User Engagement with Digital Deception

Maria Glenski, Svitlana Volkova, and Srijan Kumar

Abstract Digital deception in online social networks, particularly the viral spread of misinformation and disinformation, is a critical concern at present. Online social networks are used as a means to spread digital deception within local, national, and global communities which has led to a renewed focus on the means of detection and defense. The audience (i.e., social media users) form the first line of defense in this process and it is of utmost importance to understand the who, how, and what of audience engagement. This will shed light on how to effectively use this wisdom-of-the-audience to provide an initial defense. In this chapter, we present the key findings of the recent studies in this area to explore user engagement with trustworthy information, misinformation, and disinformation framed around three key research questions (1) Who engages with mis- and dis-information?, (2) How quickly does the audience engage with mis- and dis-information?, and (3) What feedback do users provide? These patterns and insights can be leveraged to develop better strategies to improve media literacy and informed engagement with crowd-sourced information like social news.

Keywords Disinformation · Misinformation · User engagement

Social media platforms have gone beyond means of entertainment or social networking to become commonly used mechanisms for social news consumption. As the reliability or trustworthiness of media, news organizations, and other sources is increasingly debated, the society's reliance on social media as a primary source for news, opinion, and information has triggered renewed attention on the spread of misinformation. In particular, in these online communities with increased

M. Glenski · S. Volkova
Data and Analytics Group, National Security Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
e-mail: maria.glenski@pnnl.gov; svitlana.volkova@pnnl.gov

S. Kumar (✉)
Georgia Institute of Technology, Atlanta, GA, USA

K. Shu et al. (eds.), *Disinformation, Misinformation, and Fake News in Social Media*,
Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-030-42699-6_3

importance as a means of convenient and swift but potentially unreliable information acquisition – 68% of Americans report that they get at least some of their news from social media, however 57% of social media users who consume news on one or more of these platforms expect that the news they see to be “largely inaccurate” [28]. Most studies that investigate misinformation spread in social media focus on individual events and the role of the network structure in the spread [24, 25, 31, 47] or detection of false information [33]. Many related studies have focused on the language of misinformation in social media [18, 29, 32, 34, 43, 45] to detect types of deceptive news, compare the behavior of traditional and alternative media [37], or detect rumor-spreading users [33].

These studies have found that the size and shape of (mis) information cascades within a social network is heavily dependent on the initial reactions of the audience. Fewer studies have focused on understanding how users react to news sources of varying credibility and how their various response types contribute to the spread of (mis)information. An obvious, albeit challenging, way to do so is through the audience and their complex social behavior—the individuals and society as a whole who consume, disseminate, and act on the information they receive. However, there are a few challenges that complicate the task of using the audience as reliable signals of misinformation detection.

The first major challenge is that users have a *truth bias* wherein they tend to believe that others are telling the truth [14, 39]. Furthermore, research has found that being presented with brief snippets or clips of information (the style of information most commonly consumed on social media) exacerbates this bias [39]. Although this bias is reduced as individuals make successive judgements about veracity [39], social media users tend to make shallow, single engagements with content on social media [5, 6]. In fact, recent studies have found that 59% of bitly-URLs on Twitter are shared without ever being read [5] and 73% of Reddit posts were voted on without reading the linked article [6].

The second major challenge is that humans are not perfect in identifying false information when they come across it while browsing. Recent study by Kumar et al. [22, 23] showed that when people are in the reading mode, they can effectively detect false information. In particular, an experiment done with Wikipedia hoaxes showed that humans achieved 66% accuracy in distinguishing hoaxes from non-hoaxes, compared to 50% accuracy by random guessing. While they are better than random, humans make a mistake once out of every three attempts to detect false information, which can add error to the crowd-sourced human intelligence. The real strength lies in signals from many consumers at the same time, instead of a single individual.

The third major challenge is that users attempt to counterbalance their shallow engagement with *content* with a reliance on the crowd-provided commentary for information about the content and its credibility. When users do so, they rely on the assumption that these social media platforms are able to leverage the *wisdom of the crowd* to crowd-source reliable ratings, rankings, or other curation of information so that users don’t need to expend the cognitive resources to do so themselves for the deluge of information flooding their subreddits, timelines, or news feeds. However, other research has illustrated how the *hive mind* or *herd mentality* observed when individuals’ perceptions of quality or value follow the behavior of a group can be suboptimal for the group and individual members alike [1, 15, 27]. Studies have also found that user behavior (and thus, the content that is then shown to other users) can be easily influenced and manipulated through injections of artificial ratings [7, 8, 30, 46].

The audience (i.e., social media users) is effectively the “first line of defense” against the negative impacts and spread of misinformation or digital deception. It is important not only to understand how disinformation spreads or gains rapid traction (i.e., “goes viral”) and how to identify digital deception in a variety of forms but also how individuals and the audience in general currently react, engage, and amplify the reach of deception. These patterns and insights can be leveraged to better develop strategies to improve media literacy and informed engagement with crowd-sourced information like social news. In this chapter, we highlight several recent studies that focus on the human element (the audience) of the (mis) and (dis)information ecosystem and news cycle.

As reliance on social media as a source of news remains consistently high and the reliability of news sources is increasingly debated, it is important to understand not only what (mis) and (dis)information is produced, how to identify digital deception at coarse and fine granularities, and which algorithmic or network characteristics enable its spread how, but also how users (human and automated alike) consume and contribute to the (mis) and (dis)information cycle. For example, how do users react to news sources of varied levels of credibility and what commentary or kinds of reactions are presented to other users?

In this chapter, we highlight key findings from the studies summarized in Table 1 framed around three key research questions:

RQ1: *Who* engages with (mis) and (dis)information?,
RQ2: *What* kind of feedback do users provide?, and
RQ3: *How quickly* do users engage with (mis) and (dis)information?

in Sects. 2, 3, and 4, respectively. Before we explore these research questions, we first present an overview of the *Methods and Materials* used in Sect. 1.

Table 1 Studies highlighted in this chapter and the sections that reference each study

Reference	Title	Sections
[11]	Propagation from deceptive news sources: who shares, how much, how evenly, and how quickly?	2, 4
[10]	Identifying and understanding user reactions to deceptive and trusted social news sources	3, 4
[9]	How humans versus bots react to deceptive and trusted news sources: A case study of active users	2, 3, 4

1 Methods and Materials

As we noted above, most studies that examine digital deception spread focus on individual events such as natural disasters [40], political elections [4], or crises events [38] and examine the response to the event on specific social platforms. In contrast, the studies highlighted in this chapter consider users' engagement patterns across news sources identified as spreading trustworthy information versus disinformation – highlighting distinctions in audience composition or engagement patterns that can be leveraged for robust defense against (mis) and (dis)information, educational strategies to mitigate the continued spread or negative impacts of digital deception, and more. Before we highlight key findings, we present an overview of the processes used in the studies that will be referenced in the following sections.

