such as Facebook tend to combine text and images so that an even richer, complex analysis may be warranted.

As for which actions administrators can take, we have observed the following. We are already aware of efforts within certain platforms to withdraw or label suspicious information. Reddit administrators have an option of "quarantining" entire subreddits, so that users need to explicitly consent before viewing any message on it. Topics such as hoaxes and conspiracy theories could warrant a quarantine (as well as other topics, such as offensive ones). Reddit allows individual users to flag posts which are felt to be misinformation, and then moderators have the power to remove messages which are deemed to be of particular concern. The algorithms suggested in Section 5.2 may also help to inform moderators about posts that merit further scrutiny.

We have also seen cases, for example with Facebook in Section 5.3, where certain posts have low credibility and could thus be treated as candidates for filtering. However, it may also be prudent to give users the option to enable or disable filtering. We have noticed already that administrators may draw attention to certain posts for users and these can continue to be candidates for reporting. Figure A18 presents one such example. Facebook turned to independent fact checking in order to report on this particular post so that it is now marked as inaccurate.

As for Twitter, the platform is used for its perceived "freedom of speech". So rearranging, flagging or deleting posts may all be dangerous to business. Certainly the latter once again raises the spectre of censorship, even if the posts are merely hidden due to not passing Twitter's quality filter. One constructive approach, which Twitter already appears to be using, is to suggest that users become better educated. When users attempted to search tweets under the coronavirus topic, this tactic was used [36]; by 18 March 2020 it was then announced that posts which put users at higher risk of COVID-19 transmission would simply be deleted [37].

The bottom line is that there are indeed options between flagging and educating or edging more into the territory of removal but administrators should be carefully informed about possible suspected posts before proceeding with these options (and our algorithms in Section 5 are intended to yield greater enlightenment on this). We are also suggesting that users be allowed to express some preferences for these major options, as well.

We end with a brief reflection on some of the pitfalls of rearranging or flagging, illustrated for the context of Twitter. In Twitter, if one opts to rearrange tweets, then someone could still try to set a script to load the tweets in the original order and display them on a third-party site, dragging away views (and maybe revenue) from the original site. If tweets are flagged, then cliques may simply be encouraged to spread the message in order to thwart the flagging algorithm, which may result in a contradictory effect of flagged tweets being trusted even more than tweets that are not flagged. Lastly, deleting tweets could just result in claims of censorship and additional backlash. None of this is to suggest that these approaches are not without their merits. Continuing to examine challenges to these options and then developing disincentives for various attacks should really be a thread for future research. As mentioned previously, combining these technical solutions with education appears to be the ideal strategy, as well.

### *6.3. Applying the Lens of Older Adults to Active Approaches*

We return briefly to revisit the issue of assisting older adults with respect to digital misinformation, reflecting on whether the specific proposals in Section 5 have valuable starting points for incorporating concern for this particular demographic. We will focus on Facebook (already acknowledged in Section 4 to be popular with older adults) and Twitter, because of its prevalence and use by many popular figures of the day (e.g., politicians and celebrities). For Facebook, we continue to believe that our focus on Hearts and Likes serves this user base well, because the meaning of the terms is less difficult to decipher (perhaps especially for those who were not raised on emoticons). We also feel that the design decision of reasoning about Groups and Pages is also important. In Section 4, we discussed how older adults have a strong investment in family, so that Groups certainly becomes relevant. But we also mentioned that it is important to make this group aware about fake news and the dangers of its spread. Pages may come into view (even if proposed by trusted Groups) where the content is of questionable quality and merits review. For Twitter, we acknowledged the importance of addressing the viral spread of healthcare rumours; the relevance of healthcare to this demographic was also explained in Section 4. Our suggested approach for reasoning about misinformation in Twitter supported some personalization and it would be especially valuable to imagine some specific preferences for the older adult base as part of the individual differences to consider here.

Ultimately, the best steps forward in order to provide the best solutions for a user base with specific needs is to conduct user studies to determine where their preferences lie, as we have already begun to explore [28]. Since the solutions proposed in Section 5 leverage artificial intelligence modeling integrated into each algorithm that detects and addresses misleading messages, we are well positioned to support solutions for coping with misinformation that can be tuned by parameters, such as those reflecting the preferences revealed through user studies. Basically the proposals of Section 5 all support reasoning more specifically about the kind of user who needs to be assisted with respect to misinformation. Key elements at play in our automated reasoning procedures of particular value for supporting more targeted solutions for older adults include similarity with other users and particular emphasis on determining the validity of links.

We acknowledge as well that for a data-driven study of older adults, there will be challenges in amassing specialized data sets for analysis. With the way that the data sets are tagged in Twitter, for instance, the age of users isn’t public information. In Facebook, it may be possible to examine content in the context of ads that target older adult users as one step forward. This is perhaps another reason why a study conducted from the outset with a dedicated older adult user base may prove to be most valuable.

## 7. Discussion

As we look to the future of addressing misinformation in online social networks, it is valuable to look towards what has been happening with the reporting of health information in these discussion boards. Health was the central topic in focus for the work of Ghenai [5,30,31] and interestingly, it has become the core discussion point today as we write this paper, due to the COVID-19 pandemic.

### *7.1. Healthcare Misinformation*

We begin with some observations on the models of Ghenai, to suggest where they could be adjusted or extended for future explorations. We then raise the very interesting case of diagnosis-centric misinformation, outlining some possible strategies to mitigate dramatic consequences from this, illustrated through the case of COVID-19 and with a specific challenge for search engines such as Google.

The first project of Ghenai's, using the case of the spread of information as Zika its centrepiece [5,31], combined an exploration of desired features to model in Twitter and a proposed integration of experts into the process of properly tagging the tweets as rumours (or not). The classifier had good predictive performance, with medical/domain features being of most use. Readability factors had been considered as well but one observation is that malicious users could easily mask this deficiency in their posts, so it should not be the primary consideration. Ghenai advocated examining the kinds of links which were provided as justification; but we feel that even if these sources are legitimate, users can misrepresent those sources, which points to the need for continued analysis of the texts in the tweets. Sources, in the context of medical misinformation are of course crucial. For the future, it would be valuable to carefully check whether these once reputable papers ended up being contradicted later (e.g., the link of the Mumps, Measles, and Rubella (MMR) vaccine to autism, once published in the Lancet).

The second thread of Ghenai's research sought to identify rumourmongerers in Twitter [30,31], an admirable goal in the battle against misinformation. Medical experts continued to be a resource during the labelling process. This research was able to identify a number of features that differentiate rumourmongerers from other users. The rumour environment also appeared to have more followers and more sharing of links than in the more truthful cases. One challenge with this approach is obtaining effective expert opinion, for example in cases where the medical concern is a rare one. We observe as well that in social networks with anonymity, such as Reddit, it will be more difficult to track the rumormongerers. For cases such as Facebook, with more persistent accounts, the approach will be more effective.

Before moving to highlight the specific example of COVID-19, we provide some insights into misinformation for healthcare compared to other topics of interest and some primary areas for future study, with social network misinformation. Some topics like politics may truly lack an objective ground truth, making approaches such as Ghenai's somewhat more at a disadvantage. Even for more factual topics such as health, experts may disagree (for example on whether Creutzfeldt-Jakob Disease can be contracted from ocular tonometry [38,39]). In cases where there is an authority to validate information, the authority themselves may actually be suspect, as was the case with local experts covering up "the facts" in Wuhan at the outset of the COVID-19 pandemic. Censorship may arise to increase the susceptibility of people to misinformation because when messages are removed, it is difficult to know whether this was because it was untrue or whether it was damaging to a government (again, as was the case with COVID-19 concerns in China). We also observe that older adults likely have more health problems to consider, and as such they become a particularly vulnerable group of users for health misinformation concerns. We also feel that looking at reactions to posts rather than simply the posts themselves may really prove to be quite productive to helping users; this approach was advocated by Parmentier [3].

While the kind of health concerns that Ghenai examined did focus on certain diseases and obtaining accurate information about them, one special type of healthcare discussion in online social networks merits some attention. We refer to this as "diagnosis-centric misinformation": cases where a post claims to provide a medical explanation for a combination of symptoms someone is experiencing. We feel that there is a preponderance of searches online precisely because of this particular concern, for a user. Misinformation easily pervades due to the nature of certain social networks to be populated with personal anecdotes [14], with non-experts offering medical advice, with questionable fact-checkers being cited (e.g., WebMD) and with the tendency to draw our sensationalized news stories as part of the conversation (where again this information may be coming from less reputable sources).

What makes diagnosis-centric misinformation so thorny is the fact that medical facts are often quite difficult to verify. On top of this, the average user is not a medical expert and as such is challenged in being able to verify any facts that are presented. As mentioned earlier, even reliable sources may differ in their opinion, so simply checking with a personal trusted authority may fail. The kind of user who is searching for a golden answer online is also likely to be one experiencing anxiety and thus not always reasoning rationally. It becomes all too easy as well to dismiss any contradictory evidence that is presented, if once some online result is found, the user is already biased to simply accept what they were first told.

All of this perhaps points to a greater responsibility on the part of search engine providers, who may begin to bias their users towards exploring certain sources in greater depth. If it were possible to encourage users to check some of the primary fact-based authorities or to question each source that they consulted, this small push towards a more critical interpretation of sources could be helpful. The issue of results from search engines is somewhat off topic for this paper, which has focused on the spread of misinformation among peers in social networks, but part of the problem is that the links revealed when searching may well lead users to discussion forums and then where these are hosted and perhaps especially which links are cited as evidence for arguments being made, all become relevant and are part of the step forward towards combatting misleading messages online. Some promise for the future, coming from the search engine Google, is presented in Figure A23. During the COVID-19 crisis, Google has added sidebars providing information from reliable sources whenever a user searches for something relating to the novel coronavirus.

### *7.2. Rumour Spread in Social Networks*

Tracking rumour spread in social media is a topic of current research interest. The work of Ghenai for addressing this problem has been highlighted in this paper. Appendix E provides more detail on the methods used, with Figure A28 outlining the control flow for the identification of rumor and Figure A29 revealing the data flow for isolating rumour spreaders.

In general, approaches for addressing rumour spread include misinformation detection techniques based on either content features or network features [40]. An an example of existing system that uses content features for rumour detection is TweetCred; it assesses the credibility of tweets in real-time and validates scores by asking for user feedback [41]. Another example is the pipeline proposed by Ghenai using Twitter, sentiment, linguistic, and readability features in order to assess the veracity of tweets concerning Zika virus rumours [5]. For network features, examples of existing tools are: Truthy [42], Rumorlens [43], and TwitterTrails [44] where users can check the propagation of rumors in a dashboard. There are a number of limitations to the current technologies that detect rumour spread in social media. For example, addressing the new social media challenges using a human-in-the-loop approach requires close involvement of users during the debunking phase.

We feel that it would be especially valuable for future directions on rumour spread to delve further into some of the networking properties of the social media environments. One starting direction for future research would be to delve further into the ways in which misinformation tends to spread within networks. One starting point is carefully identifying leading influencers within social networks, in an efficient manner, for these environments where the number of peers who are linked may be massive [45]. Looking at connectivity within the network is generally a valuable exercise. At this point the metaphor of disease in fact becomes relevant. Consider the case of Twitter, (interestingly, the context examined by Ghenai when conducting research on rumour spreading [5,30,31]). People within this Twitterverse have different levels of susceptibility to misinformation, and differing levels of connectivity which provide them with more or less exposure to possible misleading information. Network science and the interesting concept of emergent systems [46,47] suggests that looking for densely connected communities may prove to be an extremely valuable step forward. Fake news propagators may tend to congregate in such networks so that looking "in the right place" may provide the most value in the battle against misinformation.

For instance, some of Ghenai's steps towards flagging rumour spreaders could be examined specifically with respect to the network topology, to have a better sense of the extent of the spread and thus of its concern. A user's proximity to a known rumour spreader could suggest that they are more likely to be spreading rumours themselves, in the future. In a vast arena such as Twitter, troublesome users may be distributed far and wide across the network or could in fact be concentrated in a very limited circle of peers. For this latter scenario, we may then be even more concerned about the magnitude of the sentiment expressed and its effects on that community. Knowing where in a network the misinformation is originating would also assist in crafting solutions to assist users (e.g., to know whether they are particularly fragile or relatively safe, if they maintain their position within the network). We have been exploring the idea of merging a web crawler to understand the networking with the model of Ghenai [5] in order to learn more about how to assist users in social networks.

