# Online Misinformation: Challenges and Future Directions

Miriam Fernandez
Knowledge Media Institute, Open University
Milton Keynes, UK
miriam.fernandez@open.ac.uk

Harith Alani
Knowledge Media Institute, Open University
Milton Keynes, UK
h.alani@open.ac.uk

## ABSTRACT

Misinformation has become a common part of our digital media environments and it is compromising the ability of our societies to form informed opinions. It generates misperceptions, which have affected the decision making processes in many domains, including economy, health, environment, and elections, among others. Misinformation and its generation, propagation, impact, and management is being studied through a variety of lenses (computer science, social science, journalism, psychology, etc.) since it widely affects multiple aspects of society. In this paper we analyse the phenomenon of misinformation from a technological point of view. We study the current socio-technical advancements towards addressing the problem, identify some of the key limitations of current technologies, and propose some ideas to target such limitations. The goal of this position paper is to reflect on the current state of the art and to stimulate discussions on the future design and development of algorithms, methodologies, and applications.

## KEYWORDS

Misinformation; Technology Development

**ACM Reference Format:**
Miriam Fernandez and Harith Alani. 2018. Online Misinformation: Challenges and Future Directions. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3188730

## 1 INTRODUCTION

Misinformation generates misperceptions, which have affected many domains, including economy [4], health [24], climate change [44], foreign policy [37], etc. It has become a common part of our digital media environments [26], and it is compromising the ability of our societies to form informed opinions [22][35][11]. In 2016, **post-truth** was chosen by the Oxford Dictionary as the word of the year, after achieving a 2000% increase "in the context of the EU referendum in the United Kingdom and the presidential election in the United States".

Today, around half the world's population have access to the Internet, where they can create, propagate, and consume information instantly and globally. Although misinformation is a common problem in all media, it is exacerbated in digital social media due to the speed and ease in which posts are spread, and the difficulty of providing countervailing corrective information.[1] The social web enables people to spread information rapidly without confirmation of truth, and to paraphrase this information to fit their intentions and preset beliefs [47]. An example is this public message on Facebook that went viral in Dec 2015: *This is Dearborn Michigan after the radical Islamic attack in California! These are Isis flags and Isis supporters folks but the media has not reported because of political correctness*, the demonstration, however, was anti-Isis.[2] Recent news data analysis also showed that fake news spread far more virally than real news.[3]

Several social media platforms have recently gone under heavy criticism for becoming a ripe environment for the spread of misinformation, including fake news, mistruths, and hoaxes. It is being accused of clouding people's opinions and judgement with widely shared misinformation during major events, such the US presidential elections, and the UK's Brexit referendum.[4] In reaction, Facebook and Google announced plans for combating the spread of fake news on their platforms.[5] However, while some of these plans are materialising, they are deemed to offer partial solutions to an increasingly complex socio-technical problem. People and current technologies are yet to adapt to the age of misinformation, where incorrect or misleading information is intentionally or unintentionally spread [2].

In this paper we provide a state of the art review on the existing socio-technological solutions to combat misinformation; its detection, propagation, validation and management. We analyse the key strengths and limitations of the identified technological advancements and propose some future research directions as result of the identified limitations. The goal of this position paper is to reflect on the current state of the art and to stimulate discussions on the future design and development of algorithms, methodologies, and applications that can help to successfully address the online misinformation problem.

The rest of the paper is structured as follows: Section 2 identifies four main focuses of current technological developments including: (i) the automatic detection of online misinformation (Section 3), (ii) the investigation of misinformation propagation patterns and their prediction (Section 4), (iii) the validation and fact-checking of misinformation (Section 5) and (iv) the study of the different intervention strategies used to combat misinformation (Section 6). Section 7 summarises the limitations of the studied works and


*WWW '18 Companion, April 23–27, 2018, Lyon, France*

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3188730

[1] https://www.theguardian.com/media/greenslade/2016/nov/23/heres-the-truth-fake-news-is-not-social-medias-fault
[2] http://www.factcheck.org/2015/12/dearborns-anti-isis-rally/
[3] https://hapgood.us/2016/11/13/fake-news-does-better-on-facebook-than-real-news/
[4] https://www.ipsos-mori.com/researchpublications/researcharchive/3742/the-perils-of-perception-and-the-eu.aspx
[5] https://www.facebook.com/zuck/posts/10103269806149061;https://www.facebook.com/zuck/posts/10104445245963251?pnref=story

discusses our ideas for future research directions. Discussion and Conclusions are provided in Section 8 and Section 9 respectively.

