Front. Comput. Sci., 2021, 15(3): 153806
https://doi.org/10.1007/s11704-020-9256-0

REVIEW ARTICLE

The mass, fake news, and cognition security

Bin GUO (✉)[1], Yasan DING[1], Yueheng SUN[2], Shuai MA[3], Ke LI[1], Zhiwen YU[1]

1 School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
2 School of Cyber Security, Tianjin University, Tianjin 300350, China
3 School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract The widespread fake news in social networks is posing threats to social stability, economic development, and political democracy, etc. Numerous studies have explored the effective detection approaches of online fake news, while few works study the intrinsic propagation and cognition mechanisms of fake news. Since the development of cognitive science paves a promising way for the prevention of fake news, we present a new research area called Cognition Security (CogSec), which studies the potential impacts of fake news on human cognition, ranging from misperception, untrusted knowledge acquisition, targeted opinion/attitude formation, to biased decision making, and investigates the effective ways for fake news debunking. CogSec is a multidisciplinary research field that leverages the knowledge from social science, psychology, cognition science, neuroscience, AI and computer science. We first propose related definitions to characterize CogSec and review the literature history. We further investigate the key research challenges and techniques of CogSec, including human-content cognition mechanism, social influence and opinion diffusion, fake news detection, and malicious bot detection. Finally, we summarize the open issues and future research directions, such as the cognition mechanism of fake news, influence maximization of fact-checking information, early detection of fake news, fast refutation of fake news, and so on.

Keywords cyberspace, cognition security, fake news, crowd computing, human-content interaction

1 Introduction

The rapid popularization and development of social networks have created a direct path from content producers to consumers, changing the way users access information, debate, and form their opinions. Instead of receiving news from traditional mechanisms, such as news broadcasting or daily news programs, people are turning to social media platforms that expose themselves to a broader range of opinions and statements about daily issues. The growth of *social media* has changed patterns of information consumption deliberately and incidentally, and social media platforms have become a major source of news diffusion, such as Facebook, Twitter and YouTube. Although social networks have accelerated the dissemination of information and promoted the communication of people, contemporary social media platforms have gradually become the hotbed of spreading fake news due to their low cost, easy access, and high anonymity. Besides, the existence of social bots, botnets, and trolls hasalso been a severe problem on social media platforms. It is reported that as many as 60 million trolls could be spreading fake news on Facebook [1].

Fake news is posing threats to diverse domains, such as vaccine safety, climate change, political election, and stock stability [2]. For example, during the U.S. presidential election in 2016, *PolitiFact*, an independent fact-checker of political statements, judged 70% of all claims about Donald Trump to be false or mostly false, and found that Trump's supporters were far more likely to consume fake news than Clinton's supporters [3]. Consequently, "fake news" was named the "word of the year" by Collins Dictionary in 2017 since it had already aroused concerns around the world. In addition to political interference, fake news can also do great damage to social stability. For example, the fake news on social media about the Turkish government's implementation of capital controls led to a 20% drop in the lira against the US dollar, causing huge economic loss in Turkey. The inaccurate news which claimed that the border between Greece and North Macedonia was open induced hundreds of migrants and refugees to pour across the Greek border, further resulting in the clash between Greek police and migrants. Therefore, it can be seen that fake news is one of the current greatest threats to democracy, economy, and journalism [4].

In 2018, *Science* magazine launched a special issue about "Fake News", where they discussed the conception, network propagation mechanism and social influence of fake news [5,6]. Lazer et al. [6] identify two categories of fake news interventions, including empowering individuals to evaluate the fake news and utilizing platform-based detection algorithms. Moreover, Ruths [7] divides the dissemination of fake news into five key components: publishers, authors, articles, audiences, and rumors. Qiu et al. [8] believe that information overload contributes to the degradation of human's ability to judge the truth of news articles.

An urgent concern is that the development of Artificial Intelligence (AI) puts forward higher requirements for fake news

Received July 16, 2019; accepted January 14, 2020

E-mail: guob@nwpu.edu.cn

detection. The research on fake news will extend from text to high-quality, machine-generated and manipulated images, videos, and audios [9]. For instance, Deepfakes [10,11] hurts public feelings and affects political situations by creating audios or videos of real people they never said or did with deep neural networks, which has been widely used to forge politicians' speeches and illegal evidence [12].

To summarize, fake news can influence emotions, opinions, and other cognition activities through human-content interactions. With the idea that some information succeeds due to their content taps into general cognitive preferences [13], it is significant to understand the cognition and communication mechanisms of fake news before effectively preventing its dissemination. This paper presents a promising research area called "Cognition Security (CogSec)", which aims to *understand the interaction patterns, cognition behaviors, and social influence & diffusion mechanism between human and fake news, and investigate the efficient ways to debunk fake news and maintain human cognition security.*

CogSec is a multidisciplinary research field that leverages knowledge from social science, psychology, cognition science, neuroscience, AI, and computer science.