It is important to acknowledge that the topic area of information spread within social networks has attracted attention from a very large number of researchers (including the prolific group at Indiana University and its recent book [48]). What we are proposing here is examining further the potential value of leveraging ideas from the field of Emergent Systems, promoting the use of web crawlers to provide starting points for exploring concepts from this field of study. Our own current framework for detecting rumour spread [5] would be an especially valuable starting point for such an investigation.

## 8. Conclusions

There are at least four central conclusions from the work described in this paper. The first is that misinformation still abounds in online social networks, including some of the most popular platforms of today. The second is that each social network has specific features which may need to be considered when designing algorithms to detect misinformation using methods from AI multiagent trust modeling (e.g., if there aren't explicit ratings and voting behaviour is an element of the model). The third is that there are specific vulnerable segments of society for whom solutions may need to be finetuned (e.g., the demographic of older adults, presented in this paper). The final point that we are making is that there still are some common lessons learned from our exploration of these networks and our view to the future challenges that we are facing: from these reflections there are steps forward which be considered, towards an improved online existence. The problem of digital misinformation in online social networks is by no means an easy one to solve. Vast numbers of researchers are currently engaged in developing new solutions [49]. But our dedicated cross-platform study, under the lens of AI trust modeling, offers an important contribution to that collection of insights. We also help to explain the possible value of considering multiagent trust modeling algorithms as part of the solution for detecting and addressing misinformation; this decision-theoretic reasoning from first principles has been presented here as an effective first step in determining which posts to recommend to users. We have explained as well how this approach can still be complemented by the kind of data-driven methods which have become the more standard tactic of researchers combatting online disinformation today.

### 8.1. Impact and the Future

The work presented in this paper has impact for a number of different researchers today. We contend that a compendium of specific, clear examples of potential misinformation, drawn from several of the most popular social networking environments today, is of particular value to anyone designing novel methods for detecting misleading social media content. It is by studying the kinds of challenges for users that arise with such examples, against the backdrop of one's proposed directions for modeling misinformation, that we can each truly brainstorm on the best designs for the processing, gaining inspiration. Understanding as well how the ways that each social media environment is set up for users to provide posts and reactions is also central to true insights into the origins of this content that must be dealt with. Appendix A should therefore provide a very important resource. Other researchers going forward should also consider performing a similar kind of hands-on study, during a dedicated time period, to gauge the pulse of today's social networks, in order to move forward with solutions. If specific networking contexts are in focus for these researchers, they can of course restrict their focus to that environment.

The other general approach taken in this paper has been to begin with concrete starting points for analyzing social media content and to examine how best to adjust these so that the particular challenges exhibited from the initial study above, can be accommodated within the automated reasoning processes that embody these solutions. We propose this as well as an important method for advancing scientific study, making researchers aware that it is always valuable to see where there is common ground with those who are studying slightly different specific contexts, in order for us to strive towards solutions that retain the best practices of each particular algorithm that is developed. We have done this with our more general reflection after the study of the networks we chose to examine in more detail. We have also paused to acknowledge that special classes of users may be important to consider (e.g., older adults) and that specific hot topic issues with dire consequences (e.g., healthcare) should be accommodated. We challenge those moving forward in the future to reflect on which users need special consideration and which specific topics need to be handled especially well with the methods that are designed.

### *8.2. Summary of Research Questions in This Work: Lessons Learned*

To end this paper, we articulate the central research questions (RQs) surrounding the goals drawn out at the very beginning of the paper.

RQ1: Does misinformation exist in different popular social networking sites?

RQ2: Can existing multiagent trust modeling algorithms be effectively applied towards the detection of misinformation in social media?

RQ3: Do the algorithms developed in RQ2 support personalized solutions for key user communities such as older adults?

RQ4: Do the algorithms developed in RQ2 provide insights into how to address key concerns such as health misinformation?

RQ1 is answered affirmatively, with the evidence uncovered through our dedicated studies of each of a number of key social media environments; the various elements of Section 3 and Appendix A provide the details. RQ2 is confirmed through the solutions outlined in Section 5. We summarize key elements that should be in common to algorithms designed for any specific social network; we elaborate on ways in which specific social networks may require additional processes in order to detect misinformation (due to the design of the platform). RQ3 is raised in Section 4 and returned to in Section 6.3; how to integrate these considerations with the algorithms outlined in Section 5 is discussed. RQ4 is explained at length in Section 7.1. In all, we offer some definitive steps forward in the battle against misinformation, of use in a broad number of environments and with attention to a range of today's social issues.

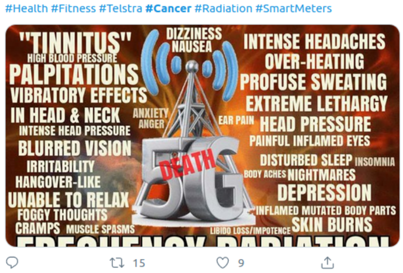
**Author Contributions:** Conceptualization, R.C. and K.M.; writing—review and editing, R.C. A.Y., and A.G.; supervision: R.C., K.M., A.P. and A.G.; writing—original draft, R.C., A.G., A.Y., M.C., G.L., R.Z., Y.J., W.T., J.P., A.P., L.G. All authors have read and agreed to the published version of the manuscript.

**Acknowledgments:** Thanks to the Natural Sciences and Engineering Research Council of Canada for support of this research. We are grateful to the anonymous reviewers for their very valuable feedback on an earlier version of this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

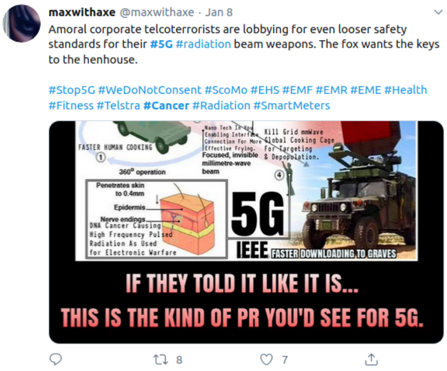
## Appendix A. Compendium of Sample Misinformation

Thisappendix displays several examples of content in social networks which may misinform, along with key descriptors of these social networks.



#Health #Fitness #Telstra **#Cancer** #Radiation #SmartMeters

**Figure A1.** #Health#Fitness#Telstra#Cancer#Radiation#SmartMeters example from Twitter.



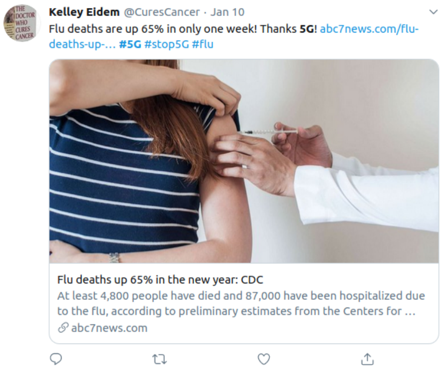
**maxwithaxe** @maxwithaxe · Jan 8

Amoral corporate telcoterrorists are lobbying for even looser safety standards for their **#5G** #radiation beam weapons. The fox wants the keys to the henhouse.

#Stop5G #WeDoNotConsent #ScoMo #EHS #EMF #EMR #EME #Health #Fitness #Telstra **#Cancer** #Radiation #SmartMeters

FASTER HUMAN COOKING
1
360° operation
Nano Tech In You Enabling Interface Connection For More Effective Frying.
Focused, invisible millimetre-wave beam
Kill Grid mmWave Global Cooking Cage For Targeting & Depopulation.
4
Penetrates skin to 0.4mm
Epidermis
Nerve endings
DNA Cancer Causing High Frequency Pulsed Radiation As Used for Electronic Warfare
5G
IEEE
FASTER DOWNLOADING TO GRAVES
IF THEY TOLD IT LIKE IT IS...
THIS IS THE KIND OF PR YOU'D SEE FOR 5G.
8
7

**Figure A2.** maxwithaxe example from Twitter.



**Kelley Eidem** @CuresCancer · Jan 10

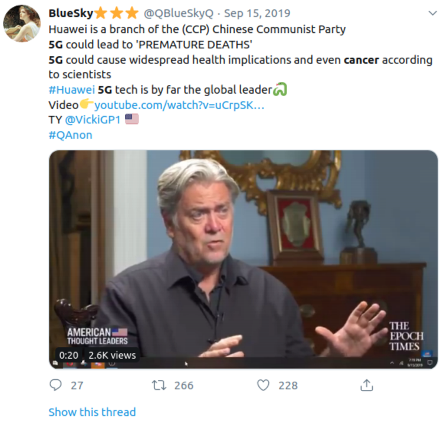
Flu deaths are up 65% in only one week! Thanks **5G**! abc7news.com/flu-deaths-up-... **#5G** #stop5G #flu

Flu deaths up 65% in the new year: CDC

At least 4,800 people have died and 87,000 have been hospitalized due to the flu, according to preliminary estimates from the Centers for ...

abc7news.com

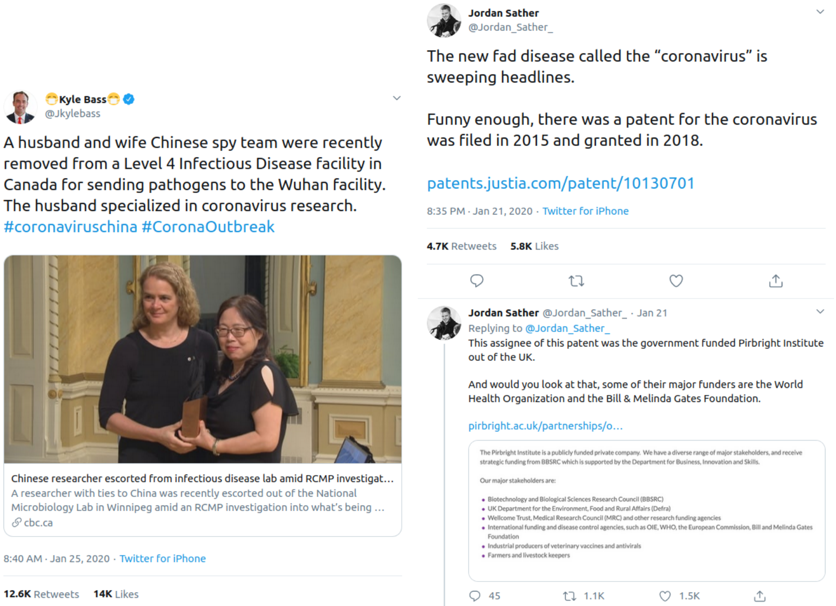
**Figure A3.** Kelley Eidem @CuresCancer example from Twitter.



**BlueSky**★★★ @QBlueSkyQ · Sep 15, 2019

Huawei is a branch of the (CCP) Chinese Communist Party
**5G** could lead to 'PREMATURE DEATHS'
**5G** could cause widespread health implications and even **cancer** according to scientists
#Huawei **5G** tech is by far the global leader🐍
Video👉youtube.com/watch?v=uCrpSK...
TY @VickiGP1 🇺🇸
#QAnon

**Figure A4.** BlueSky example from Twitter.



😷Kyle Bass😷 @Jkylebass

A husband and wife Chinese spy team were recently removed from a Level 4 Infectious Disease facility in Canada for sending pathogens to the Wuhan facility. The husband specialized in coronavirus research. #coronaviruschina #CoronaOutbreak

Chinese researcher escorted from infectious disease lab amid RCMP investigat…
A researcher with ties to China was recently escorted out of the National Microbiology Lab in Winnipeg amid an RCMP investigation into what's being …
cbc.ca

8:40 AM · Jan 25, 2020 · Twitter for iPhone

12.6K Retweets 14K Likes

Jordan Sather @Jordan_Sather_

The new fad disease called the "coronavirus" is sweeping headlines.

Funny enough, there was a patent for the coronavirus was filed in 2015 and granted in 2018.

patents.justia.com/patent/10130701

8:35 PM · Jan 21, 2020 · Twitter for iPhone

4.7K Retweets 5.8K Likes

Jordan Sather @Jordan_Sather_ · Jan 21
Replying to @Jordan_Sather_

This assignee of this patent was the government funded Pirbright Institute out of the UK.