## 2 DIMENSIONS OF COMBATING ONLINE MISINFORMATION

Aiming to provide a clear picture of the current state of the art approaches to combat online misinformation, we did an extensive review of existing relevant technologies and characterised them according to the following four dimensions:

- **Misinformation content detection**: Are misinformation content and sources automatically identified? Are streams of information automatically monitored? Is relevant corrective information identified as well?
- **Misinformation dynamics**: Are patterns of misinformation flow identified and predicted? Is demographic and behavioural information considered to understand and predict misinformation dynamics?
- **Content Validation**: Is misinformation validated and fact-checked? Are the users involved in the content validation process?
- **Misinformation management**: Are citizens' perceptions and behaviour with regards to processing and sharing misinformation studied and monitored? Are intervention strategies put in place to handle the effects of misinformation?

Figure 1 presents a general reflective comparison of eleven of the most popular platforms developed to aid in the battle against misinformation. This comparison reflects our view of how much attention and focus each platform gives to the four dimensions above. To generate this figure two independent assessors have assigned a score (from 0 to 10) to each the four dimensions for each tool. The image reflects the average of the two assessors for each dimension.

## 3 MISINFORMATION DETECTION

Large amounts of misinformation have been observed to spread online in viral fashion. Examples include rumours [26], false news [10], hoaxes [39], and even elaborate conspiracy theories [5]. Several approaches and tools have emerged in recent years to automatically or semi-automatically identify misinformation based on the characteristics of the **content** (text as well as multimedia images/videos), or the source of the misinformation and the **network** of that source. **Contextual** information, including a compiled list of **misleading sites** and **microblog-specific features**, such as hashtags or mentions in Twitter, are often used to complement the above.

Works of Castillo and Colleagues [11][12][31] studied information credibility on Twitter mainly based on content features, and created supervised machine learning classifiers to detect this credibility. Their studies concluded that credible tweets tend to include more URLs, and are longer than non-credible tweets. Additionally, question and exclamation marks tend to concentrate on non-credible tweets, frequently using first and third-person pronouns. These studies derived on the creation of the **TweetCred** system,[6] a real-time, web-based system (available as browser extension) to assess credibility of content on Twitter. The system provides a score of credibility for each tweet, based on the previously generated classifiers and it validates this score by asking user feedback. Similar tools developed as browsing extensions include **Fake News Alert**[7] and **B.S. Detector**,[8] which rely on manually compiled lists of misleading websites, such as the one generated by Zimdars [72] and **Dispute Finder** [20], which is based on a database of known disputed claims generated by crawling websites that already maintain a list of disputed claims. Qazvinian and colleagues [55] also studied content features for misinformation detection. They concluded that lexical and Part of Speech (POS) patterns are key for correctly identifying rumours. Hashtags can result in high precision, but lead to low recall.

In addition to the analysis of content, other works and systems, focus on the use of network analysis techniques to detect misinformation [57][59][31][26]. The studies show that different diffusion patterns exist that characterise misinformation vs. legitimate memes, with misinformation patterns propagating in a more viral way [26] and often being generated by bots and not humans [57]. On the other hand credible news tend to originate at a single or a few users in the network, have many re-posts and propagate through authors who have previously written a large number of messages and register more friends [31].

Tools to detect and display the diffusion of misinformation include **Truthy** [57], **RumorLens** [58] and **Twitter trails** [46]. These tools are based on a semi-automatic approach where users can explore the propagation of a rumour with an interactive dashboard. However, they do not monitor the social media stream automatically to detect misinformation, but require the user to input a specific rumour to investigate. Aiming to address this issue Shao [60] and colleagues developed **Hoaxy** [60], a platform that automatically monitors the social stream, detects, and analyses online misinformation. Following this trend **Facebook** has recently released new tools to help combat the spread of fabricated news stories. As opposed to Hoaxy, Facebook tools not only use a combination of content and network analysis but also include user feedback to accurately identify fake news. This system is under continuous development and testing.[9] However, current efforts to combat misinformation have been criticized because they fall short on preventing misuse of the platform.[10] **Google's** proposal to tackle misinformation also includes asking users for feedback[11] by providing a link at the bottom of the snippet box.

As it can be observed, some of the limitations of current systems for misinformation detection include: (a) providing alerts without any rationale or explanation of their decisions, and (b) generally disengaging users by regarding them as passive consumers rather than active co-creators and detectors of misinformation. Another element to consider is that, automatic systems for misinformation detection based on known features can potentially be fooled, and carefully crafted misinformation may go undetected.