1.1 Attributing News Sources

In several of the studies highlighted in the following section [9–11], credibility annotations partition news sources into (1) fine-grained or (2) coarse labeled sets based on the hierarchy of types of information spread in Fig. 1. Fine-grained labeled news sources are partitioned into five classes of news media. That is, news sources identified as a:

- *trustworthy* news source that provided factual information with no intent to deceive;

or one of several classes of *deceptive* news sources:

- *clickbait*: attention-grabbing, misleading, or vague headlines to attract an audience;

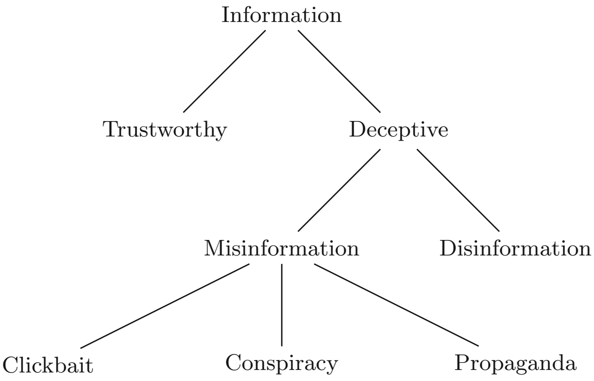

Fig. 1 Hierarchy of information, misinformation, and disinformation used in news source annotations

- *conspiracy theories*: uncorroborated or unreliable information to explain events or circumstances;
- *propaganda*: intentionally misleading information to advance a social or political agenda; or
- *disinformation*: fabricated and factually incorrect information spread with an intention to deceive the audience.

Coarse-grained labeled sets build off of these fine-grained annotations to look at more abstract groupings of:

- *trustworthy* news sources that provide factual information with no intent to deceive;
- *misinformation* news sources identified as spreading clickbait, conspiracy theories, and propaganda; or
- *misinformation* + *disinformation* news sources identified as spreading clickbait, conspiracy theories, propaganda, *and intentional disinformation*.

New sources identified as spreading disinformation were collected from EUvsDisinfo.eu[1] while all others were obtained from a list compiled by Volkova et al. [43] through a combination of crowd-sourcing and public resources.[2] As of November 2016, EUvsDisinfo reports included almost 1,992 confirmed disinformation campaigns found in news reports from around Europe and beyond.

1.2 Inferring User Account Types: Automated Versus Manual

Classification of user accounts as manually run by individuals (i.e. human) or account that is automated (i.e. bot) is done through thresholding of botometer scores. Botometer scores [2] indicate the likelihood of a user account being an automated bot account and are collected for a given user. The label of 'bot' is then assigned if the score is at or above the bot threshold of 0.5, otherwise the label of 'human' is assigned to the user.

1.3 Predicting User Demographics

To infer gender, age, income, and education demographics of users identified to be individual, manually-run accounts, Glenski et al. [11] employed a neural network model trained on a large, previously annotated Twitter dataset [42]. Following

[1] News sources collected from EUvsDisinfor.eu were identified as spreaders of *disinformation* by the European Union's East Strategic Communications Task Force.

[2] Example resources used by Volkova et al [43] to compile deceptive news sources: http://www.fakenewswatch.com/, http://www.propornot.com/p/the-list.html.

previous methodology [42], each demographic attribute was assigned one of two mutually exclusive classes. Gender was classified as either male (M) or female (F), age as either younger than 25 (Y) or 25 and older (O), income as below (B) or at and above (A) $35,000 a year, and education as having only a high school education (H) or at least some college education (C).[3]

1.4 Measuring Inequality of User Engagement

In order to measure the inequality of engagement with trustworthy information versus deceptive news, we leverage three measures commonly used to measure income inequality: Lorenz curves, Gini coefficients, and Palma ratios. Rather than measuring how shares of a region, nation, or other population's income is spread across the individuals within the population, these metrics can be adapted to quantify and illustrate how interactions or the volume of engagement is spread across the population of users who engage with (mis) and (dis)information. This allows us to compare inequality of engagement with news sources across types of information (trustworthy news, conspiracy, disinformation etc.) in a approach to the way economists compare income inequality across countries.

Lorenz curves (an example of which is illustrated in Fig. 2) are often used as a graphical representation of income or wealth distributions [17]. In those domains, the curves plot the cumulative percentage of wealth, income, or some other variable

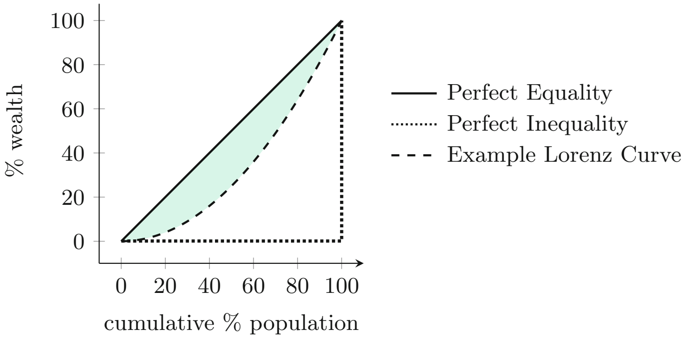

Fig. 2 Lorenz curves illustrate inequality within a frequency distribution graphically by plotting the cumulative share (from least to greatest) of the variable under comparison (e.g. income or wealth) as a function of the population under consideration. The proportion of the area under the diagonal (representing perfect equality) that is captured above the Lorenz curve represents how far the population pulls from perfect equality and is called the Gini Coefficient

[3] The area under the ROC curve (AUC) for 10-fold cross-validation experiments were 0.89 for gender, 0.72 for age, 0.72 for income, and 0.76 for education.

to be compared against the cumulative (in increasing shares) percentage of a corresponding population. The inequality present is illustrated by the degree to which the curve deviates from the straight diagonal ($y = x$) representative of perfect equality. There are two metrics that summarize Lorenz curves as a single statistic: (1) the Gini coefficient is defined as the proportion of the area under the line of perfect equality that is captured above the Lorenz curve, and (2) the Palma ratio, defined as the ratio of the share of the top 10% to the bottom 40% of users in the population.

Again, using wealth inequality as an example, if each individual in the population had a equal amount of wealth (perfect equality), the Lorenz curve would fall along the diagonal in Fig. 2, the Gini coefficient would be 0, and the Palma ratio would be 0.25 (10/40). Paired together, the Gini coefficient and Palma ratio provide a balanced understanding of the degree to which a Lorenz curve deviates from perfect equality. Gini coefficients are most sensitive to changes within the mid-range of the lorenz curve while the Palma is more sensitive to changes at the extremes.