And would you look at that, some of their major funders are the World Health Organization and the Bill & Melinda Gates Foundation.

pirbright.ac.uk/partnerships/o…

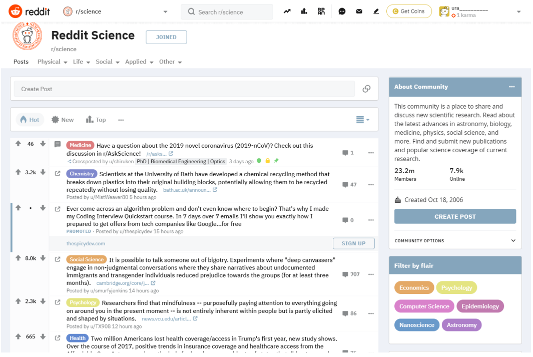
The Pirbright Institute is a publicly funded private company. We have a diverse range of major stakeholders, and receive strategic funding from BBSRC which is supported by the Department for Business, Innovation and Skills.

Our major stakeholders are:

- Biotechnology and Biological Sciences Research Council (BBSRC)
- UK Department for the Environment, Food and Rural Affairs (Defra)
- Wellcome Trust, Medical Research Council (MRC) and other research funding agencies
- International funding and disease control agencies, such as OIE, WHO, the European Commission, Bill and Melinda Gates Foundation
- Industrial producers of veterinary vaccines and antivirals
- Farmers and livestock keepers

45 1.1K 1.5K

**Figure A5.** KyleBass and Jordon Sather examples from Twitter (coronavirus).



46 Medicine Have a question about the 2019 novel coronavirus (2019-nCoV)? Check out this discussion in r/AskScience! /r/asks… Crossposted by u/shiruken PhD | Biomedical Engineering | Optics 3 days ago 1

3.2k Chemistry Scientists at the University of Bath have developed a chemical recycling method that breaks down plastics into their original building blocks, potentially allowing them to be recycled repeatedly without losing quality. bath.ac.uk/announ… Posted by u/MistWeaver80 5 hours ago 47

Ever come across an algorithm problem and don't even know where to begin? That's why I made my Coding Interview Quickstart course. In 7 days over 7 emails I'll show you exactly how I prepared to get offers from tech companies like Google…for free PROMOTED · Posted by u/thespicydev 15 hours ago 0 thespicydev.com

8.0k Social Science It is possible to talk someone out of bigotry. Experiments where "deep canvassers" engage in non-judgmental conversations where they share narratives about undocumented immigrants and transgender individuals reduced prejudice towards the groups (for at least three months). cambridge.org/core/j… Posted by u/smurfyjenkins 14 hours ago 707

2.3k Psychology Researchers find that mindfulness -- purposefully paying attention to everything going on around you in the present moment -- is not entirely inherent within people but is partly elicited and shaped by situations. news.vcu.edu/articl… Posted by u/TX908 12 hours ago 86

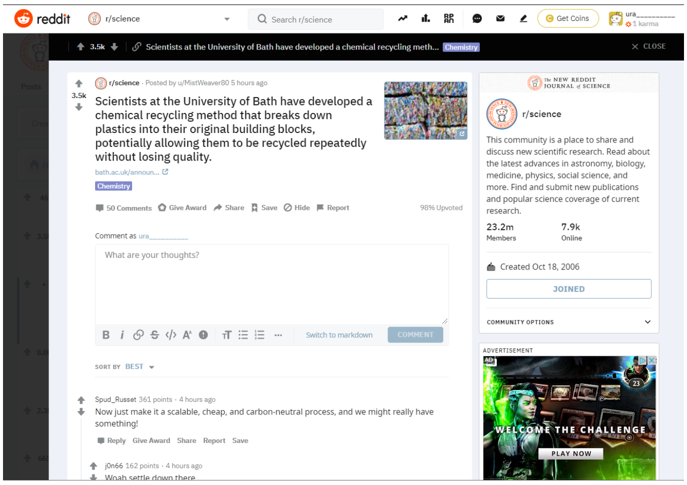
665 Health Two million Americans lost health coverage/access in Trump's first year, new study shows. Over the course of 2017, positive trends in insurance coverage and healthcare access from the

This community is a place to share and discuss new scientific research. Read about the latest advances in astronomy, biology, medicine, physics, social science, and more. Find and submit new publications and popular science coverage of current research.

23.2m Members 7.9k Online

Created Oct 18, 2006

**Figure A6.** Front page of r/science subreddit from Reddit.



3.5k Scientists at the University of Bath have developed a chemical recycling meth… Chemistry

r/science · Posted by u/MistWeaver80 5 hours ago

3.5k Scientists at the University of Bath have developed a chemical recycling method that breaks down plastics into their original building blocks, potentially allowing them to be recycled repeatedly without losing quality.

bath.ac.uk/announ…

98% Upvoted

Comment as ura__________

What are your thoughts?

Spud_Russet 361 points · 4 hours ago
Now just make it a scalable, cheap, and carbon-neutral process, and we might really have something!

j0n66 162 points · 4 hours ago
Woah settle down there

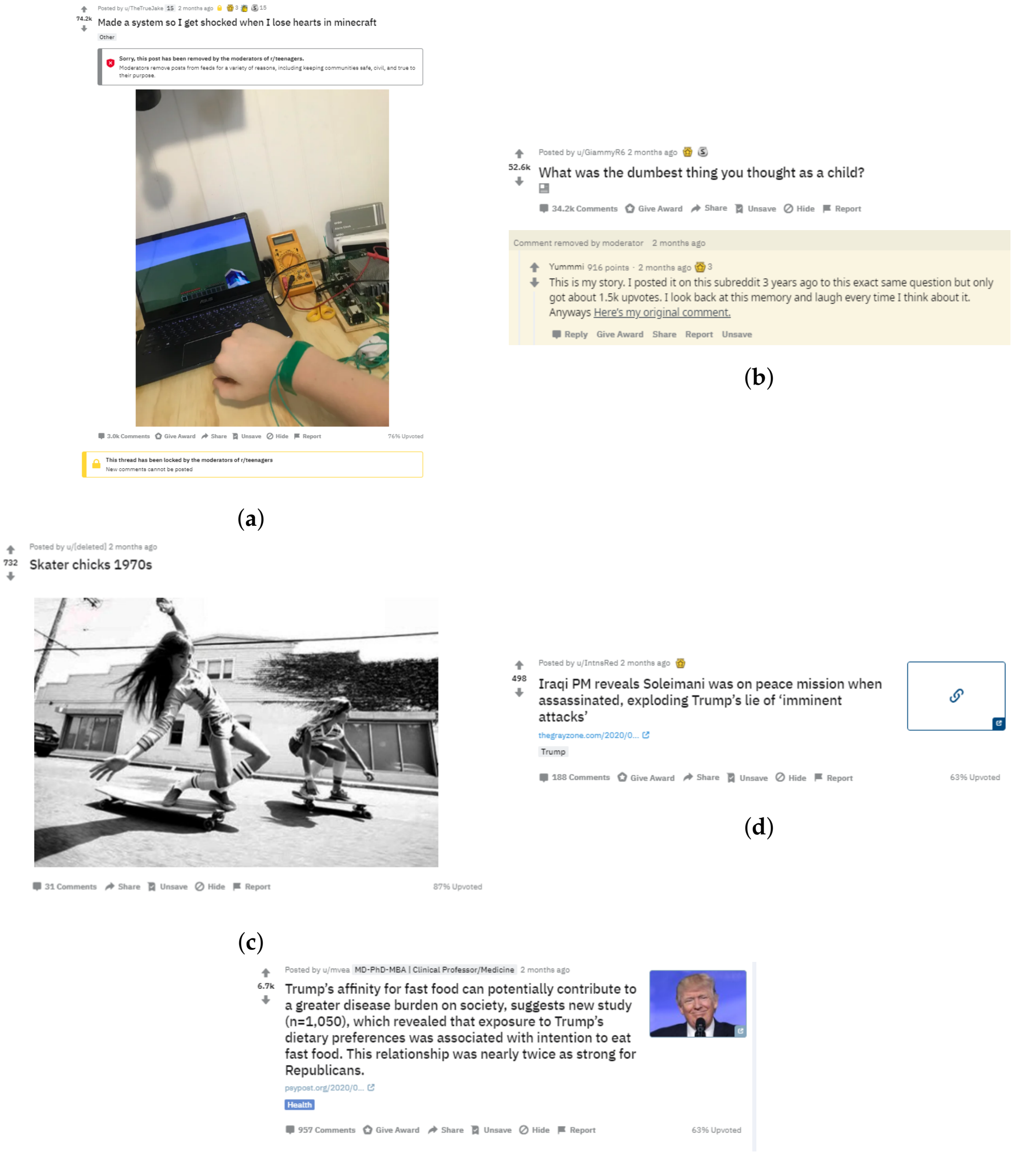
r/science

This community is a place to share and discuss new scientific research. Read about the latest advances in astronomy, biology, medicine, physics, social science, and more. Find and submit new publications and popular science coverage of current research.

23.2m Members 7.9k Online

Created Oct 18, 2006

**Figure A7.** Sample individual page on Reddit.



Posted by u/TheTrueJake 15 2 months ago 3 15

74.2k

Made a system so I get shocked when I lose hearts in minecraft

Other

**Sorry, this post has been removed by the moderators of r/teenagers.**
Moderators remove posts from feeds for a variety of reasons, including keeping communities safe, civil, and true to their purpose.

3.0k Comments Give Award Share Unsave Hide Report 76% Upvoted

**This thread has been locked by the moderators of r/teenagers**
New comments cannot be posted

(**a**)

Posted by u/GiammyR6 2 months ago

52.6k

What was the dumbest thing you thought as a child?

34.2k Comments Give Award Share Unsave Hide Report

Comment removed by moderator 2 months ago

Yummmi 916 points · 2 months ago 3

This is my story. I posted it on this subreddit 3 years ago to this exact same question but only got about 1.5k upvotes. I look back at this memory and laugh every time I think about it. Anyways Here's my original comment.

Reply Give Award Share Report Unsave

(**b**)

Posted by u/[deleted] 2 months ago

732

Skater chicks 1970s

31 Comments Share Unsave Hide Report 87% Upvoted

(**c**)

Posted by u/IntnaRed 2 months ago

498

Iraqi PM reveals Soleimani was on peace mission when assassinated, exploding Trump's lie of 'imminent attacks'

thegrayzone.com/2020/0...

Trump

188 Comments Give Award Share Unsave Hide Report 63% Upvoted

(**d**)

Posted by u/mvea MD-PhD-MBA | Clinical Professor/Medicine 2 months ago

6.7k

Trump's affinity for fast food can potentially contribute to a greater disease burden on society, suggests new study (n=1,050), which revealed that exposure to Trump's dietary preferences was associated with intention to eat fast food. This relationship was nearly twice as strong for Republicans.

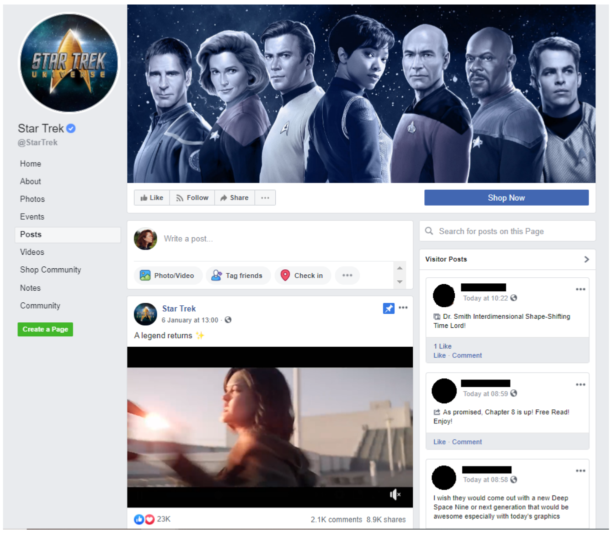
psypost.org/2020/0...