[6] https://chrome.google.com/webstore/detail/tweetcred/fboklljinlogeihdnkikeeneiankdgikg?hl=en

[7] https://chrome.google.com/webstore/detail/fake-news-alert/aickfmgnhocegpdbfnpfnedpeionfkbh?hl=en

[8] http://bsdetector.tech/

[9] https://www.theverge.com/2017/12/21/16804912/facebook-disputed-flags%2Dmisinformation-newsfeed-fake-news

[10] https://www.engadget.com/2018/01/19/facebooks-fake-war-on-fake-news/

[11] https://www.engadget.com/2018/01/31/google-tackles-fake-news-in-snippets/

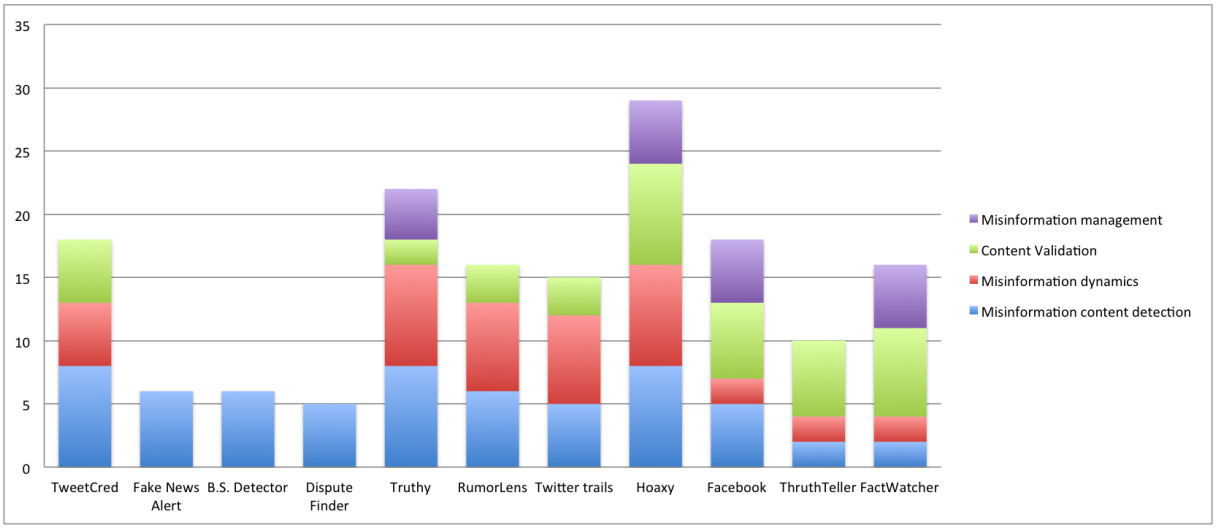


**Figure 1: Comparison of relevant platforms according to the four identified dimensions**

## 4 MISINFORMATION DYNAMICS

Online social networks are characterised by **homophily** [45], **polarisation** [17], **algorithmic ranking/personalisation** [53][29][3], and **social bubbles** [49]. These characteristics create information environments with low content diversity and strong social reinforcement, which has an effect on the information users are exposed to and on how information propagates. All of these factors, coupled with the fast news life cycle [14], influence the dynamics of social news sharing, and particularly the ways misinformation initiates and propagates.

Ratkiewicz [57] analysed the spread of misinformation in the context of political campaigns and showed how, in its initial phase, the propagation of misinformation exhibits pathological diffusion graphs. These graphs can take many forms, including high numbers of unique injection points with few or no connected components or strong star-like topologies. However, once the community has accepted the misinformation, its propagation cascade will quickly become indistinguishable, hence early identification of misinformation is critical. This works also highlights the relevance of **bots** initiating the process of misinformation spread. Boshmaf [8] and Freitas[25] reported that simple automated mechanisms that produce contents and boost followers yield successful infiltration strategies of misinformation. However, nobody knows exactly how many social bots populate social media, or what share of content, and particularly misinformation, can be attributed to bots [21]. A similar dangerous phenomenon is **crowdturfing**, where crowdworkers are hired to support and propagate arguments or claims, simulating grassroots social movements. [68][40][41].

Del Vicario [18] studied misinformation propagation and reported that users mostly tend to select and share content based on homogeneity (**echo chambers**), causing reinforcement and fostering confirmation bias, segregation, and polarisation [5], an effect exacerbated by the platforms; personalisation and ranking algorithms [53][29][3]. This is also confirmed by [62], who concluded that rumor spreaders form strong partisan structures. Del Vicario also shows that different types of misinformation propagate differently. While misinformation around scientific news reach a higher level of diffusion faster, it also decays faster. On the other hand, conspiracy theories are assimilated more slowly but are propagated over longer time periods. Friggeri [26] studied the propagation of rumors within Facebook and concluded that: (i) misinformation cascades run deeper than non-misinformation cascades within the network, (ii) even when denied, the rumour cascade continues to propagate, as there are many more non-denied re-shares than denied ones and, (iii) a rumour can lie dormant for weeks or months, and then it can become popular again. More recent work by [74][60] has also found that misinformation spreads faster and more widely across the network, with fact-checking content typically lagging that of misinformation by 10-20 hours. These works also suggest that misinformation-mongering is dominated by few very active accounts that bear the brunt of the promotion and spreading of misinformation, whereas the propagation of fact checking is a more distributed, grass-roots activity.