1.5 Predicting User Reactions to Deceptive News

Discourse acts, or speech acts, can be used to identify the *use* of language within a conversation, e.g. agreement, question, or answer. In these studies, user reactions are classified as one of eight types of discourse acts analyzed in the context of social media discussions in previous work by Zhang et al.[49]: agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, or question, or as none of the given labels, denoted“other”, using linguistically-infused neural network models [10].

1.6 Data

News Propagation and Influence from Deceptive Sources [11] Two datasets are used in this study. First, 11 million direct interactions (i.e. retweet and @mention) by almost 2 million Twitter users' who engaged with a set of 282 credibility-annotated news sources (using the approach described above) from January 2016 through January 2017. Second, a subset of these interactions for the 66,171 users who met all three of the following requirements: actively engaged (at least five times) with deceptive news sources, were identified as individual user accounts, and met the activity threshold of predictive models used to infer gender, age, income, and education demographics of the users. This dataset uses the fine-grained classifications of news sources identified as spreading trustworthy news, clickbait, conspiracy, propaganda, or disinformation.

Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources [10] User reactions to news sources were inferred for two popular platforms, Reddit and Twitter. The Reddit dataset comprises all Reddit posts submitted during the 13 month period from January 2016 through January 2017 that linked to domains associated with a previously annotated set of trustworthy versus deceptive news sources [11, 43] and the immediate comments (i.e. that directly responded to one of the posts). The Twitter dataset contains all tweets posted in the same 13 month period that directly @mentioned or retweeted content from one of these source's Twitter accounts. Coarse-grained news source classifications sets are used in this study: trustworthy, deceptive, and misinformation and disinformation.

How Humans versus Bots React to Deceptive and Trusted News Sources: A Case Study of Active Users [9] The dataset used in this study comprises a 431,771 tweets sample identified as English-content in the Twitter metadata of tweets posted between January 2016 and January 2017 that @mentioned or retweeted content from one of the annotated news sources (described above) also used for cross-platform and demographics-based engagement studies [10, 11]. This study focused on users who frequently interacted (at least five times) with deceptive news sources and considered fine-grained classifications of news sources. Each tweet was assigned a reaction type and user account type (bot or human) using the annotation processes described above – inferred via linguistically infused models [10] or based on botometer scores of users who authored each post [2].

2 *Who* Engages with (mis) and (dis)information?

Some studies model misinformation or rumor diffusion as belief exchange caused by influence from a users network, ego-network, or friends, e.g. the Tipping Model [35] and several previous studies have investigated the characteristics of users that spread or promote information as a way to identify those who spread rumors or disinformation [33]. For example, a 2015 study by Wu et al. [47] highlighted the type of user who shared content as one of their most important features in predictive models that were able to detect false rumors on Weibo with 90% confidence as quickly as 24 hours after the content was initially broadcast on the social network. Ferrara [3] found that users with high followings generated highly-infectious cascades for propaganda information. Recent work has also found that accounts spreading disinformation are significantly more likely to be automated accounts [36].

In this section, we focus on *who* engages with misinformation and disinformation and highlight key findings from several recent studies [9, 11, 21] related to user engagement with news sources categorized using the fine-grained classifications of: Trustworthy, Clickbait, Conspiracy, Propaganda, and Disinformation.

2.1 The Population Who Engage with Misinformation and Disinformation

Studies have identified that when an individual believes in one conspiracy theory, that individual is also likely to believe in others [12, 26]. At an aggregate level, one can consider whether this pattern might also hold for propagation or engagement with disinformation online – if a user engages or spreads mis and disinformation once, are they likely to engage again? if a user engages with news sources who publish one kind of deceptive content (e.g. clickbait), are they also likely to engage with another (e.g. intentional disinformation)? When investigated as a population as a whole, Glenski et al. [11] found that there were overlaps between populations of users engaging with news sources of varied degree of deception (illustrated in Fig. 3) but that the increased likelihood of sharing another type of deceptive news given that you engaged with another was not always reciprocal. For example, users who engage with news sources who spread clickbait and conspiracy theories are likely to also engage propaganda sources, but not the other way around.

Figure 4 highlights the degree to which engagement with news sources is evenly spread (or not) across the population who engage with news sources spreading trustworthy information versus mis- or disinformation. Unsurprisingly, disinformation sources are most highly retweeted from a small group of users that actively engage with those sources regularly. Effectively, a disproportionate amount of the engagement, promotion, or propagation of content published by news sources who were identified as spreading intentional disinformation from a subset of highly active, vocal users. Propaganda is the next most unevenly engaged with news, followed by trustworthy news, conspiracy, and clickbait.

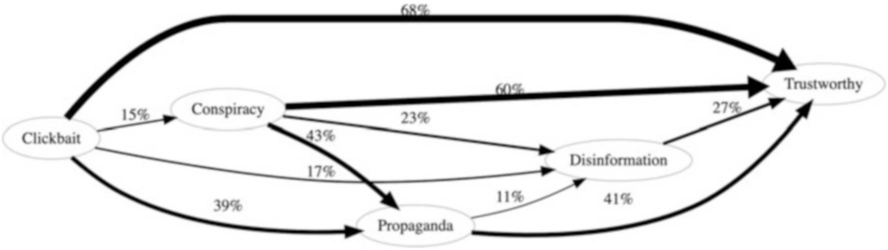

Fig. 3 Overlaps of users who engage with news sources across the spectrum of credibility as a directed graph for overlaps of at least 10% of users. Edges illustrate the tendency of users who engage with a news source spreading one type of deceptive content to also engage with a news source spreading another type of deceptive content. For example, the edge from Clickbait to Trustworthy illustrates that 68% of users who engage with news sources that spread clickbait, also engage with trustworthy news. Note: in total, 1.4 M users engaged with Trustworthy news sources, 19 k with Clickbait, 35.8 k with Conspiracy, 233.8 k with Propaganda, and 292.4 k with Disinformation

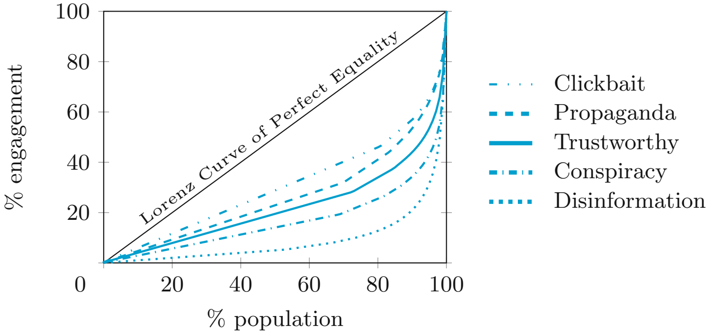

Fig. 4 Lorenz curves for inequality in engagement with news sources identified as spreading trustworthy news, clickbait, conspiracy theories, propaganda, and intentional disinformation. Legend (at right) is ordered from closest to furthest from the diagonal (representing engagement that is equally distributed across the population)