Health

957 Comments Give Award Share Unsave Hide Report 63% Upvoted

(**e**)

**Figure A8.** Misinformation on Reddit. (**a**) User claimed to have developed a certain electronic system. The post received a 74.2 thousand net score and 3 thousand comments, which is considered far above average for the r/teenagers community with 1.8 million subscribers. Many initial responses were positive. Only after a few dozen threads did negative threads begin to show up. User began to notice inconsistencies with the setup in the picture. (**b**) User stole and reposted a story made by another user, claiming as own. Before the community response was removed, it received a 19.2 thousand net score, which is considered a lot for the r/AskReddit community. Initial responses were unsuspecting. However, another user commented and provided definitive evidence that the author stole and reused their story. (**c**) User claimed picture was taken in 1970s but was clearly modern. The post received a 732 net score and 31 comments, which is not a lot compared to top posts in the subreddit but still a decent amount. After just 2 comment threads, a user picked up a detail in the picture that showed it must have been taken much more recently. (**d**) User referred to an article from a questionable source. The post received a 498 net score and 188 comments, which is not a lot compared to top posts in the subreddit, but still a decent amount. A few comment threads in, users started to point out the cited article didn't seem to be very credible and the news wasn't true. (**e**) User posted a link to a questionable source The post received a 6.7 thousand net score and 957 comments, a farily large amount for the subbreddit. Despite the post's high rating, users pointed out that the study was unreliable and inconclusive quite early on. Many also claimed that the author had a history of blindly linking as many studies as possible without verifying which ones seemed reasonable.



Star Trek
6 January at 13:00 ·

A legend returns ✨

23K 2.1K comments 8.9K shares

[redacted]
Today at 10:22

Dr. Smith Interdimensional Shape-Shifting Time Lord!

1 Like
Like · Comment

[redacted]
Today at 08:59

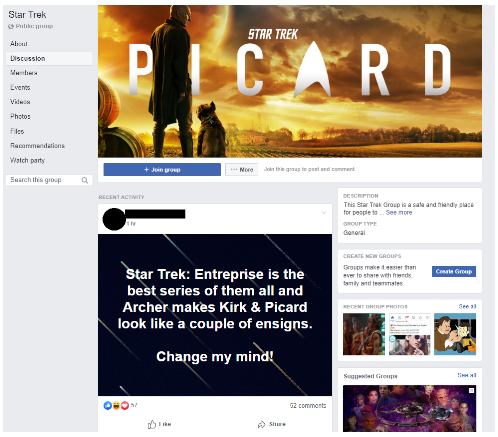
As promised, Chapter 8 is up! Free Read! Enjoy!

Like · Comment

[redacted]
Today at 08:58

I wish they would come out with a new Deep Space Nine or next generation that would be awesome especially with today's graphics

**Figure A9.** Sample Facebook Page.



[redacted]
1 hr

Star Trek: Entreprise is the best series of them all and Archer makes Kirk & Picard look like a couple of ensigns.

Change my mind!

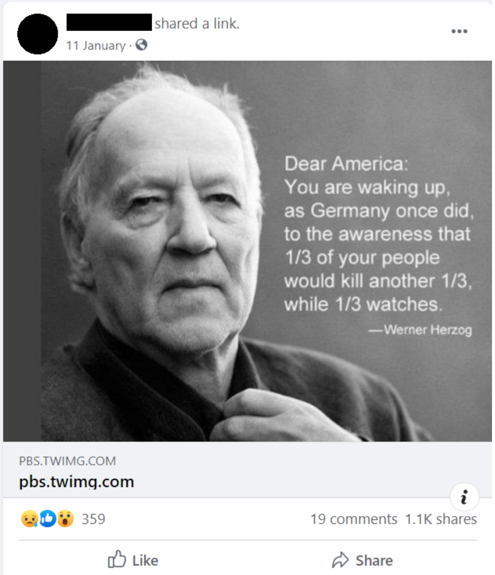
57 52 comments

This Star Trek Group is a safe and friendly place for people to ... See more

General

Groups make it easier than ever to share with friends, family and teammates.

**Figure A10.** Sample Facebook Group.



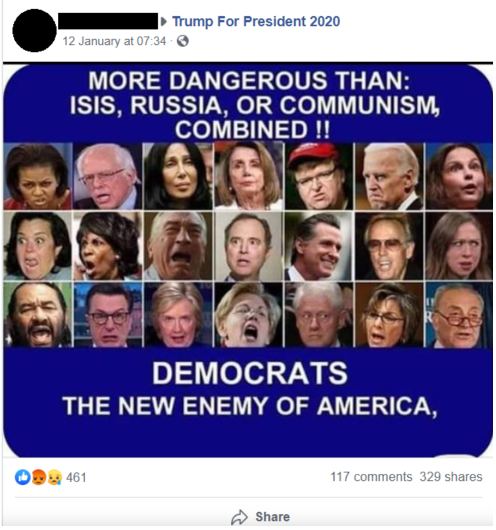
[redacted] shared a link.
11 January ·

Dear America:
You are waking up,
as Germany once did,
to the awareness that
1/3 of your people
would kill another 1/3,
while 1/3 watches.
—Werner Herzog

PBS.TWIMG.COM
pbs.twimg.com

359 19 comments 1.1K shares

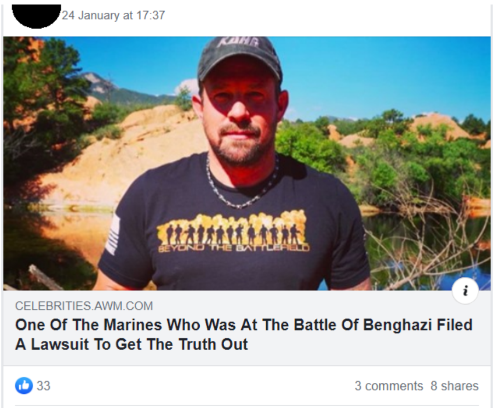
**Figure A11.** Facebook misattributed quote.



[redacted] [redacted] ▸ Trump For President 2020
12 January at 07:34 ·

461 117 comments 329 shares

**Figure A12.** Facebook inflammatory post.

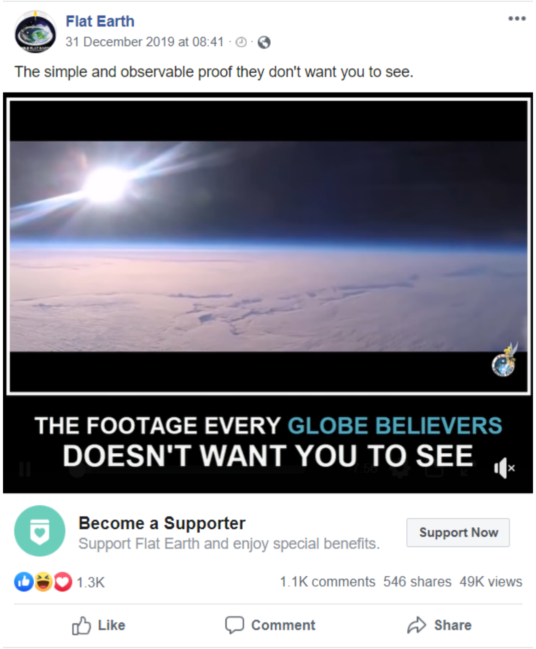


[redacted] 24 January at 17:37

CELEBRITIES.AWM.COM
One Of The Marines Who Was At The Battle Of Benghazi Filed A Lawsuit To Get The Truth Out

33 3 comments 8 shares

**Figure A13.** Facebook suspected content, low response.



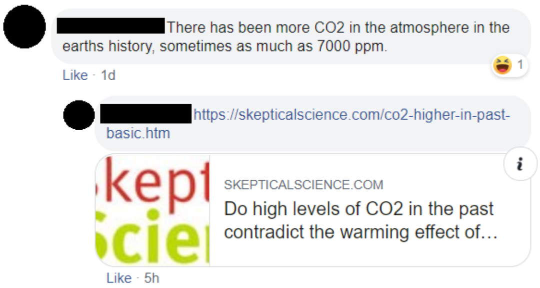
Flat Earth
31 December 2019 at 08:41 ·

The simple and observable proof they don't want you to see.

Become a Supporter
Support Flat Earth and enjoy special benefits.

1.3K 1.1K comments 546 shares 49K views

**Figure A14.** Facebook suspected content, joke.



[redacted] [redacted] There has been more CO2 in the atmosphere in the earths history, sometimes as much as 7000 ppm.

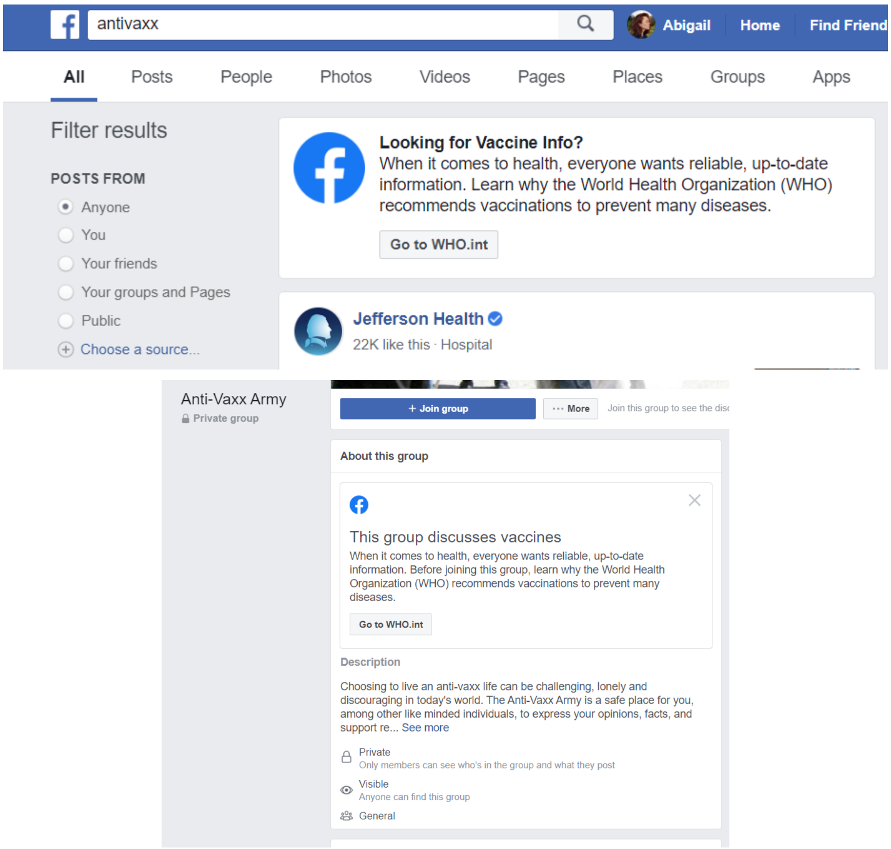
Like · 1d

[redacted] [redacted] https://skepticalscience.com/co2-higher-in-past-basic.htm

SKEPTICALSCIENCE.COM
Do high levels of CO2 in the past contradict the warming effect of…

Like · 5h

**Figure A15.** Facebook example of ideology.



antivaxx

Abigail Home Find Friend

All Posts People Photos Videos Pages Places Groups Apps

Filter results

POSTS FROM

● Anyone
○ You
○ Your friends
○ Your groups and Pages
○ Public
⊕ Choose a source...

**Looking for Vaccine Info?**
When it comes to health, everyone wants reliable, up-to-date information. Learn why the World Health Organization (WHO) recommends vaccinations to prevent many diseases.

Go to WHO.int

Jefferson Health
22K like this · Hospital

Anti-Vaxx Army
Private group

+ Join group ··· More Join this group to see the disc

About this group

This group discusses vaccines
When it comes to health, everyone wants reliable, up-to-date information. Before joining this group, learn why the World Health Organization (WHO) recommends vaccinations to prevent many diseases.