Understanding users [67] and their motivations [13] are also key aspects to understand misinformation dynamics. Wagner and colleagues [67] studied the susceptibility of users to interact with bots and spammers. This study, conducted over Twitter, concludes that susceptible users tend to communicate with many different users, use more social words and show more affection than non-susceptible users. Similarly, [71][13] showed that, personality aspects influence misinformation dynamics. Extroverts and individuals with high cooperativeness and high reward dependence are founded more prone to share misinformation, but no significant differences were found in terms of gender. The key motivations behind misinformation spreading include information seeking and socialising. Psychology also shows that individuals with higher anxiety levels are more likely to spread misinformation [34].

The effect of finite memory and attention on the spread of missinformation has been studied by Tambuscio [63] and Qiu [56]. These studies conclude that in social media environments, where users

are influenced by high information load and finite attention, low quality information is likely to go viral.

While all these studies provide important insights on how misinformation propagates, they do not analyse in depth the topology and the typology of the social network that is consuming and sharing misinformation. Similarly, deeper studies are needed to understand how, not only demographics (age, gender, geographical location), but also user behaviour influences the spread of misinformation.

## 5 CONTENT VALIDATION

Information validation practices are key to identify misinformation.[12] More than 110 independent **fact-checking groups** and organisations emerged online around the world over the past decade, and half of them were established in European countries [30], (e.g., Full Fact in the UK, Snopes and Root Claim in the US, FactCheckNI in Northern Ireland, and Pagella Politica in Italy, to name just a few). These groups and organisations aim to provide a frontline service in dealing with false information online following guidelines, such as the ones captured by The Verification Handbook.[13]

However, fact checking is a time-consuming verification practice that makes it near impossible to compete with the speed of social media. **Computational fact-checker** initiatives have also emerged in the last few years with the aim of enhancing our ability to evaluate the veracity of dubious information. Among these works Ciampaglia [15] exploited implicit information from the topology of the Wikipedia Graph. Their results show that network analytics methods, in conjunction with large-scale knowledge repositories, are effective towards automatic fact-checking methods. Baoxu [61] follows a similar approach, but proposes a path mining approach over large-knowledge graphs (DBpedia[14] and SemMedDB[15]) to leverage a collection of factual statements for automatic fact-checking. Besides the analysis of textual sources, works like the one of Boididou and colleagues [7] focus on the automatic verification of unreliable media content by building classifiers from multiple user and content features.

An additional problem of fact-checking initiatives is that they are often disconnected from where the crowds read, debate, and share misinformation with little or no awareness of any invalidation offered by the fact checkers. To address these issues several initiatives have emerged that aim to bring the results of fact-checking initiatives closer to the public. Examples are **TruthTeller**,[16] developed by the Washington Post, which transcribes political videos and checks them against a database that draws on PolitiFact[17] and FactCheck.org.[18] The program tells viewers which statements are true or false. **Truth Googles**[19] implements a similar approach in a brower plug-in, also based on these databases. **Hoaxy** [60] integrates the efforts of fact-checking with a continuous monitoring of the social stream, making the social media information and the fact-checking information simultaneously available for the user. **FactWatcher** [33] complements previous approaches by considering different types of facts, including situational facts, one-of-the-few facts, and prominent streaks. As opposed to previous tools that are oriented to the general public, FactWatcher[20] is focused on supporting journalist with the creation of news stories. In the same fashion, ClaimBuster[21] [32] provides computational tools to assist professionals in understanding and verifying claims. Particularly, it assigns scores to factual claims indicating whether they should be checked, providing a priority ranking to help fact-checkers.

**Crowdsourcing** initiatives have also been considered to validate and verify information [73]. One of the most recent initiatives by Facebook integrates crowdsourcing with fact-checkers (Poynter and Politifact, among others) to fight fake news. Users can mark stories as fake and see warnings that indicate the story has been disputed by third-party fact-checkers. Systems like **TweetCred**[22] and **Trudhy**[23] use crowdworkers to annotate data and train machine learning algorithms that can learn from human annotations when assessing the credibility of tweets.