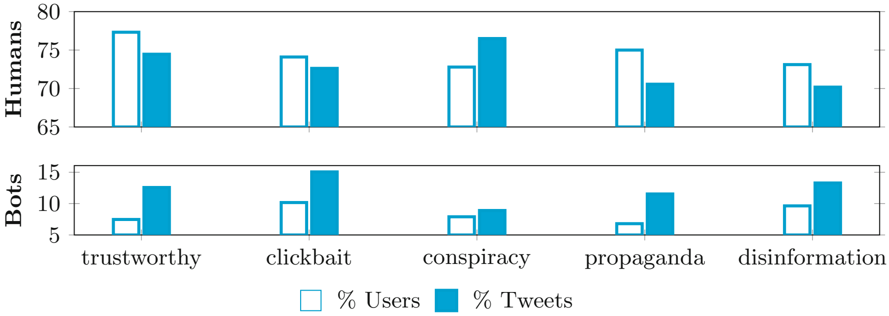

Fig. 5 Prevalence of humans (above) and bots (below) within tweets responding to news sources of varied credibility (% tweets, as solid bars) and within the populations of user accounts who authored the response-tweets (% users, as white-filled bars)

2.2 Automated Versus Manual Accounts

Glenski et al. [9] found that automated bot user accounts were responsible for approximately 9–15% of the direct responses to news sources spreading (mis) and (dis)information across all five fine-grained classifications of trustworthy news sources and news sources identified as spreading misinformation – clickbait, conspiracy, or propaganda – and intentional disinformation but only comprise around 7–10% of the users responsible for those response-tweets. Figure 5 illustrates the prevalence of automated (i.e. bots) versus manually-run (i.e. human) accounts among users who react to news sources spreading (dis)information within each category and the responses themselves (i.e. the percentage of tweets authored by user accounts inferred as automated versus manually run accounts).

Although news sources who spread conspiracy have the lowest presence of human users (72.8% of user accounts who authored reaction tweets), they have a disproportionately high proportion of reactions tweets authored by human-users (76.5% of tweets)—the highest proportion of human-authored reaction tweets across all five classes of news sources including the Trustworthy news sources which have the highest relative presence of human users. Interestingly, clickbait sources have the highest presence of bots with 10.17% of users identified as bots (who were responsible for 15.06% of the reaction-tweets) while news sources who spread disinformation have the second highest proportions of bots for users who reacted as well as reaction tweets posted.

2.3 Sockpuppets: Multiple Accounts for Deception

While bots are effective in spreading deceptive information at a fast speed and a large scale, the technology is not advanced enough to make their conversations and behavior believable as humans. This makes them barely effective in one-on-one conversations. Thus, bad actors adopt a smart strategy to deceive the audience: they create multiple accounts and operate them simultaneously to converse with the audience [21]. Kumar et al. showed that puppetmasters typically operate two or more 'sockpuppet' accounts, with the primary goal of deceiving others. These sockpuppet accounts typically support one another and create an illusion of magnified consensus. However, sometimes their strategies are more complex—instead of overtly supporting one another, some sockpuppet accounts oppose one another to create an illusion of argument. This attracts more attention and gets the audience involved as well. These crafty arguments are eventually used to influence people's opinions and deceive them.

Thus, the complex deceptive ecosystem created by the sockpuppets leads to increased attention to and the spread of false information and propaganda.

2.4 Demographic Sub-populations

When considering only the user accounts that frequently interacted with deceptive news sources on Twitter and the users' inferred demographics [11], the population was found to be primarily predicted to be male (96%), older (95%), with higher incomes (81%), college-educated (82%), and classified as "regular users" who followed more accounts than they had followers (59%), illustrated in Fig. 6. Although intuitively, this sample would not be expected to be a representative sample of Twitter users overall, the sample's majority demographic aligned with that found in a Pew Resarch Center survey conducted during the time period covered by the study – the Pew Research center survey found that 17% of Twitter users had a high school education or less, 38% were between 18 and 29 years old, and 47% were male [13] – although the study's sample was more heavily skewed towards the majority demographic than the Pew Research Center's findings.

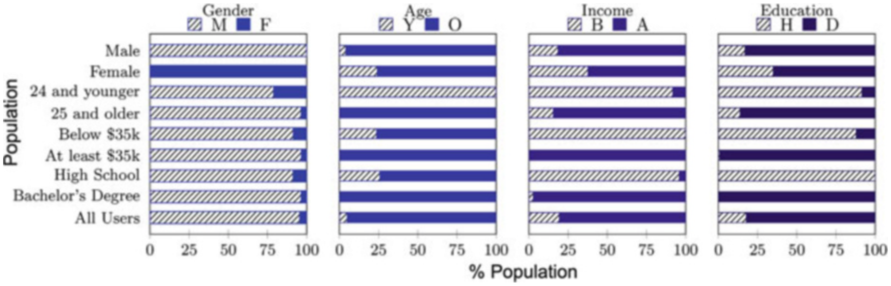

Fig. 6 Inferred demographics of users who frequently engage with deceptive news sources on Twitter [11]

Table 2 Inequality in user authorship of feedback to news sources who spread information (trustworthy news), misinformation (clickbait, conspiracy, and propaganda), and disinformation. Illustrated using the Palma Ratio – the ratio of feedback posted by the top 10% most active users to the 40% least active

		Trustworthy	Clickbait	Conspiracy	Propaganda	Disinformation
Gender	Male	4.47	1.72	6.58	3.58	25.06
	Female	3.08	1.40	1.44	3.04	18.03
Age	≤ 24	2.18	1.37	2.28	2.91	41.34
	≥ 25	4.48	1.72	6.58	3.58	18.15
Income	$< \$35k$	4.67	2.38	8.75	4.41	35.49
	$\geq \$35k$	4.35	1.66	6.00	3.39	8.99
Education	High school	4.09	2.19	8.88	4.17	35.52
	College	4.49	1.68	6.02	3.45	10.66
User Role	Follow	4.80	1.79	7.73	3.74	28.26
	Lead	4.00	1.65	5.37	3.42	21.29

There are significant differences in how equally users contribute to the feedback provided to news sources who spread information, misinformation, and disinformation online. We highlight the inequality in authorship of feedback (the extent to which a subset of highly active users contribute a disproportionate amount of the feedback within sub-populations by demographic) from Glenski et al. [11] in Table 2. Of note, the largest disparity in participation of feedback is in the sub-population of users inferred to be 24 years old or younger – the most active 10% of these users which provide feedback to disinformation sources via retweets or mentions author 41.34 as much as the least active 40% of users within this subpopulation. In contrast, the older sub-population (≥25 years old) have a much smaller palma ratio of 18.15. Overall, there is much greater inequality in participation of users who respond to news sources identified as spreading disinformation. Interestingly, the set of news sources which elicit the closest to uniform participation from responding users is clickbait, the least deceptive of the news sources who spread misinformation, rather than news sources identified as spreading trustworthy *information*.