Go to WHO.int

Description

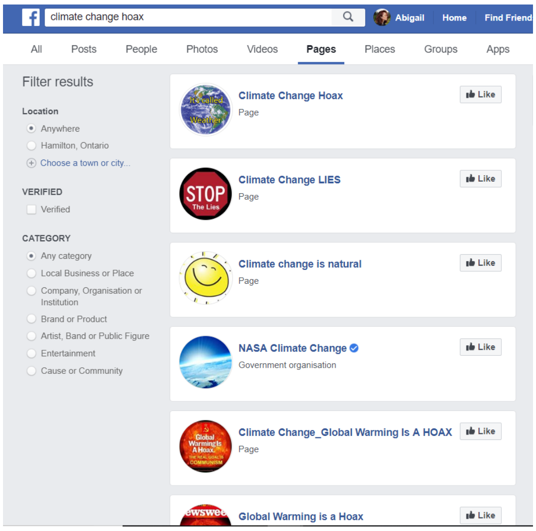
Choosing to live an anti-vaxx life can be challenging, lonely and discouraging in today's world. The Anti-Vaxx Army is a safe place for you, among other like minded individuals, to express your opinions, facts, and support re... See more

Private
Only members can see who's in the group and what they post

Visible
Anyone can find this group

General

**Figure A16.** Facebook healthcare example.



climate change hoax

Abigail Home Find Friend

All Posts People Photos Videos **Pages** Places Groups Apps

Filter results

Location
● Anywhere
○ Hamilton, Ontario
⊕ Choose a town or city...

VERIFIED
□ Verified

CATEGORY
● Any category
○ Local Business or Place
○ Company, Organisation or Institution
○ Brand or Product
○ Artist, Band or Public Figure
○ Entertainment
○ Cause or Community

Climate Change Hoax
Page
Like

Climate Change LIES
Page
Like

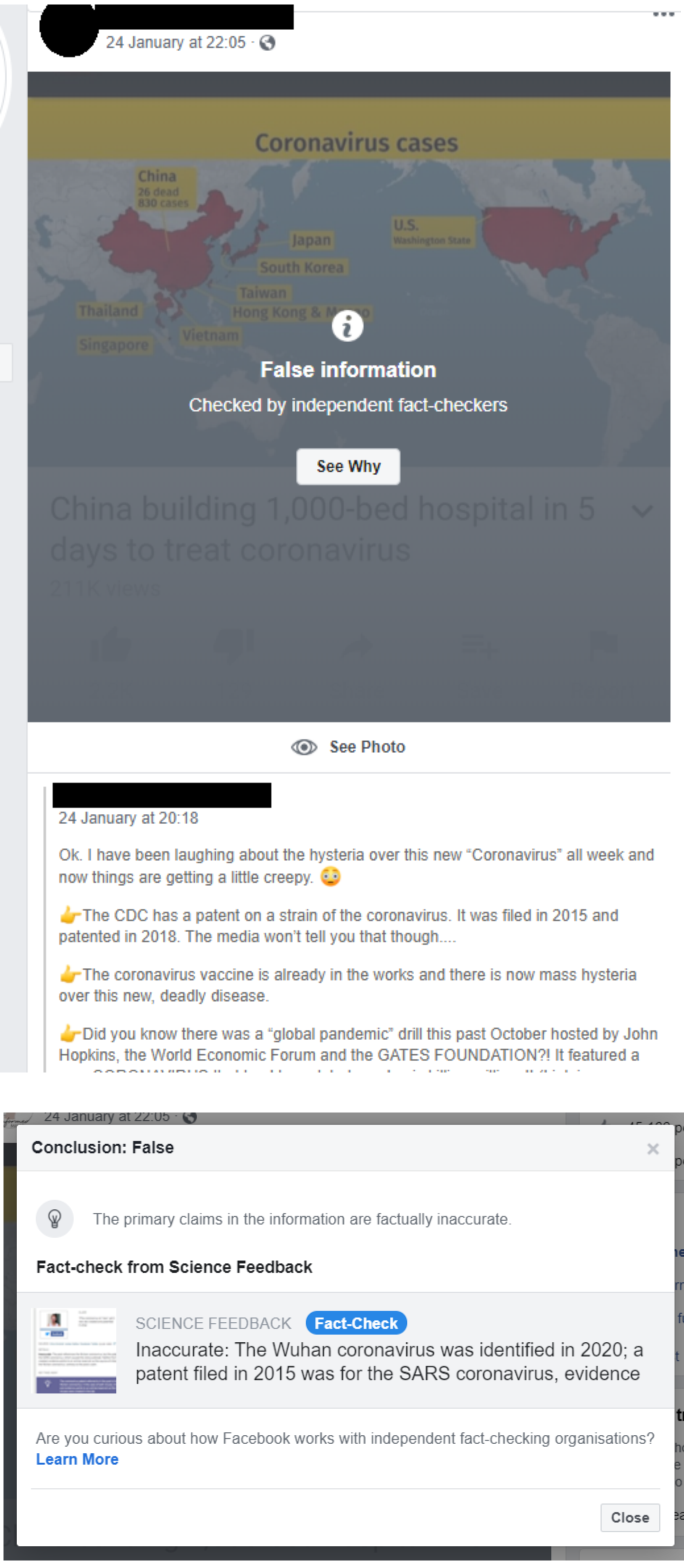
Climate change is natural
Page
Like

NASA Climate Change
Government organisation
Like

Climate Change_Global Warming Is A HOAX
Page
Like

Global Warming is a Hoax
Like

**Figure A17.** Facebook climate change example.



[redacted]

24 January at 22:05 ·

**False information**

Checked by independent fact-checkers

See Why

China building 1,000-bed hospital in 5 days to treat coronavirus

21.1K views

See Photo

[redacted]

24 January at 20:18

Ok. I have been laughing about the hysteria over this new "Coronavirus" all week and now things are getting a little creepy. 😳

👉The CDC has a patent on a strain of the coronavirus. It was filed in 2015 and patented in 2018. The media won't tell you that though....

👉The coronavirus vaccine is already in the works and there is now mass hysteria over this new, deadly disease.

👉Did you know there was a "global pandemic" drill this past October hosted by John Hopkins, the World Economic Forum and the GATES FOUNDATION?! It featured a

24 January at 22:05 ·

**Conclusion: False** ×

The primary claims in the information are factually inaccurate.

**Fact-check from Science Feedback**

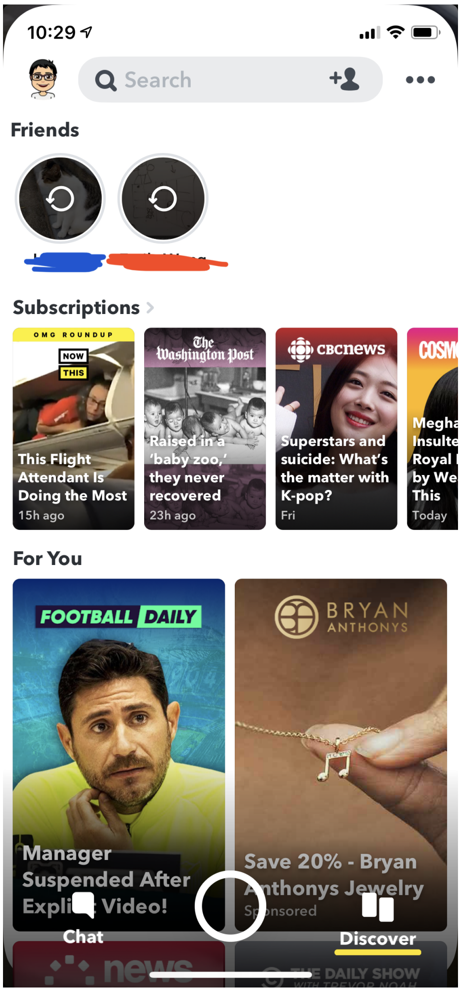
SCIENCE FEEDBACK **Fact-Check**

Inaccurate: The Wuhan coronavirus was identified in 2020; a patent filed in 2015 was for the SARS coronavirus, evidence

Are you curious about how Facebook works with independent fact-checking organisations? **Learn More**

Close

**Figure A18.** Facebook coronavirus example (fact checking).



10:29

Search

Friends

[redacted] [redacted]

Subscriptions

OMG ROUNDUP NOW THIS
This Flight Attendant Is Doing the Most
15h ago

The Washington Post
Raised in a 'baby zoo,' they never recovered
23h ago

CBCnews
Superstars and suicide: What's the matter with K-pop?
Fri

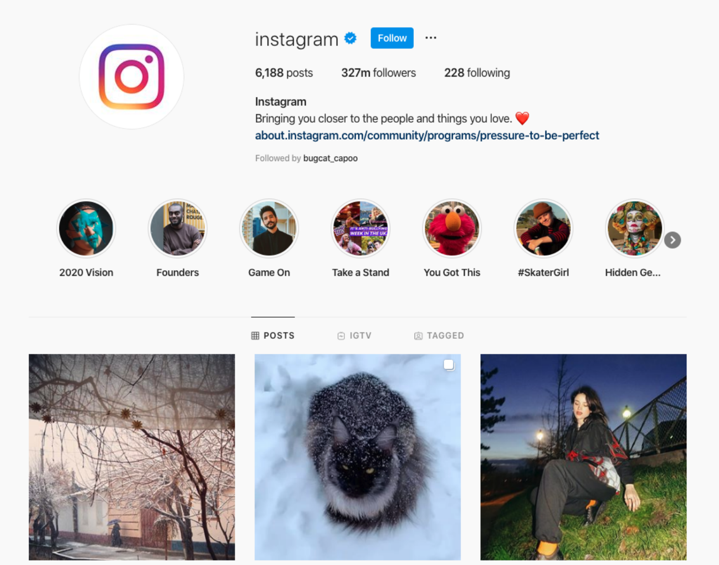
COSM
Megha Insulte Royal by We This
Today

For You

FOOTBALL DAILY
Manager Suspended After Expli Video!

BRYAN ANTHONYS
Save 20% - Bryan Anthonys Jewelry
Sponsored

**Figure A19.** Snapchat samples.



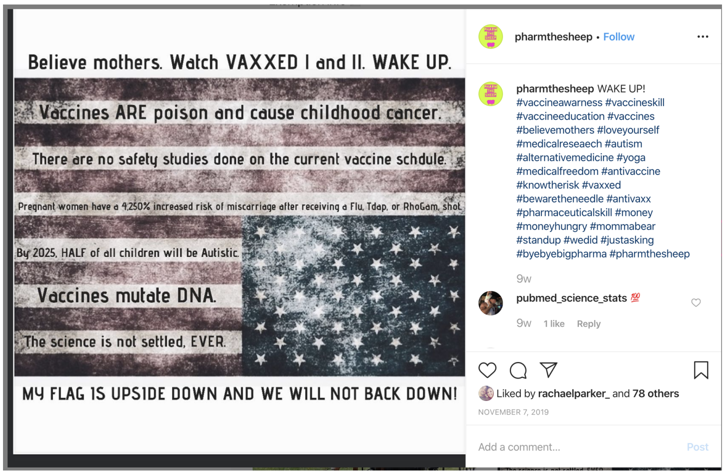
instagram Follow ···

6,188 posts 327m followers 228 following

Instagram
Bringing you closer to the people and things you love. ❤️
about.instagram.com/community/programs/pressure-to-be-perfect

Followed by bugcat_capoo

**Figure A20.** Instagram samples.



Believe mothers. Watch VAXXED I and II. WAKE UP.

Vaccines ARE poison and cause childhood cancer.

There are no safety studies done on the current vaccine schdule.

Pregnant women have a 4,250% increased risk of miscarriage after receiving a Flu, Tdap, or RhoGam, shot.

By 2025, HALF of all children will be Autistic.

Vaccines mutate DNA.

The science is not settled, EVER.