Although the work of fact checkers and crowdsourcing initiatives is really valuable in correcting misinformation, they are faced by a number of complex challenges, which limits their ability to change existing misperceptions. Not only they are unable to keep with the high volume of misinformation generated online, or are disconnected from where users read, debate and share misinformation, but simply publishing corrective information by fact checkers is often regarded as insufficient for changing misinformed beliefs and opinions [1]. Whether a claim is accepted by an individual is strongly influenced by the individual's believe system, since it is common to look for information that confirm our believes (confirmation bias) and don't scrutinize contrary ideas to avoid or lessen cognitive dissonance (motivated reasoning) [38] [71]. Moreover, Penny Cook and colleagues also highlighted in a recent study the problem of the "Illusory Truth Effect" when it comes fake news and corrective information [54]. Their study shows how repetition can increase the perceived accuracy of plausible but false statements. Garret and colleagues also show how, compared to post-exposure corrections, real-time corrections may cause users to be more resistant to factual information [28]. It is therefore important to consider not only which corrective information should be provided, but when, how and to whom should it be provided.

## 6 MISINFORMATION MANAGEMENT

Combating misinformation is a complex task, and there is consensus in psychology literature that simply presenting people with corrective information is likely to fail in changing their salient beliefs and opinions, or may, even, reinforce them [23][50][51] [28][54]. People often struggle to change their beliefs even after finding out that the information they already accepted is incorrect [16][64]. Nevertheless, some strategies have been found to be effective in correcting misperceptions [43], such as providing an explanation

[12] http://www.poynter.org/2016/366-links-to-understand-fact-checking-in-2016/440618/
[13] http://verificationhandbook.com
[14] http://wiki.dbpedia.org/
[15] https://skr3.nlm.nih.gov/SemMedDB
[16] www.washingtonpost.com/news/ask-the-post/wp/2013/09/25/announcing-truth-teller-beta-a-better-way-to-watch-political-speech/?utm_term=.b78b7c187228
[17] http://www.politifact.com/
[18] http://www.factcheck.org/
[19] https://www.media.mit.edu/projects/truth-goggles/overview/
[20] http://idir.uta.edu/factwatcher/nba.php
[21] http://idir-server2.uta.edu/claimbuster/
[22] http://twitdigest.iiitd.edu.in/TweetCred
[23] http://truthy.indiana.edu/

rather than a simple refute [52], exposing to related but disconfirming stories [6], and revealing the demographic similarity of the opposing group [27]. Recent work by Cambridge University is also considering the use of "fake news vaccine" to immunise users against the problem by "pre-emptively exposing" readers to a small "dose" of the misinformation [65]. An online game[24] has been released as part of this research to let players experience what is like to create and spread misinformation so that they are more likely to identify it. An alternative approach for dealing with pervasive misinformation is to seek more direct behavioral interventions that encourage people to make certain decisions over others [42].

Works that have attempted to stop the spread of misinformation in social networks generally use three main strategies: (i) combating it with **facts** [9][48][70], (ii) **malicious account detection** in early stage [69][19][41] and (iii) the use of **ranking and selection strategies** based on corrective information.

Among the works that have attempted to combat the spread of misinformation with **facts** Budak et al. [9] introduced the notion of competing campaigns to counteract the effect of misinformation. With this purpose, they designed the Multi-Campaign Independence Cascade Model (MCICM) and studied multiple methods to choose the optimal subset of users as seeds to propagate the "good" campaign. Similar efforts include the works of [48] and [70]. The first work aims to find the "Node Protectors", i.e., the smallest set of highly influential nodes whose "decontamination" with good information helps to contain the viral spread of misinformation. The second work aims to identify the most important disseminators of misinformation to "inject correct information" in the diffusion. These models of information propagation present however several limitations. First they are based on the assumption that once a user is "contaminated" with "good" information she will propagate this information among her network. However, persuading users to adopt certain beliefs, and propagate them is not trivial [43]. Secondly, these works assume that the models of diffusion of "good" and "bad" information are coincident, when in reality; they may actually not spread at the same rate. Indeed, several recent works have found that misinformation spreads wider and faster [74][60]. This type of misinformation management approach has also being recently used by Twitter. The company notified more than 1.4 million people about the fact that they interacted during the US elections with accounts generated by the Russian government-linked organisation Internet Research Agency. While Twitter mentions that a survey will be send to a small group of people to gain feedback, little is known so far about the effects of this initiative.[25]

Regarding the methods focused on the **early detection of malicious accounts** we can highlight works that aim to identify spammers [69], bots [21], crowdturfing [68][40] and malicious accounts in general [19][41]. These techniques generally focus on the analysis of various user, temporal, geographical and linguistic features in order to successfully identify these accounts. However, it is unclear what intervention strategies to use in order to stop the spread of misinformation once these accounts have been identified. Twitter, for example, is currently suspending accounts associated with duplicative or suspicious activity.[26]

A third type of misinformation management approach, currently used by organisations like Google and Facebook, is the collection of feedback from users regarding misinformation content, and the use of this feedback as a factor to enhance **information selection and ranking** mechanisms. By doing so, these platforms aim to avoid and/or limit displaying and recommending content that has been previously tagged as 'missinformation' by other users.