3 *What* Kind of Feedback Do Users Provide?

In this section focusing on *what* kind of feedback users provide to news sources who spread information, misinformation, and disinformation, we highlight key findings from two recent studies [9, 10] related to the kinds of reactions (asking questions; expressing agreement, disagreement, or appreciation; providing answers; etc.) users post in response to social media news sources categorized using both the coarse-grained classifications of: Trustworthy, Deceptive, or Deceptive+Disinformation news sources [10] across two popular, and very different, social media platforms (Twitter and Reddit) and fine-grained classifications of: Trustworthy, Clickbait, Conspiracy, Propaganda, and intentional Disinformation [9] across user account characteristics (whether account is automated—i.e. a bot—or manually run).

3.1 Across Multiple Platforms

Glenski et al. [10] found that the predominant kinds of feedback elicited by any type of news source—from trustworthy sources sharing factual information without an intent to deceive the audience to deceptive news sources who spread intentional disinformation—across both Twitter and Reddit were answers, expressions of appreciation, elaboration on content posted by the news source, and questions. Figure 7 illustrates the distribution of these types of feedback, denoted reaction types, among Reddit comments (top plot) or tweets (bottom plot) responding to each category of news source (using the coarse classification as trustworthy versus deceptive or deceptive + disinformation) as a percentage of all comments/tweets

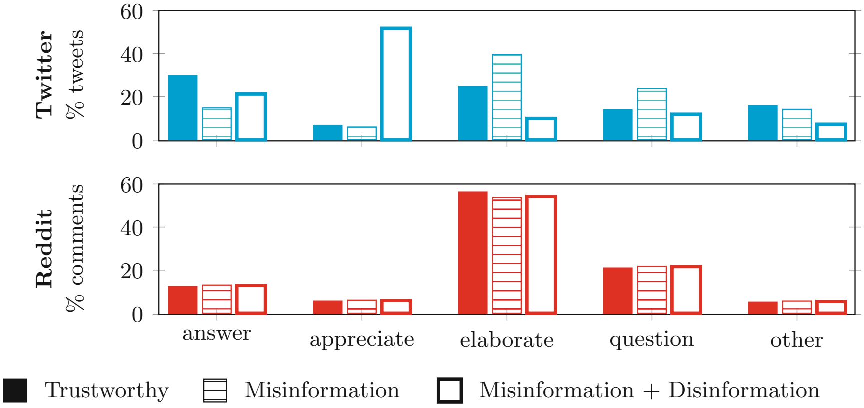

Fig. 7 Distributions of the five most frequently occurring ways in which users engage with news sources on Twitter (above) and Reddit (below) for coarse-grained partitions of trustworthy, misinformation, and misinformation + disinformation spreading news sources

reacting to sources of the given type (i.e. trusted, all deceptive, and deceptive excluding disinformation sources).

There were clear differences in the kinds of feedback posed to news sources on Twitter. As shown by the misinformation + disinformation bars, the misinformation news sources, when including disinformation-spreading sources, have a much higher rate of appreciation reactions and a lower rate of elaboration responses, compared to trustworthy news sources. Feedback from users towards disinformation spreading news sources are more likely to offer expressions of appreciation than elaboration. Differences are still significant ($p < 0.01$) but the trends reverse when the set of misinformation news sources do not include those that spread disinformation (including only those that spread clickbait, conspiracy, and propaganda). There is also an increase in the rate of question-reactions compared to trustworthy news sources when disinformation-spreading news sources are excluded from the set of deceptive news sources.

Feedback provided via user engagement on Reddit appears to follow a very similar distribution across different types of feedback for trustworthy versus misinformation/disinformation sources. However, Mann-Whitney U tests on Reddit-based user engagement still found that the illustrated differences between trusted and misinformation + disinformation news sources were statistically significant ($p < 0.01$)—regardless of whether we include or exclude disinformation sources. Posts that link to misinformation + disinformation sources have higher rates of expressions of appreciation and posing or answering questions while posts that link to trustworthy sources have higher relative rates of providing additional information or details via elaborations, expressions of agreement, and expressions of disagreement.

3.2 Across User-Account Characteristics

When the distributions of each class are compared, we find several key differences in what kind of feedback (i.e. reaction indicated from the primary discourse act of user response) is elicited. Conspiracy news sources have the highest relative rate of elaboration responses, i.e. *"On the next day, radiation level has gone up. [url]"* – with a more pronounced difference within the bot population – and the lowest relative rate of feedback in the manner of providing answers within the bot population but not within manually run accounts (i.e. human users). Clickbait news sources, on the other hand, have the highest relative rate of feedback where users provide answers and the lowest rate of where users pose questions across both populations of user account types (Fig. 8).

Conspiracy and propaganda news sources have higher rates within the population of manual accounts of accounts raising questions in response to the news sources than providing answers; manually run "human" accounts who respond to these types

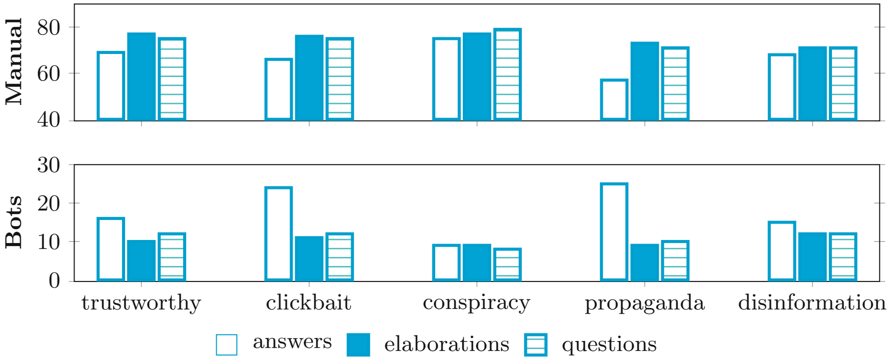

Fig. 8 Percentages of feedback of a given type (i.e. answers, elaborations, or questions) that were posted by manually run individual (above) and automated bot (below) user-accounts for each of the fine-grained classifications of news sources: those who spread trustworthy news, clickbait, conspiracy, propaganda, or disinformation

of news sources question the content posted by the source more often than they provide answers in response to a news source's posting. When reactions authored by bot-accounts are examined, there is a similar trend for conspiracy sources but a higher relative rate of answer reactions than question reactions to propaganda sources.

4 *How Quickly* Do Users Engage with (mis) and (dis)information?