MY FLAG IS UPSIDE DOWN AND WE WILL NOT BACK DOWN!

pharmthesheep • Follow

pharmthesheep WAKE UP! #vaccineawarness #vaccineskill #vaccineeducation #vaccines #believemothers #loveyourself #medicalreseaech #autism #alternativemedicine #yoga #medicalfreedom #antivaccine #knowtherisk #vaxxed #bewaretheneedle #antivaxx #pharmaceuticalskill #money #moneyhungry #mommabear #standup #wedid #justasking #byebyebigpharma #pharmthesheep

9w

pubmed_science_stats 💯

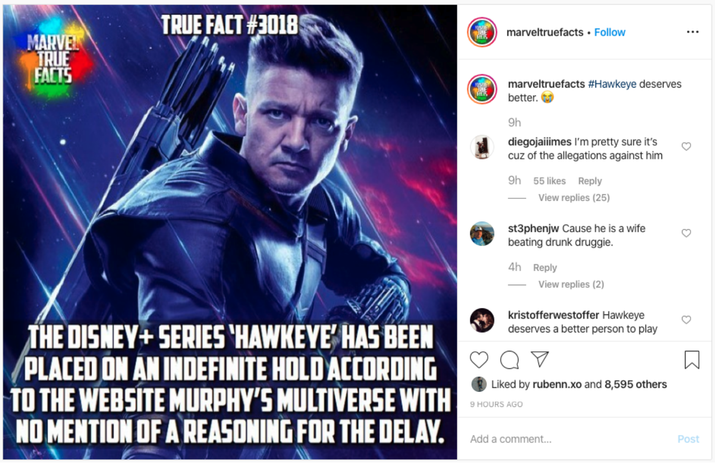
9w 1 like Reply

Liked by rachaelparker_ and 78 others

NOVEMBER 7, 2019

Add a comment... Post

**Figure A21.** Instagram healthcare example.



MARVEL TRUE FACTS

TRUE FACT #3018

THE DISNEY+ SERIES 'HAWKEYE' HAS BEEN PLACED ON AN INDEFINITE HOLD ACCORDING TO THE WEBSITE MURPHY'S MULTIVERSE WITH NO MENTION OF A REASONING FOR THE DELAY.

marveltruefacts • Follow

marveltruefacts #Hawkeye deserves better. 😭

9h

diegojaiiimes I'm pretty sure it's cuz of the allegations against him

9h 55 likes Reply

View replies (25)

st3phenjw Cause he is a wife beating drunk druggie.

4h Reply

View replies (2)

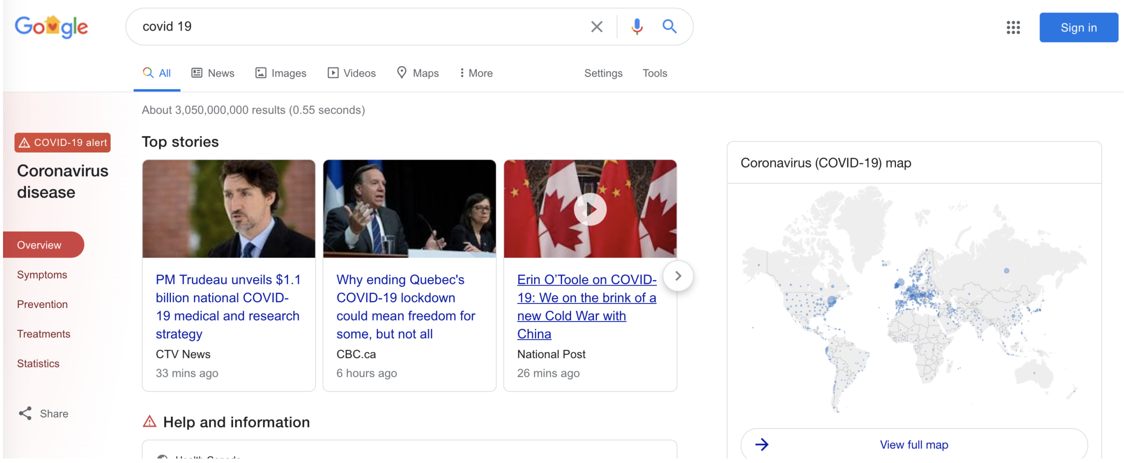
kristofferwestoffer Hawkeye deserves a better person to play

Liked by rubenn.xo and 8,595 others

9 HOURS AGO

Add a comment... Post

**Figure A22.** Instagram Disney example.



covid 19

About 3,050,000,000 results (0.55 seconds)

COVID-19 alert

Coronavirus disease

Top stories

PM Trudeau unveils $1.1 billion national COVID-19 medical and research strategy
CTV News
33 mins ago

Why ending Quebec's COVID-19 lockdown could mean freedom for some, but not all
CBC.ca
6 hours ago

Erin O'Toole on COVID-19: We on the brink of a new Cold War with China
National Post
26 mins ago

Help and information

Coronavirus (COVID-19) map

View full map

**Figure A23.** Google acting on misinformation (coronavirus).

## Appendix B. Key Processes for Reasoning about Misinformation

In this appendix we display in more detail the proposed algorithm to run for Twitter and the key decision process for the algorithm to run for Reddit, when detecting misinformation. Ghenai's Fake Cures model is as in [30,31] whereas Ghenai's Zika model is presented in [5,31]. The former aims to identify users who spread rumours, while the latter aims to isolate misinformation.

```
$cur_user <- current user
select a subset of users $U from all users on Twitter
for each user $u in $U
generate credibility $u_c for user $u with Ghenai's ''Fake Cures'' model
generate similarity $u_s for user $u with respect to $cur_user with modified PCC
end for

for each user $u in $U
select all tweets $t_u under user $u
for each tweet $t in $t_u
$rating <- NULL
if $t has a response tree of depth 2 with at least two leaves
$rating <- trustworthiness estimated by the Reddit model
given tweet $t's response tree
else
$rating <- Ghenai's Zika model applied to tweet $t
end if
$trust <- Sardana's POMDPTrust with ($rating, $u_s, $u_c)
if $trust < 0.45 // some threshold here
flag($t)
end if
end for
end for
```

**Figure A24.** Twitter algorithm to detect misinformation.

```
$estimate_users <- {}
for each tweet $t in first $T search results $results
$estimate_users <- union($estimate_users, user corresponding to $t)
end for

run previous algorithm on $estimate_users to establish the rating of their tweets

$misinformative_tweets <- 0
for each tweet $t in $results
// tweet could be flagged by the algorithm above
if flagged($t)
$misinformed_tweets <- $misinformed_tweets + 1
end if
end for
if $misinformed_tweets > $T / 2:
post warning at top of search results list
end if
```

**Figure A25.** Twitter algorithm to inform users.

**Table A1.** Reddit factors to consider to detect misinformation.

| Feature | Possible Values |
|---|---|
| Message type | {post, comment} |
| Net score | integer |
| Number of comments | integer |
| Highest score on follow-up comment with sentiment of distrust | integer |
| Intentional? | {yes, no} |
| Author karma | integer |
| Author is repeat offender? | {yes, no} |
| Easily verifiable? | {yes, no} |
| Supporting evidence provided? | {yes, no} |
| Informal use of language? | {yes, no} |

**Table A2.** Description of Reddit factors to consider to detect misinformation.

| Feature | Description |
|---|---|
| Message type | Whether the message in question is a post or a comment. |
| Net score | The net score of a post/comment. This is an indication of the positivity/negativity in the reactions of the community to the message. As this is approximately number of upvotes minus number of downvotes (i.e., a difference), it should be compared with the number of comments to estimate a proportion. |
| Number of comments | This is an indication of the amount of attention the message received. Since the size of a subreddit influences visibility of its content, to get a more relative estimate, one could compare the number of comments with the number of followers of the subreddit, or the number of comments in the highest rated posts on that subreddit. |
| Highest score on follow-up comment with sentiment of distrust | This is an indication of how distrustful the community is to the message. This should be compared with the number of comments or net score for a more relative estimate. |
| Intentional? | Whether the author accidentally or intentionally tried to spread misinformation. For the most part, this can only come as an impression. An author that apologizes in follow-up comments could indicate an accident, whereas an author that denies untruthfulness with additional poor arguments could indicate they acted intentionally. |
| Author karma | The author's total karma (i.e., score) for past posts/comments. This is an indication of how credible an author is. |
| Author is repeat offender? | This is another indication of how credible an author is. |
| Easily verifiable? | Whether it is easy to verify the truthfulness of the message. This may influence the highest score on a follow-up comment with a sentiment of distrust as fewer members of the community would be able to verify the message. Things such as general facts or public events would be easier to verify, whereas things such as personal stories would be more difficult. |
| Supporting evidence provided? | This is an indication for the credibility of a post. |
| Informal use of language? | This is an indication for the credibility of a post. |

## Appendix C. Sardana's Model

BayesTrust is a recent multiagent trust-based message recommendation system that filters messages for a particular agent based on the advice of other agents in the network. The goal of this system is to save the time and attention of social network readers by reasoning about which messages will be most beneficial to them, and filtering out or flagging messages which do not appear to be beneficial. This system is modeled as a Partially Observable Markov Decision Process (POMDP) that considers one message at a time in a purely episodic manner. The POMDP is formally defined by a tuple of parameters: $(S, A, O, T, \Omega, R, \gamma, h)$.

- $S$: The state space for a message, defined as {good, bad}
- $A$: The action space for the agent, defined as {accept, reject, elicit_advice}
- $O$: The set of observations associated with actions. In this work, the accept and reject actions always result in a nil observation, while requesting advice results in a tuple $(r, m, c)$, where $r$ is the rating a peer has given to the message, $m$ is the similarity that peer has to the recomendee, and $c$ is the credibility of the peer.
- $T = Pr(s'|s, a)$: The transition function for message state given the current state and the chosen action.
- $\Omega = Pr(o'|s', a)$: The probability of seeing observation $o$ given action $a$ is taken at $t_i$ and the state at $t_{i+1}$ is $s'$.
- $R : S \times A \to \mathbb{R}$: The reward function representing the utility of each state-action pair for a particular agent.
- $\gamma$: Reward discount factor.
- $h$: Horizon (finite or infinite).

Given this POMDP formulation, the system reasons about the expected utility of showing a message to a user, by repeatedly requesting advice from the peer network, iteratively updating the belief over the state of the message.

$$b_{t+1}(s_{t+1}) \propto Pr(o_{t+1}|s_{t+1}, a_t) \sum_{s_t} Pr(s_{t+1}|s_t, a_t) b_t(s_t)$$

where $b_t$ is the belief that the message is of the given state at time $t$. After advice from the peer network is exhausted, it is straightforward to reason about the expected utility of either showing (recommending) or hiding (rejecting) the message from the user:

$$\begin{aligned} EU_{show} &= b(G) \cdot R(show, G) + b(B) \cdot R(show, B) \\ EU_{hide} &= b(G) \cdot R(hide, G) + b(B) \cdot R(hide, B) \end{aligned}$$

where $b$ is the belief that the message is either $G$, good or $B$, bad (i.e., the product of the probabilities of seeing the observed sequence of advice given the underlying state, according to $\Omega$) and $R$ is the reward function that encodes the reward of showing or hiding good and bad messages respectively. A user will be shown the message only if $EU_{show} > EU_{hide}$

## Appendix D. Illustration on Sample Misinformation

We return to the model that was sketched in Section 5.2, for Reddit. In order to drill down to demonstrate our methods operating on a specific example we include display the algorithms that would operate in more detail.

One distinguishing point of the model is its effort to distinguish distrust from negativity. Algorithm A1 begins with a step making use of a Sardana-inspired POMDP.

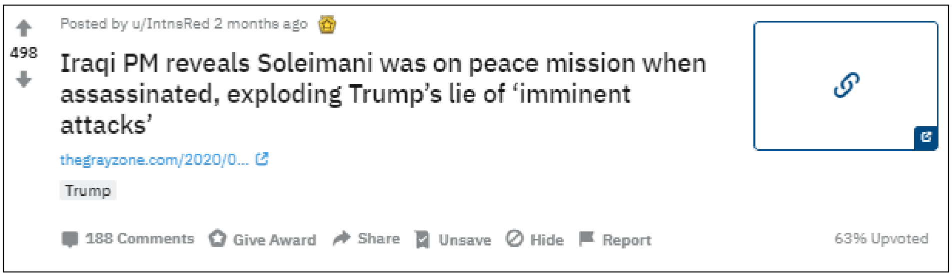
**Algorithm A1:** Belief Updating

```
input  : $A$: set of advisors who responded to the message
         $C$: credibility of advisors
         $R$: responses to the message (votes and comments)
         $V$: past ratings given by advisers
output : $b$: good/bad belief for message (pair of [0,1] values)
$b \leftarrow$ initial beliefs;
for $a \in A$ do
    $o \leftarrow$ getObservation() ; // tuple obtained by polling advisor for advice
    $b \leftarrow$ updateBeliefs($b, C, R, V$) ; // from o using POMDP approach
end
return $b$;
```

Once an agent has generated beliefs about a message, it returns an observation on a message. For this, a reasoning process would produce an observation tuple of (rating, similarity, distrust rating, distrust similarity, credibility) where advisors who both voted and commented could have distrust assessed, and where the rating history of an agent and advisor can be examined for similarity. The final step, that of deciding whether to display the message or not, would be based on a weighted combination of likelihood of misinformation, severity and user tolerance so that when above a certain threshold, the message will not be displayed.