In particular, the main contribution of this work is three folds.

- Characterizing the Cognition Security (CogSec) research area, ranging from its concept model to research scope.
- Investigating the main research challenges of CogSec and presenting the state-of-the-art techniques to address these issues.
- Discussing the open issues and future research directions of CogSec.

2 Characterizing cognition security

In addition to fake news, there are other types of information spreading on social media platforms threatening the CogSec, such as *rumor*, *hoax*, *click-bait*, *disinformation*, and *misinformation*. The widely-recognized definitions are summarized in Table 1.

For characterizing the research area of cognition security, this section firstly presents the problem statement about CogSec. In this paper, we follow the definition of *fake news* used in recent papers [18,19].

Definition 1 Fake News: *A news article that is intentionally and verifiably false.*

Abundant social media users generate a massive number of contents based on social interactions.Users interact with such online content and their perceptions, behaviors, and knowledge are implicitly influenced [20,21]. We define the human-content interaction as follows.

Definition 2 Human-Content Interaction: *Publish, share, like, and comment on online content* (*e.g.*, *news articles*, *posts*, *photos*, *videos*, *etc*.).

We further give the definitions of cognition security and cognition security protection.

Definition 3 Cognition Security: *Cognition Security (CogSec) refers to the potential impacts of fake news on human cognition, ranging from misperception, untrusted knowledge acquisition, targeted opinion/attitude formation, to biased decision making.*

Definition 4 Cognition Security Protection: *CogSec protection is committed to credible interventions to ensure humans' CogSec, including the techniques of cognition mechanism investigation, diffusion pattern mining, early fake news detection, malicious bot detection, and so on.*

CogSec extends the security paradigm. For example, traditional vision of *network security* [22] mainly emphasizes data and information security, while CogSec focuses on the complex interaction mechanism between human cognition and multimodal contents, expanding from the traditional "*machine*" security to "*human-machine*" fusion security, as presented in Table 2. In the field of security, there is another term, *Cognitive Security*, which looks similar to CogSec. However, the term *cognitive security* means using AI and other technologies to protect the safety of digital devices from hackers, which focuses on the problem of machine security [23,24]. In conclusion, the CogSec is different from the term Cognitive Security, which studies more about the interaction mechanism between human and contents in the cyberspace.

Recently, there have been several related studies and important findings, and representative ones are presented below.

(1) Echo chamber [25,26] It traps users by exposing them to opinions and beliefs that they have agreed on [27]. Echo chamber is compounded by the rise of algorithmic news recommendation and content filtering [28], which makes users always

Table 1 Definitions of some types of false information

Term	Definition
Rumor	"An item of circulating information whose veracity status is yet to be verified at the time of posting." [14]
Hoax	"A deliberately fabricated falsehood made to masquerade as truth." [15]
Click-bait	"A piece of low-quality journalism which is intended to attract traffic and monetize via advertising revenue." [16]
Disinformation	"Fake or inaccurate information which is intentionally false and deliberately spread." [17]
Misinformation	"Fake or inaccurate information which is unintentionally spread." [17]
Fake news	"A news article that is intentionally and verifiably false." [18]

Table 2 Differences with other concepts

Term	Research focus	Security paradigm
Network security	Data and content security	Machine security
Cognitive security	Digital devices security	Machine security
Cognition security	The interaction and cognition mechanism between human and contents in the cyberspace	Human-machine security

browse their favorite information and implicitly influences users' cognition behaviors. For example, Barberá et al. [29] observe that information is mainly exchanged among users with similar ideological preferences in the case of political issues. Similarly, Bessi [30] demonstrates that the echo chamber reinforces selective exposure and group polarization. People tend to concentrate on confirming claims and ignore obvious objections because they mainly focus on their preferred information. Moreover, Zajonc et al. [31] assume that the perceived accuracy of false information increases linearly with the frequency of exposure to the same false information, which means that fake news repeatedly appearing in echo chambers may gradually be accepted as true news. In particular, a highly homogeneous echo chamber in social networks can decrease people's ability to identify the credibility of news and increase their misperceptions, which contributes to the dissemination of fake news [32].

(2) Online gatekeeper [33,34] It refers to the information controller during the information dissemination, usually acting as a content filter, selector or manipulator [35]. Xu and Feng [36] believe that users are more likely to perform as gatekeepers. Similarly, Garimella et al. [37] explore the role of gatekeepers in the creation of political echo chambers, and they find these gatekeepers usually have a lower clustering coefficient. Although online gatekeepers consume information with different viewpoints, they usually share a certain viewpoint to strengthen the homogeneity of target communities and form closed fields of public opinions, which contribute to the dissemination of fake news [38]. Therefore, the utilization of gatekeepers to prevent the spread of fake news needs to be further studied.