Information diffusion studies have often used epidemiological models, originally formulated to model the spread of disease within a population, in the context of social media [16, 41, 48]. For example, Tambuscio et al. [41] used such a model to determine a threshold of fact-checkers needed to eliminate a hoax. In this context, users are *infected* when they spread information to other users. A recent study by Vosoughi et al. [44] found that news that was fact-checked (post-hoc) and found to be false had spread faster and to more people than news items that were fact-checked and found to be true. In this section, we highlight key findings on the speed at which users react to content posted by news sources of varying credibility and comparative analyses of the delays of different types of responses. By contrasting the speed of reactions of different types, from different types of users (bot and human), and in response to sources of varying credibility, one is able to determine whether deceptive or trusted *sources* have slower immediate share-times overall or within combinations of classes of user account or news sources.

4.1 Across Multiple Platforms

In [10], Glenski et al. examine the speed and volume of user engagement with social news using coarse-grained partitioning of sources as trustworthy or deceptive (e.g. news sources that spread a variety of disinformation). A key finding was the differences in the pace and longitude of engagement with the same deceptive news sources across differing social platforms: Twitter and Reddit. The duration of engagement with content across trustworthy and deceptive news sources alike was found to be typically more prolonged for engagement with information spread on Twitter compared to Reddit. Intuitively, this could be due to the different manner in which users engage with content in general when using one platform versus another. Users are able to pinpoint specific *users* (or news source accounts) to follow, regularly consume content from, or easily engage with on Twitter whereas users "follow" topics, areas of interest, or communities of users through the Reddit mechanism of subscribing to subreddits. While news sources have content spreading across both, there is a greater difficulty to consistently engage with a single news sources content over time on Reddit.

Cumulative density function plots for three means of engagement are illustrated in Fig. 9 for the sets of trustworthy, misinformation, and misinformation +

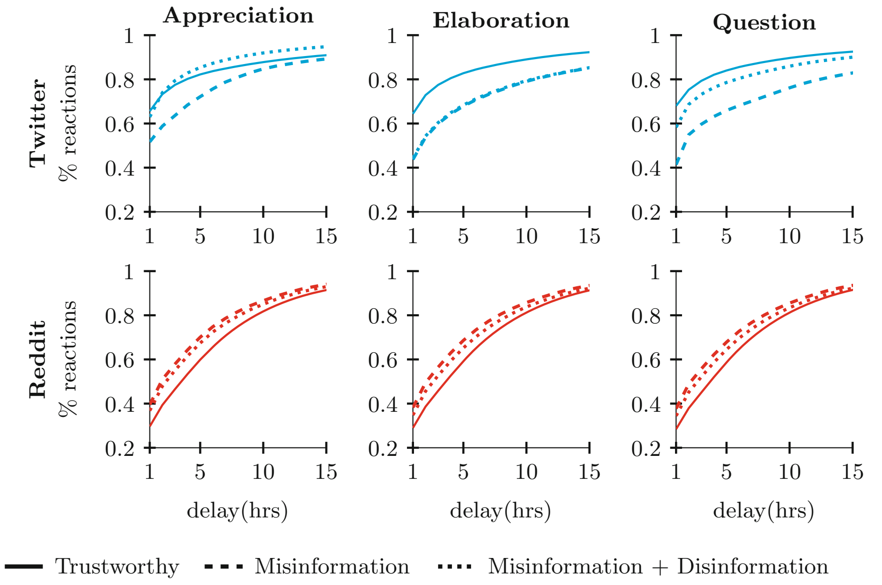

Fig. 9 Cumulative density function plots for three means of engagement (where users express *appreciation* towards the news source, *elaborate* on content published by the news source, or *question* the content published by a news source) for the sets of trustworthy, misinformation, and misinformation + disinformation news sources when users engaged via the Twitter (above) and Reddit (below) platforms

disinformation spreading news sources when users engaged via the Twitter and Reddit platforms. In addition to the differences in scale of duration of engagement, *engagement with trustworthy social news sources are less heavily concentrated within the first 12 to 15 h after content is initially published by the social news source on Reddit whereas the opposite is found on Twitter. While Twitter social news sources may have a larger range of delays before a user engages with content, they are also more heavily skewed with larger concentrations immediately following a source's publication of content (*$p < 0.01$*).*

If delays in providing feedback are examined using more fine-grained classifications of misinformation [11], the populations of users who provide feedback to news sources on Twitter to news sources who spread trustworthy information, conspiracy, and disinformation news have similarly short delays between the time when a news source posts new content on Twitter and when a user provides feedback via @mentioning or retweeting the news source. However, delays are significantly longer for news sources identified as spreading clickbait and propaganda misinformation ($p < 0.01$).

4.2 Across User-Account Characteristics

Next, we highlight the speed with which bot and human users react to news sources [9]. As would be expected, this study found that response activity is heavily concentrated in the window of time soon after a news source posts when considering any combination of type of information being spread or feedback being provided. Mann Whitney U tests that compared distributions of response delays found that manually-run accounts will pose questions and provide elaborations of information posted by news sources those that spread clickbait faster than automated bot accounts do ($p < 0.01$); There is a heavier concentration (at least 80%) of reactions from manually-run accounts that have response delays with at most a 6 h delay compared to automated bot accounts that have approximately 60–70% of their elaboration and question based responses falling within that initial 6 h window, shown in Fig. 10.

There are similar trends for all the other combinations of feedback provided to and type of information spread by news sources with a few notable exceptions: (1) automated bot accounts provide answer-responses to news sources identified as spreading propaganda content with significantly shorter delays than manually-run accounts ($p < 0.01$) and (2) MWU tests comparing sub-populations of automated and manual accounts authoring feedback providing answers to news sources who spread either clickbait or disinformation were not found to differ with statistical significance.

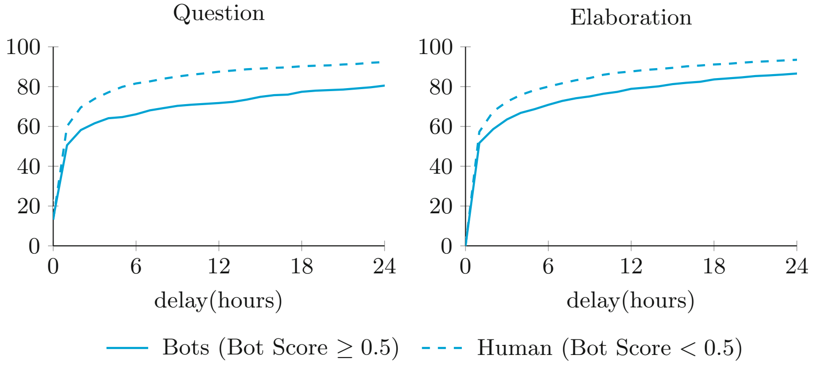

Fig. 10 Cumulative distribution function (CDF) plots of the volume by the delay between when a news source identified as spreading clickbait posted content and when a response that posed a question (left) or elaboration (right) was posted for bots and human user accounts, using a step size of one day