*Example*

Consider the following instance of misinformation found on Reddit in the r/worldnews subreddit, where a user author made a post about recent news and referenced an article.



Posted by u/IntnsRed 2 months ago

498

Iraqi PM reveals Soleimani was on peace mission when assassinated, exploding Trump's lie of 'imminent attacks'

thegrayzone.com/2020/0...

Trump

188 Comments

63% Upvoted

**Figure A26.** Example of Reddit Misinformation for our Algorithm.

The post received 498 net score and 188 comments, which is not a lot compared to top posts in the subreddit, but still a decent amount. Within a few comment threads in, users started to point out the cited article did not seem to be very credible and the news was not true.

To detect misinformation, we will first supply the algorithms with some input:

- the rating information—which users voted, what did they vote
- comments on the post
- past ratings of users

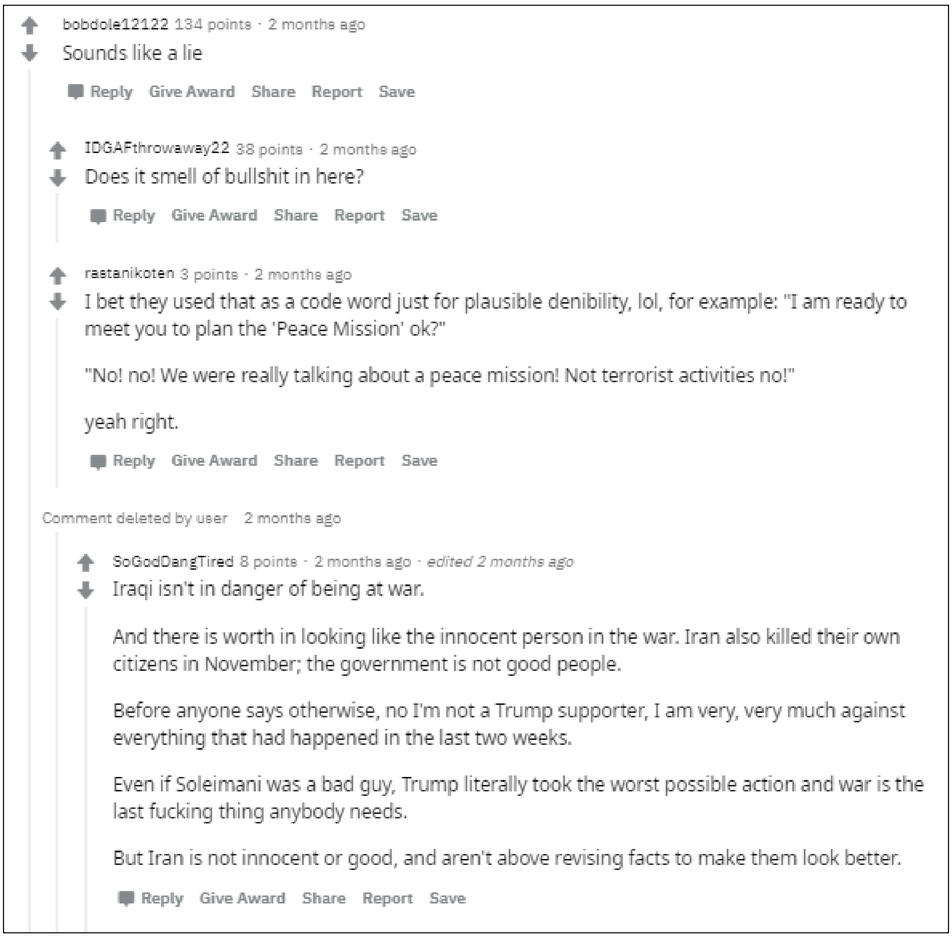
The agent responsible for making a decision about the post will use Algorithm A1 which will run the extended POMDP model to assess the belief of whether this message is good or bad. This involves polling advisors for advice and using the received advice to update the computed beliefs. In this case, each advisor represents a user who responded to the post (either voted or commented or both). The returned advice is an observation tuple of the form (rating, similarity, distrust rating, distrust similarity, credibility).

Each advisor that has been polled for advice will generate an observation tuple. The most important fields are the rating and distrust rating. The rating is primarily based on voting. If an advisor upvoted this post, a rating of 1 is assigned. If an advisor downvoted this post, a rating of 0 is assigned. In this Example, about 63% of users upvoted the post. On the other hand, the distrust rating is based on whether comments have a sentiment of distrust. In this case several comments express distrust, although the majority are fairly neutral and focused on discussion of the topic.

Once the agent has assessed beliefs about this message, it will address the message by reasoning about:

- likelihood that the post contains misinformation—the belief that this message is bad, the output from detecting misinformation
- severity—some value based on how severe the consequences may be if this post is misinterpreted. As the subreddit is about world news and this post covers a serious political event, this value would be somewhat high here.
- user tolerance for misinformation—depends on the user

The agent will then make a simple decision on whether to display the message or not.



bobdole12122 134 points · 2 months ago

Sounds like a lie

IDGAFthrowaway22 38 points · 2 months ago

Does it smell of bullshit in here?

rastanikoten 3 points · 2 months ago

I bet they used that as a code word just for plausible denibility, lol, for example: "I am ready to meet you to plan the 'Peace Mission' ok?"

"No! no! We were really talking about a peace mission! Not terrorist activities no!"

yeah right.

Comment deleted by user 2 months ago

SoGodDangTired 8 points · 2 months ago · *edited 2 months ago*

Iraqi isn't in danger of being at war.

And there is worth in looking like the innocent person in the war. Iran also killed their own citizens in November; the government is not good people.

Before anyone says otherwise, no I'm not a Trump supporter, I am very, very much against everything that had happened in the last two weeks.

Even if Soleimani was a bad guy, Trump literally took the worst possible action and war is the last fucking thing anybody needs.

But Iran is not innocent or good, and aren't above revising facts to make them look better.

**Figure A27.** Display of Thread for Example.

The main interest is how the agent evaluates the likelihood that the post contains misinformation. Although the proportion of users who did not respond negatively and did not respond distrustfully is higher for this Example, there is a fairly large proportion of users who did one of those. So, there is likely to be a noticeable difference between the rating and distrust rating of advisors for this post than for a post with no misinformation. In addition, having a distrust rating further distinguishes posts with distrustful responses from posts with generally negative responses. Having many advisors with a higher distrust rating is one of the more convincing arguments for the presence of misinformation.

So, depending on how well the well the model is configured and how much weight is given to some types of observations over others, there is reason to believe this algorithm could perform well.

## Appendix E. Ghenai's Rumour Spread Model

This appendix provides more detail on the rumour spread model of Ghenai [5]. This model presents a pipeline incorporating expert knowledge, crowdsourcing and machine learning four health-related rumor discovery in a social media stream. The steps taken to build this tool are shown in below. The study shows that tracking health misinformation in social media is not trivial, and requires some expert supervision. This can then be augmented by "crowd" workers in order to provide additional annotation of the captured rumour-related tweets. The proposed work in [5] shows the importance for collaborations between health professionals and data researchers in order to quickly understand and mitigate health misinformation on social media.

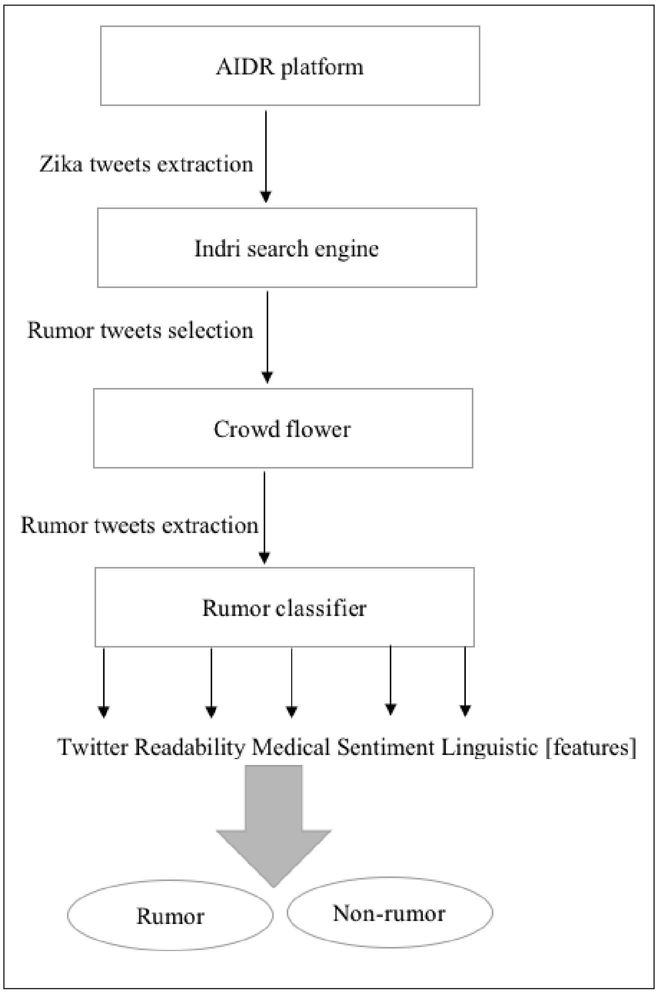


**Figure A28.** Control flow of the model.

In order to identify users who may be spreading rumors in Twitter, Ghenai conducted a study focusing on the issue of promoting fake cures for cancer [30,31]. A rumor group was promoting questionable treatments whereas the control group posted about cancer, but without discussing any of the rumor topics. Queries were crafted, with the assistance of an oncologist, in order to properly characterize the set of possible rumors. Several steps were performed as well to remove from focus tweets arriving from bots and those that matched the queries but were in fact out of focus for the topic; this included some filtering on the basis of average tweeting rate of the user (retaining those with less than 24 tweets per day). Ghenai also focused on the most recent tweets when assembling the set of data. Crowdsourced labelling was employed but simply to confirm that the tweets were discussing cures (not to pass judgment on the veracity of the claims). After this, logistic regression classifier users were trained, where rumor users were characterized as ones claiming a cure for cancer (vs. simply discussing prevention or debunking). Ghenai examined various behavioural indicators such as number of followers and number being followed, average mentions and user age. Focus was on examining posts before rumors had been spread significantly. The list of features employed for the analysis also built upon those introduced (e.g., including language) in [5]. Ghenai discovered that users identified as spreading rumors in fact used more sophisticated language, circulate in health domains before posting a rumor and are likely not personally involved in the illness. This suggests that actors other than patients are responsible for the promotion of cancer misinformation. Figure A29 provides some detail on the data flow in the study.