## 7 RESEARCH DIRECTIONS

In this section we summarise the main limitations we identified, according to the four dimensions we studied, and propose some ideas to target such limitations.

- **Misinformation Identification:** Current misinformation identification approaches tend to focus on (a) alerting users without **rationale** or explanation of their decisions, and (b) **disengaging users** by regarding them as **passive consumers** rather than active co-creators and detectors of misinformation.
- **Misinformation Dynamics:** Most current studies on misinformation dynamics (a) do not analyse the influence of the **topology** and the **typology** of the social network on the consumption and sharing of misinformation, and (b) do not take into account how **the misinformation-handling behaviour of users** influences the spread of misinformation.
- **Content Validation:** Current fact checkers and crowdsourcing initiatives for content validation (a) are not able to cope with the **high volume** of misinformation generated online, and (b) are often **disconnected** from where the users tend to read, debate and share misinformation.
- **Misinformation Management:** Common misinformation management strategies (a) do not go beyond the generation of **facts** and the early detection of malicious accounts, and (b) tend to focus on the **technical** and not on the **human aspects** of the problem (i.e., the motivations and behaviours of the users when generating and spreading misinformation).

As we can observe from the above summary, the limitations of current technologies are numerous and diverse, which highlight various directions for further discussions and research. Tackling the new societal challenge of misinformation requires closely involving the users and strengthening their resilience to misinformation. Future technology should therefore help promoting: (1) **Empowerment**, by raising individual and collective awareness of current misinformation content and sources, (2) **Engagement**, by fostering networking and cross-communication between users, (3) **Education**, by informing users of advanced misinformation analysis results and predictions, and (4) **Encouragement** of all users to play a role in detecting, in/validating, and combating misinformation.[27] More specifically, further advancements are required in the following dimensions.

[24] https://www.getbadnews.com
[25] https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html
[26] https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html
[27] https://www.demos.co.uk/files/Resilient_Nation_-_web-1.pdf

*7.0.1 User Involvement.* While digital literacy and media literacy initiatives have emerged in the last few years to help users identify misinformation, [28] most of the technologies we surveyed do not appear to closely involve the users to battle misinformation, hence considering users as passive consumers rather than active co-creators and detectors of misinformation. Only a few systems, such as TweetCred, involves users, but mainly to validate the results of their misinformation detection algorithms. Our hypothesis is that, to advance the state of the art, we need to closely involve users in the process of misinformation detection and management.

Many insights have been provided from social science research with regards to what works and what does not, to correct or to limit the spread of misinformation. However, translating such insights and successful approaches into delivery tools would require the participation of all stakeholders, including end users, social scientists, computer scientists, educators, etc., in the co-design of their functions, user interfaces, and delivery methods. This would increase the acceptance of such tools, and thus their impact of combating misinformation.

*7.0.2 Misinformation Dynamics.* With regards to the exploration of misinformation dynamics, we believe that the topology and typology of the network could play an important role in how misinformation spreads. Works should therefore study similarities and differences of misinformation spread, across different platforms, and how platform-specific and network-specific features influence the dynamics of misinformation.

Understanding these dynamics, and the user, topological, and typological factors that influence them, can be used to develop models that predict how, where, and by whom certain misinformation are likely to spread.

*7.0.3 User Behaviour.* User-behaviour may be a key factor in how misinformation is spread. Investigating the behavioural patterns that are commonly associated with the propagation of misinformation could help to better predict and control the cascade of misinformation.

With technology, we would be able to study the impact of various misinformation interventions and correction techniques at large scale, to better understand their impact on user behaviour towards misinformation. Many such studies have already been reported in social science literature. However, executing them on very large numbers of users (e.g., hundreds of thousands), and monitoring their results over longer periods of time (e.g., several weeks, months, and years) would required a high degree of automation. Such large-scale and longer-term experiments could yield new or more representative insights that are very difficult to obtain manually.

*7.0.4 Content Validation.* Validating content is a complex part of the misinformation control cycle. Corroborating and refuting facts is not a trivial task, particularly considering the volume and the velocity at which online information is often generated. We can however aim to embed fact checkers into the environments where users tend to read, debate, and share misinformation. For example, this can be achieved by developing browser and social-media platform plug-ins that are able to assess existing discussions and shared articles, and highlight related factual or corrective information that is available from any of the known fact-check sites.

In spite of recent research and technological developments, and the rise of fact-checking sites, there is still a clear lack of tools to support users who would like to validate any piece of information. Such validation could, for example, include searching various fact-checking sites for related articles, and assessing the legitimacy of the information source (e.g., whether it is from a known fakenews site). Many lay-users might be unaware of such validation actions and possibilities, or lack the basic skill to perform them effectively.