Table 3 Users by demographics who respond faster to news sources identified as spreading *trustworthy* information, several classes of misinformation (*clickbait, conspiracy, propaganda*), and *disinformation*. A dash (—) indicates no significant differences were found between sub-populations for a given demographic with all other results being statistically significant (MWU $p < 0.01$)

	Trustworthy	Clickbait	Conspiracy	Propaganda	Disinformation
Gender	Male	—	Male	Male	Male
Age	≥ 25	—	≤ 24	≤ 24	≤ 24
Income	$< 35k$	$< 35k$	$< 35k$	$< 35k$	$< 35k$
Education	High school	High school	High school	High school	High school
Role	Leader	Follower	Follower	Leader	Follower

4.3 Demographic Sub-populations

In Table 3, we highlight the speed of response comparisons by demographic sub-population from [11]. Older users were found to retweet news sources identified as spreading trustworthy news more quickly than their younger counterparts but slower to share all other types – that is, younger users ($\leq$ 24 years old) engage with the deceptive news sources (misinformation and disinformation spreading alike) more quickly, sooner after the news source is posting content. Users inferred to have only high school education engage faster than those with a college education across the board. Except for comparisons between predicted gender or age brackets for clickbait sources, there are statistically significant differences in delays for all information type and demographic combinations.

5 Discussion and Conclusions

In this chapter, we have highlighted key findings of several recent studies that examined the human element of the digital deception ecosystem and news cycles—the audience who engage with, spread, and consume the misinformation and disinformation present on the online social platforms that society has come to rely on for quick and convenient consumption of information, opinion, and news. Framed to answer each of our three key research questions, we have presented the key findings of several recent studies and how the results pair together to present a comprehensive understanding of user engagement with multiple scales or resolutions of deception (from coarse to fine-grained credibility annotations).

However, in each of the studies referenced above, we have analyzed audience engagement from the position of knowing whether news sources or content is deceptive or trustworthy. Often, if not always, individual users are not given such clear labels of deception versus not. Rather, they are faced with the opposite, where deceptive and trustworthy content and news sources alike portray themselves as trustworthy. A key premise of studying the audience reaction to misinformation and disinformation is that they should be able identify false information when they come across it in social media. But how effective are readers in identifying false information? To answer this question, Kumar et al. [23] conducted a human experiment using hoax articles on Wikipedia as disinformation pieces and non-hoax articles as non-deceptive pieces.

Hoax articles on Wikipedia contain completely fabricated information and are created with the intention of deceiving others. Identifying them on Wikipedia requires a meticulously manual process that guarantees the ground truth. In a human experiment, Kumar et al. showed a pair of articles to Mechanical Turk workers—one article was a hoax article and another was a non-hoax article—and the workers were told to identify the hoax article. In this scenario, random guess would yield a 50% accuracy while the workers got the answer correct 66% of the times. This shows that humans are able to identify false information better than random though they are not perfect. Analysis of their mistakes showed that well-formatted, long, and well-referenced hoax articles fooled humans into thinking it is true. This shows that humans *can* be able to identify false information when they come across it, as shown in this setting of Wikipedia content. However, the real power comes when leveraging feedback at a large scale from a sizeable audience in social media.

Similarly, Karduni et al. [20] conducted human experiments to study user decision-making processes around misinformation on Twitter and how uncertainty and confirmation bias (the tendency to ignore contradicting information) affect users decision-making. The authors developed visual analytic system – Verifi[4] designed provide users with the ability to characterize and distinguish misinformation from legitimate news. Verifi explicitly presents a user with the cues to make decisions

[4] https://verifi.herokuapp.com/

about the veracity of news media sources on Twitter including account-level temporal trends, social network and linguistic features e.g., biased language, subjectivity, emotions etc. The authors then used Verifi to measure how users assess the veracity of the news media accounts on Twitter (focusing on textual content rather than images) and what role confirmation bias plays in this process. Their analysis shows that certain cues significantly affected users decisions about the veracity of news sources more than others, for example specific named entities, fear and negative language and opinionated language. However, similar to Kumar et al. study, user accuracy rate ranges between 54% and 74% depending on different experimental conditions.

Verifi2 [19], a visual analytic system that enables users to explore news in an informed way by presenting a variety of factors that contribute to its veracity. It allows to contrast (1) language used by real and suspicious news sources, (2) understand the relationship between different news sources, (3) understand top names entities, and (4) compare real vs. suspicious news sources preferences on images. The authors conduct interviews with experts in digital media, communications, education, and psychology who study misinformation in order to help real users make decisions about misinformation in real-world scenarios. All of their interviewees acknowledged the challenge in defining misinformation, as well as the complexity of the issue which involves both news outlets with different intents, as well as audiences with different biases. Finally, Verifi2 expert users suggested to define a spectrum of trustworthiness rather than binary classes (real vs. suspicious news sources), and identified the potentials for Verifi2 to be used in scenarios where experts educate individuals about differences between real and suspicious sources of news.

A well-rounded understanding of existing patterns, trends, and tendencies of user engagement is a necessary basis for the development of effective strategies to defend against the evolving threat of digital deception. Key findings highlighted here in the context of multiple studies at varied resolutions of credibility of information or sources, user account characteristics, and social platforms under consideration can be used to inform models and simulations of (dis)information spread within and across communities of users, social platforms, geolocations, languages, and types of content. Further, they can be used to advise direct interventions with individuals or groups of users to improve their manual detection skills. Some open challenges include how to effectively combine feedback from large audience in real-time and how to improve detection of complex multimedia disinformation using audience feedback.

Acknowledgements This work was supported in part by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. This research was also supported by the Defense Advanced Research Projects Agency (DARPA), contract W911NF-17-C-0094. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. This work has also been supported in part by Adobe Faculty Research Award , Microsoft, IDEaS, and Georgia Institute of Technology.