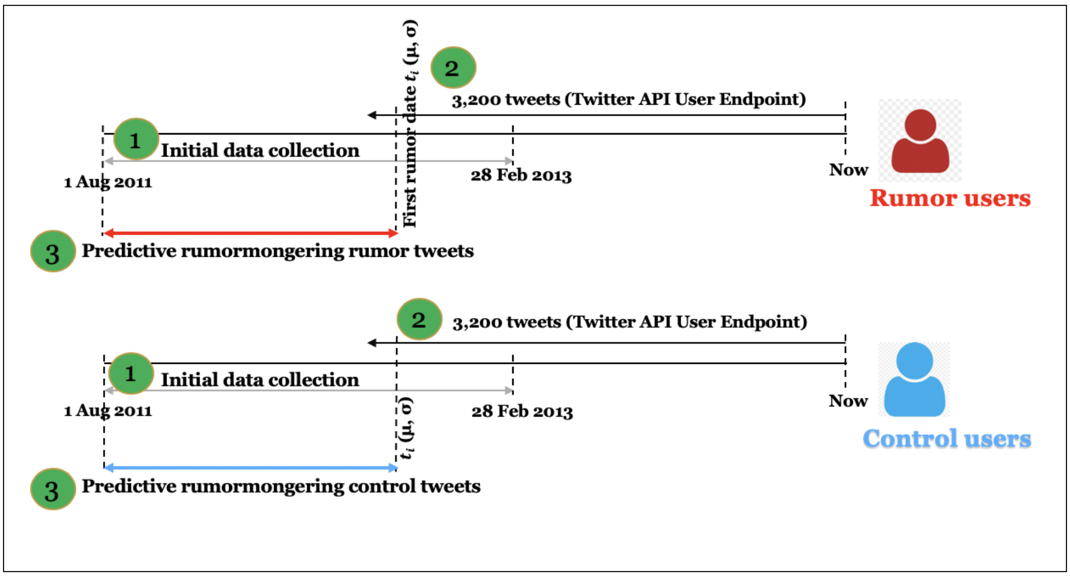

2
3,200 tweets (Twitter API User Endpoint)
First rumor date $t_i$ (μ, σ)
1
Initial data collection
1 Aug 2011
28 Feb 2013
Now
Rumor users
3
Predictive rumormongering rumor tweets
2
3,200 tweets (Twitter API User Endpoint)
1
Initial data collection
1 Aug 2011
28 Feb 2013
$t_i$ (μ, σ)
Now
Control users
3
Predictive rumormongering control tweets

**Figure A29.** Data flow of the model.

## References

1. Shao, C.; Ciampaglia, G.; Varol, O.; Flammini, A.; Menczer, F.; Yang, K.C. The spread of low-credibility content by social bots. *Nat. Commun.* **2018**, *9*, 4787. [CrossRef] [PubMed]
2. Sardana, N.; Cohen, R.; Zhang, J.; Chen, S. A Bayesian Multiagent Trust Model for Social Networks. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 995–1008.
3. Parmentier, A.; Cohen, R. Learning User Reputation on Reddit. In Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2019), Thessaloniki, Greece, 14–17 October 2019; Barnaghi, P.M., Gottlob, G., Manolopoulos, Y., Tzouramanis, T., Vakali, A., Eds.; ACM: New York, NY, USA, 2019; pp. 242–247. [CrossRef]
4. Parmentier, A.; Cohen, R. Personalized Multi-Faceted Trust Modeling in Social Networks. In Proceedings of the Advances in Artificial Intelligence—33rd Canadian Conference on Artificial Intelligence (Canadian AI 2020), Ottawa, ON, Canada, 13–15 May 2020; pp. 445–450. [CrossRef]
5. Ghenai, A.; Mejova, Y. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI 2017), Park City, UT, USA, 23–26 August 2017; p. 518. [CrossRef]
6. Wang, Y.; Singh, M.P. Evidence-Based Trust: A Mathematical Model Geared for Multiagent Systems. *ACM Trans. Auton. Adapt. Syst.* **2010**, *5*, 1–28. [CrossRef]
7. Teacy, W.; Patel, J.; Jennings, N.; Luck, M. TRAVOS: Trust and Reputation in the Context of Inaccurate Information Sources. *Auton. Agents Multi Agent Syst.* **2006**, *12*, 183–198. [CrossRef]

8. Burnett, C.; Norman, T.J.; Sycara, K.P. Trust Decision-Making in Multi-Agent Systems. In Proceedings of the IJCAI 2011 the 22nd International Joint Conference on Artificial Intelligence, Catalonia, Spain, 16–22 July 2011; pp. 115–120. [CrossRef]
9. Sabater-Mir, J.; Sierra, C. Review on Computational Trust and Reputation Models. *Artif. Intell. Rev.* **2005**, *24*, 33–60. [CrossRef]
10. Granatyr, J.; Botelho, V.; Lessing, O.R.; Scalabrin, E.E.; Barthès, J.P.; Enembreck, F. Trust and Reputation Models for Multiagent Systems. *ACM Comput. Surv.* **2015**, *48*, 1–42. [CrossRef]
11. Sen, S.; Rahaman, Z.; Crawford, C.; Yücel, O. *Agents for Social (Media) Change*; International Foundation for Autonomous Agents and Multiagent Systems: Richland, DC, USA, 2018; pp. 1198–1202.
12. Sapienza, A.; Falcone, R. How to Manage the Information Sources' Trustworthiness in a Scenario of Hydrogeological Risks. In Proceedings of the 18th International Workshop on Trust in Agent Societies co-located with the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), Singapore, 10 May 2016; pp. 71–82.
13. Cormier, M.; Moffatt, K.; Cohen, R.; Mann, R. Purely Vision-Based Segmentation of Web Pages for Assistive Technology. *Comput. Vis. Image Underst.* **2016**, *148*, 46–66.
14. Ohashi, D.; Cohen, R.; Fu, X. The Current State of Online Social Networking for the Health Community: Where Trust Modeling Research May Be of Value. In Proceedings of the 2017 International Conference on Digital Health, Association for Computing Machinery, New York, NY, USA, 2–5 July 2017; pp. 23–32. [CrossRef]
15. Manjoo, F. *How Twitter Is Being Gamed to Feed Misinformation*; The New York Times: New York, NY, USA, 2017.
16. Press Association. *Twitter Directs Users to Government Information on Coronavirus with a Link to the Department of Health and Social Care That also Provides Official Updates about the Deadly Virus*; Press Association: London, UK, 2020.
17. Zadrozny, B.; Rosenblatt, K.; Collins, B. *Coronavirus Misinformation Surges, Fueled by Clout Chasers*; NBC News: New York, NY, USA, 2020.
18. Schultz, A. How Does Facebook Measure Fake Accounts? 2019. Available online: https://about.fb.com/news/2019/05/fake-accounts/ (accessed on 20 August 2020).
19. Facebook. What Types of ID Does Facebook Accept? Available online: https://www.facebook.com/help/159096464162185 (accessed on 20 August 2020).
20. Thomala, L.L. Number of Sina Weibo Users in China 2017–2021. 2019. Available online: https://www.statista.com/statistics/941456/china-number-of-sina-weibo-users/ (accessed on 19 February 2020).
21. Chokshid, N. *Older People Shared Fake News on Facebook More Than Others in 2016 Race, Study Says*; The New York Times: New York, NY, USA, 2019.
22. Guess, A.; Nagler, J.; Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **2019**, *5*, eaau4586. [CrossRef] [PubMed]
23. Wylie, L.E.; Patihis, L.; McCuller, L.; Davis, D.; Brank, E.; Loftus, E.F.; Bornstein, B. Misinformation Effect in Older versus Younger Adults: A Meta-Analysis and Review. In *The Elderly Eyewitness in Court*; Psychology Press: New York, NY, USA, 2014.
24. Bosak, K.; Park, S.H. Characteristics of Adults' Use of Facebook and the Potential Impact on Health Behavior: Secondary Data Analysis. *Interact. J. Med. Res.* **2018**, *7*, e11. [CrossRef] [PubMed]
25. Wood, S.; Lichtenberg, P.A. Financial Capacity and Financial Exploitation of Older Adults: Research Findings, Policy Recommendations and Clinical Implications. *Clin. Gerontol.* **2017**, *40*, 3–13. [CrossRef] [PubMed]
26. Jung, E.H.; Sundar, S.S. Senior citizens on Facebook: How do they interact and why? *Comput. Hum. Behav.* **2016**, *61*, 27–35. [CrossRef]
27. Moffatt, K.; David, J.; Baecker, R.M. Connecting Grandparents and Grandchildren. In *Connecting Families: The Impact of New Communication Technologies on Domestic Life*; Neustaedter, C., Harrison, S., Sellen, A., Eds.; Springer: London, UK, 2013; pp. 173–193. [CrossRef]
28. Yu, J.; Moffatt, K. Improving the Accessibility of Social Media for Older Adults. In Proceedings of the CSCW'19 Workshop on Addressing the Accessibility of Social Media, Austin, TX, USA, 9–13 November 2019; pp. 1–6.
29. John, O.P.; Donahue, E.M.; Kentle, R.L. *The Big Five Inventory: Versions 4a and 54*; Institute of Personality and Social Research, University of California: Berkeley, CA, USA, 1991.

30. Ghenai, A.; Mejova, Y. Fake Cures: User-Centric Modeling of Health Misinformation in Social Media. *Proc. ACM Hum. Comput. Interact.* **2018**, *2*, 1–20. [CrossRef]
31. Ghenai, A. Health Misinformation in Search and Social Media. Ph.D. Thesis, University of Waterloo, Waterloo, ON, Canada, 2019.
32. Zhang, J.; Cohen, R. Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. *Electron. Commer. Res. Appl.* **2008**, *7*, 330–340. [CrossRef]
33. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
34. Parmentier, A.; PNg, J.; Tan, W.; Cohen, R. Learning Reddit user reputation using graphical attention networks. In Proceedings of the Accepted to Future Technologies Conference 2020, Vancouver, BC, Canada, 5–6 November 2020; pp. 1–13.
35. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543. [CrossRef]
36. Chu, J.; McDonald, J. Helping the World Find Credible Information about Novel #Coronavirus. 2020. Available online: https://blog.twitter.com/en_us/topics/company/2020/authoritative-information-about-novel-coronavirus.html (accessed on 31 March 2020).
37. TwitterSafety. Content that Increases the Chance that Someone Contracts or Transmits the Virus, Including: Denial of Expert Guidance—Encouragement to Use Fake or Ineffective Treatments, Preventions, and Diagnostic Techniques—Misleading Content Purporting to be from Experts or Authorities. 2020. Available online: https://twitter.com/TwitterSafety/status/1240418440982040579 (accessed on 31 March 2020).
38. Davanipour, Z.; Sobel, E.; Ziogas, A.; Smoak, C.G.; Bohr, T.; Doram, K.; Liwnicz, B. Ocular Tonometry and Sporadic Creutzfeldt-Jakob Disease (sCJD): A Confirmatory Case-Control Study. *Br. J. Med. Med Res.* **2014**, *4*, 2322–2333. [CrossRef] [PubMed]
39. Abelson, M.B.; Lilyestrom, L. Mad Eye Disease: Should You Worry? *Rev. Ophthalmol.* **2008**, *15*. Available online: https://www.reviewofophthalmology.com/article/mad-eye-disease-should-you-worry (accessed on 6 September 2020).
40. Fernandez, M.; Alani, H. Online misinformation: Challenges and future directions. In Proceedings of the Companion Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 595–602.
41. Gupta, A.; Kumaraguru, P.; Castillo, C.; Meier, P. Tweetcred: Real-time credibility assessment of content on Twitter. In *International Conference on Social Informatics*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 228–243.
42. Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonça lves, B.; Patil, S.; Flammini, A.; Menczer, F. Truthy: Mapping the spread of astroturf in microblog streams. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 249–252.
43. Resnick, P.; Carton, S.; Park, S.; Shen, Y.; Zeffer, N. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In Proceedings of the Computational Journalism Conference, New York, NY, USA, 24 October 2014; Volume 5, p. 7.
44. Metaxas, P.T.; Finn, S.; Mustafaraj, E. Using twittertrails.com to investigate rumor propagation. In Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing, Vancouver, BC, Canada, 14–18 March 2015; pp. 69–72.
45. Agarwal, R.R.; Cohen, R.; Golab, L.; Tsang, A. Locating Influential Agents in Social Networks: Budget-Constrained Seed Set Selection. In Proceedings of the Advances in Artificial Intelligence—33rd Canadian Conference on Artificial Intelligence (Canadian AI 2020), Ottawa, ON, Canada, 13–15 May 2020. [CrossRef]
46. Johnson, N. *Simply Complexity: A Clear Guide to Complexity Theory*; Oneworld Publications: London, UK, 2009.
47. Watts, D.J. *Six Degrees: The Science of a Connected Age*; Norton: New York, NY, USA, 2007.

48. Menczer, F.; Fortunato, S.; Davis, C.A. *A First Course in Network Science*; Cambridge University Press: Cambridge, UK, 2020. [CrossRef]
49. Ciampaglia, G.L.; Mantzarlis, A.; Maus, G.; Menczer, F. Research Challenges of Digital Misinformation: Toward a Trustworthy Web. *AI Mag.* **2018**, *39*, 65–74. [CrossRef]