*7.0.5 Misinformation Management.* Regarding the generation of effective misinformation management strategies, we believe that understanding how citizens behave towards misinformation, what opinions they form about it, and how these opinions evolve over time, are key to successfully manage the impact of misinformation.

Technology can be used to test the effectiveness of various misinformation management policies and techniques, as well as to deploy them at scale.

## 8 DISCUSSION

In this work we have provided an overview of the current technology developments towards battling the problem of online misinformation. Due to the relevance of this problem, new works are constantly emerging from a variety of disciplines (social science, computer science, communication, political science, etc.). We are therefore aware that a high number of works are not captured in this paper. However, we hope that the current compilation provides a simple and clear overview of the multiple dimensions of the problem, the existing technological solutions, and their limitations.

We have also proposed multiple research directions as result of the conducted analysis. All of these directions are based on a strong user-focus. It is our view that the solution to the new societal challenge of misinformation is not for social media platforms to become the arbiters of truth, which raises various ethical and philosophical dilemmas, but to closely involve the users as part of the solution.

As mention earlier, misinformation is a complex problem involving human, societal and technological factors. We can therefore not look at the problem with a unique lens. Multidisciplinary research is needed to design and develop methodologies, practices, policies and technologies able to effectively combat misinformation.

## 9 CONCLUSIONS

In this position paper we have investigated the existing technological developments towards combating the problem of online misinformation. We have analysed these works following four key dimensions: (i) misinformation detection, (ii) misinformation dynamics, (iii) content validation and, (iv) misinformation management. We have investigated the limitations of these works and identified the lack of user involvement and consideration as a key limitation in all four dimensions. We have subsequently proposed various research directions focused on involving users as participants and co-creators of misinformation technology. We hope that this paper stimulates discussions across disciplines on how to enhance the

[28] https://webliteracy.pressbooks.com/;https://fakenews.publicdatalab.org/;https://thetrustproject.org/

current landscape of technology development to effectively target the problem of online misinformation.

**Acknowledgments.** Research funded by Co-Inform, H2020 program of the European Union, grant agreement 770302.

## REFERENCES


[1] Michelle A Amazeen. 2013. A Critical Assessment of Fact-checking in 2012. (2013).

[2] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. 2015. A model for identifying misinformation in online social networks. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 473–482.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[4] Larry M Bartels. 2002. Beyond the running tally: Partisan bias in political perceptions. *Political Behavior* 24, 2 (2002), 117–150.

[5] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 355–356.

[6] Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

[7] Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. 2014. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 743–748.

[8] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference*. ACM, 93–102.

[9] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*. ACM, 665–674.

[10] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. 2011. The persistent effects of a false news shock. *Journal of Empirical Finance* 18, 4 (2011), 597–615.

[11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588.

[13] Xinran Chen and Sei-Ching Joanna Sin. 2013. âĂŸMisinformation? What of it?âĂŹMotivations and individual differences in misinformation sharing on social media. *Proceedings of the Association for Information Science and Technology* 50, 1 (2013), 1–4.

[14] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2015. The production of information in the attention economy. *Scientific reports* 5 (2015).

[15] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.

[16] Michael D Cobb, Brendan Nyhan, and Jason Reifler. 2013. Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology* 34, 3 (2013), 307–326.

[17] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM* 133 (2011), 89–96.

[18] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[19] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2013. Compa: Detecting compromised accounts on social networks.. In *NDSS*.

[20] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*. ACM, 341–350.

[21] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[22] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 515–540.

[23] DJ Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology* 38, S1 (2017), 127–150.

[24] Gary L Freed, Sarah J Clark, Amy T Butchart, Dianne C Singer, and Matthew M Davis. 2010. Parental vaccine safety concerns in 2009. *Pediatrics* 125, 4 (2010), 654–659.

[25] Carlos Freitas, Fabricio Benevenuto, Saptarshi Ghosh, and Adriano Veloso. 2015. Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 25–32.

[26] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades.. In *Proceedings of the 2014 International Conference on Web and Social Media. ICWSM 2014*.

[27] R Kelly Garrett, Erik C Nisbet, and Emily K Lynch. 2013. Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication* 63, 4 (2013), 617–637.

[28] R Kelly Garrett and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1047–1058.

[29] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014).

[30] LUCAS Graves and FEDERICA Cherubini. 2016. The rise of fact-checking sites in Europe. (2016).

[31] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter. In *Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain*.

[32] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1803–1812.

[33] Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. 2014. Data in, fact out: automated monitoring of facts by FactWatcher. *Proceedings of the VLDB Endowment* 7, 13 (2014), 1557–1560.