References

1. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. **100**(5), 992–1026 (1992). https://doi.org/10.2307/2138632, http://www.jstor.org/stable/2138632
2. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: a system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274. International World Wide Web Conferences Steering Committee (2016)
3. Ferrara, E.: Contagion dynamics of extremist propaganda in social networks. Inf. Sci. **418**, 1–12 (2017)
4. Ferrara, E.: Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday **22**(8) (2017). https://doi.org/10.5210/fm.v22i8.8005, http://journals.uic.edu/ojs/index.php/fm/article/view/8005
5. Gabielkov, M., Ramachandran, A., Chaintreau, A., Legout, A.: Social clicks: what and who gets read on twitter? ACM SIGMETRICS Perform. Eval. Rev. **44**(1), 179–192 (2016)
6. Glenski, M., Pennycuff, C., Weninger, T.: Consumers and curators: browsing and voting patterns on reddit. IEEE Trans. Comput. Soc. Syst. **4**(4), 196–206 (2017)
7. Glenski, M., Weninger, T.: Predicting user-interactions on reddit. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE/ACM (2017)
8. Glenski, M., Weninger, T.: Rating effects on social news posts and comments. ACM Trans. Intell. Syst. Technol. (TIST) **8**(6), 1–9 (2017)
9. Glenski, M., Weninger, T., Volkova, S.: How humans versus bots react to deceptive and trusted news sources: a case study of active users. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE/ACM (2018)
10. Glenski, M., Weninger, T., Volkova, S.: Identifying and understanding user reactions to deceptive and trusted social news sources. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 176–181 (2018)
11. Glenski, M., Weninger, T., Volkova, S.: Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? IEEE Trans. Comput. Soc. Syst. **5**(4), 1071–1082 (2018)
12. Goertzel, T.: Belief in conspiracy theories. Polit. Psychol. **15**(4), 731–742 (1994)
13. Gottfried, J., Shearer, E.: News use across social media platforms 2016. Pew Research Center (2016). http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/
14. Hasson, U., Simmons, J.P., Todorov, A.: Believe it or not: on the possibility of suspending belief. Psychol. Sci. **16**(7), 566–571 (2005)
15. Hirshleifer, D.A.: The blind leading the blind: social influence, fads and informational cascades. In: Ierulli, K., Tommasi, M. (eds.) The New Economics of Human Behaviour, chap 12, pp. 188–215. Cambridge University Press, Cambridge (1995)
16. Jin, F., Dougherty, E., Saraf, P., Cao, Y., Ramakrishnan, N.: Epidemiological modeling of news and rumors on twitter. In: Proceedings of the Seventh Workshop on Social Network Mining and Analysis, p. 8. ACM (2013)
17. Kakwani, N.C., Podder, N.: On the estimation of lorenz curves from grouped observations. Int. Econ. Rev. **14**(2), 278–292 (1973)

18. Karadzhov, G., Gencheva, P., Nakov, P., Koychev, I.: We built a fake news & click-bait filter: what happened next will blow your mind! In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (2017)
19. Karduni, A., Cho, I., Wesslen, R., Santhanam, S., Volkova, S., Arendt, D.L., Shaikh, S., Dou, W.: Vulnerable to misinformation? Verifi! In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 312–323. ACM (2019)
20. Karduni, A., Wesslen, R., Santhanam, S., Cho, I., Volkova, S., Arendt, D., Shaikh, S., Dou, W.: Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In: Twelfth International AAAI Conference on Web and Social Media (2018)
21. Kumar, S., Cheng, J., Leskovec, J., Subrahmanian, V.: An army of me: sockpuppets in online discussion communities. In: Proceedings of the 26th International Conference on World Wide Web, pp. 857–866. International World Wide Web Conferences Steering Committee (2017)
22. Kumar, S., Shah, N.: False information on web and social media: a survey. arXiv preprint arXiv:1804.08559 (2018)
23. Kumar, S., West, R., Leskovec, J.: Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of the 25th International Conference on World Wide Web, pp. 591–602. International World Wide Web Conferences Steering Committee (2016)
24. Kwon, S., Cha, M., Jung, K.: Rumor detection over varying time windows. PLoS One **12**(1), e0168344 (2017)
25. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: Proceedings of the 13th International Conference on Data Mining (ICDM), pp. 1103–1108. IEEE (2013)
26. Lewandowsky, S., Oberauer, K., Gignac, G.E.: Nasa faked the moon landing – therefore,(climate) science is a hoax: an anatomy of the motivated rejection of science. Psychol. Sci. **24**(5), 622–633 (2013)
27. Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D.: How social influence can undermine the wisdom of crowd effect. Proc. Natl. Acad. Sci. **108**(22), 9020–9025 (2011)
28. Matsa, K.E., Shearer, E.: News use across social media platforms 2018. Pew Research Center (2018). http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
29. Mitra, T., Wright, G.P., Gilbert, E.: A parsimonious language model of social media credibility across disparate events. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW), pp. 126–145. ACM (2017)
30. Muchnik, L., Aral, S., Taylor, S.J.: Social influence bias: a randomized experiment. Science **341**(6146), 647–651 (2013)
31. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1589–1599. Association for Computational Linguistics (2011)
32. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2921–2927 (2017). https://aclanthology.info/papers/D17-1317/d17-1317
33. Rath, B., Gao, W., Ma, J., Srivastava, J.: From retweet to believability: utilizing trust to identify rumor spreaders on twitter. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2017)
34. Rubin, V.L., Conroy, N.J., Chen, Y., Cornwell, S.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of NAACL-HLT, pp. 7–17 (2016)
35. Schelling, T.C.: Micromotives and macrobehavior. WW Norton & Company, New York (2006)
36. Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F.: The spread of fake news by social bots. arXiv preprint arXiv:1707.07592 (2017)
37. Starbird, K.: Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM). AAAI (2017)

38. Starbird, K., Maddock, J., Orand, M., Achterman, P., Mason, R.M.: Rumors, false flags, and digital vigilantes: misinformation on twitter after the 2013 Boston marathon bombing. iConference 2014 Proceedings (2014)
39. Street, C.N., Masip, J.: The source of the truth bias: Heuristic processing? Scand. J. Psychol. **56**(3), 254–263 (2015)
40. Takahashi, B., Tandoc, E.C., Carmichael, C.: Communicating on twitter during a disaster: an analysis of tweets during typhoon haiyan in the philippines. Comput. Human Behav. **50**, 392–398 (2015)
41. Tambuscio, M., Ruffo, G., Flammini, A., Menczer, F.: Fact-checking effect on viral hoaxes: a model of misinformation spread in social networks. In: Proceedings of the 24th International Conference on World Wide Web, pp. 977–982. ACM (2015)
42. Volkova, S., Bachrach, Y.: Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1 (2016)
43. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 647–653 (2017)
44. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). https://doi.org/10.1126/science.aap9559
45. Wang, W.Y.: “Liar, liar pants on fire”: a new benchmark dataset for fake news detection (2017)
46. Weninger, T., Johnston, T.J., Glenski, M.: Random voting effects in social-digital spaces: a case study of reddit post submissions. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 293–297. HT ’15, ACM, New York (2015). https://doi.org/10.1145/2700171.2791054
47. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: Proceedings of the 31st International Conference on Data Engineering (ICDE), pp. 651–662. IEEE (2015)
48. Wu, L., Morstatter, F., Hu, X., Liu, H.: Mining misinformation in social media. In: Big Data in Complex and Social Networks, CRC Press, pp. 123–152 (2016)
49. Zhang, A., Culbertson, B., Paritosh, P.: Characterizing online discussion using coarse discourse sequences. In: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM). AAAI (2017)