[34] Marianne E Jaeger, Susan Anthony, and Ralph L Rosnow. 1980. Who hears what from whom and with what effect: A study of rumor. *Personality and Social Psychology Bulletin* 6, 3 (1980), 473–478.

[35] Anna Kata. 2010. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine* 28, 7 (2010), 1709–1716.

[36] Per Kristensson, Jonas Matthing, and Niklas Johansson. 2008. Key strategies for the successful involvement of customers in the co-creation of new technology-based services. *International journal of service industry management* 19, 4 (2008), 474–491.

[37] Steven Kull, Clay Ramsay, and Evan Lewis. 2003. Misperceptions, the media, and the Iraq war. *Political Science Quarterly* 118, 4 (2003), 569–598.

[38] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.

[39] Tom Lauricella, Christopher S. Stewart, and Shira Ovide. 2013. Twitter hoax sparks swift stock swoon. *The Wall Street Journal* 23 (2013).

[40] Kyumin Lee, Prithivi Tamilarasan, and James Caverlee. 2013. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media.. In *ICWSM*.

[41] Sangho Lee and Jong Kim. 2014. Early filtering of ephemeral malicious accounts on Twitter. *Computer Communications* 54 (2014), 48–57.

[42] Thomas C Leonard. 2008. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy* 19, 4 (2008), 356–360.

[43] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131.

[44] Aaron M McCright and Riley E Dunlap. 2011. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly* 52, 2 (2011), 155–194.

[45] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[46] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. ACM, 69–72.

[47] Dung T Nguyen, Nam P Nguyen, and My T Thai. 2012. Sources of misinformation in online social networks: Who to suspect?. In *Military Communications Conference, 2012-MILCOM 2012*. IEEE, 1–6.

[48] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 213–222.

[49] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. Measuring online social bubbles. *PeerJ Computer Science* 1 (2015), e38.

[50] Brendan Nyhan. 2010. Why the" Death Panel" Myth Wouldn't Die: Misinformation in the Health Care Reform Debate. In *The Forum*, Vol. 8.

[51] Brendan Nyhan and Jason Reifler. 2013. Which Corrections Work? Research results and practice recommendations. *New America Foundation Media Policy*

*Initiative Research Paper* (2013).

[52] Brendan Nyhan and Jason Reifler. 2015. Displacing misinformation about events: An experimental test of causal corrections. *Journal of Experimental Political Science* 2, 1 (2015), 81–93.

[53] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[54] Gordon Pennycook, T Cannon, and D Rand. 2017. Implausibility and Illusory Truth: Prior Exposure Increases Perceived Accuracy of Fake News but Has No Effect on Entirely Implausible Statements. *Unpublished Paper Manuscript, December* 11 (2017), 2017.

[55] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[56] Xiaoyan Qiu, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour* 1, 7 (2017), 0132.

[57] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 249–252.

[58] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*.

[59] Eunsoo Seo, Prasant Mohapatra, and Tarek Abdelzaher. 2012. Identifying rumors and their sources in social networks. *SPIE defense, security, and sensing* (2012), 83891I–83891I.

[60] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 745–750.

[61] Baoxu Shi and Tim Weninger. 2015. Fact checking in large knowledge graphs: A discriminative predict path mining approach. *arXiv preprint arXiv:1510.05911* (2015).

[62] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *new media & society* 19, 8 (2017), 1214–1235.

[63] Marcella Tambuscio, Diego FM Oliveira, Giovanni Luca Ciampaglia, and Giancarlo Ruffo. 2016. Network segregation in a model of misinformation and fact checking. *arXiv preprint arXiv:1610.04170* (2016).

[64] Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33, 3 (2016), 460–480.

[65] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1, 2 (2017).

[66] William H Voorberg, Viktor JJM Bekkers, and Lars G Tummers. 2015. A systematic review of co-creation and co-production: Embarking on the social innovation journey. *Public Management Review* 17, 9 (2015), 1333–1357.

[67] Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)* 2, 4 (2012), 1951–1959.

[68] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2012. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 679–688.

[69] Steve Webb, James Caverlee, and Calton Pu. 2008. Social Honeypots: Making Friends With A Spammer Near You.. In *CEAS*.

[70] Huiyuan Zhang, Huiling Zhang, Xiang Li, and My T Thai. 2015. Limiting the spread of misinformation while effectively raising awareness in social networks. In *International Conference on Computational Social Networks*. Springer, 35–47.

[71] Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Chongde Lin, Qinghua He, Chunhui Chen, He Li, Robert K Moyzis, Jared Lessard, and Qi Dong. 2010. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual Differences* 48, 8 (2010), 889–894.

[72] Melissa Zimdars. 2016. False, misleading, clickbait-y, and satirical news sources. https://docs.google.com/document/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/preview. (2016).

[73] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 347–353.

[74] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.