(3) Media bias [39,40] It is one type of cognition bias, which means that journalists are unable to report news events objectively due to their partial opinions [41]. As Jamieson and Campbell [42] recognized, news outlets do not just report the facts but are often affected by governments or audiences' preference. Under the comprehensive impact of various aspects, media outlets often release claims without authorization, which provides an opportunity for the dissemination of fake news. Puglisi [43] estimates that *New York Times* may lean democratic. Furthermore, Gerber et al. [44] demonstrate that voters who read the *Washington Post* regularly are more likely to vote democratic candidates in the 2005 governor election in Virginia.Many researchers fear that unregulated media will have a major impact on our society [45], but competition among different media outlets can eliminate ideological bias in some cases [46].

(4) The spread of fake news [47,48] Many factors contribute to the dissemination of fake news, such as the cognitive limitation of readers [49], accessibility of social media platforms [50], and demographics of audiences [51]. Some studies have been conducted on the propagation features and structures of fake news. For example, DiFonzo et al. [52] find that social posts containing negative emotions are more likely to be spread. Guess et al. [53] state that conservatives are more likely to share fake news, and that older Facebook users (mainly over 65 years) spread about seven times as much fake news as the young during the 2016 US presidential election. Moreover, Budak et al. [54] demonstrate that the popularity of fake news is the result of news production and consumption.

3 Key research challenges and techniques

Having characterized the concepts of CogSec and reviewed some related studies, this section investigates some key research challenges and techniques of this research area, including *human-content cognition mechanism*, *social influence and opinion diffusion*,*fake news detection*, and *malicious bot detection*.

3.1 Human-content cognition mechanism

Understanding the mechanisms that people access, share, and repost online contents is critical to protect the CogSec, which relies on the knowledge from psychology, cognition science, and neuroscience [55].

(1) Personality, content sharing, and debunking The content sharing centric interpersonal social interactions enable information to be spread quickly [56]. Content sharing behaviors among social media users, such as publishing, reposting, and liking, will gradually affect the scope of news dissemination [57]. Several studies aim to learn the information-sharing mechanism on social media. For instance, Scholz et al. [58] present a neurocognitive framework to understand the online content sharing mechanism. Based on a news dataset from *New York Times*, they find that the core functions of sharing relate to both self-expression and social bonds. Hodas et al. [59] reveal a systematic link between the personality type and the social post topic. They observe that the user preference might be predicted from both personality and transitory mood state. In addition, Falk et al. [60] focus on the neural responses of information consumers' brains. They find that individuals are more able to spread their opinions to others, thus generating greater mentalizing-system activities in the initial process of information sharing.

Some works predict content reposts in social networks. For example, Hu et al. [61] predict the popularity of images and their diffusion paths based on Diffusion-LSTM, a memory-based deep recurrent neural network model. A combination of social features and image features is used to characterize individual reposting behaviors. Similarly, Zhang et al. [62] propose an attention-based deep neural network to integrate textual and social contextual information for retweeting prediction.

Different from the above works, Lewandowsky et al. [63] study the role of cognition factors in false information debunking. They further divide human cognition problems in the face of low-credibility information into four categories, including *continued influence effect*, *familiarity backfire effect*, *overkill backfire effect*, and *worldview backfire effect*, which provides a theoretical basis for CogSec protection.

(2) Neuroscience inhuman-content interaction Neuroscience has also been widely used in many research areas (e.g., healthcare [64], emotional retention [65], education [66], artificial intelligence [67], and economics [68]) related to human-machine interaction. As presented by Poldrack and Farah [69], the utilization of new tools, such as Electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), and Magnetoencephalography (MEG), for imaging and manipulating the brain will continue to advance our understanding of how the human brain cultivates to thought and action.

Regarding the CogSec, neuroscience has been used for understanding the human-content interaction. Some efforts have been conducted to understand or predict population-level behaviors (e.g., scoring and sharing on social media) based on a small number of individuals' neural responses. For example, Dmochowski et al. [70]find that the naturalistic stimulus, such as viewing multimodal contents, evokes highly reliable brain activities of the audience. Falk et al. [71] further conclude that the neural responses of a small group of users can be used to predict behaviors of large-scale populations. In particular, neural activities of the medial prefrontal region can predict the population response. Hasson et al. [72] report that the brains of different individuals tend to be highly consistent when viewing complex scenes (e.g., movies). Adolphs [73] identifies a series of neural structures involved in users' perceptions and judgments of content stimulus, and analyzes the decision-making mechanisms. In general, neuroscience provides a theoretical basis for understanding human-content interaction, which plays a practical role in the protection of public CogSec.

3.2 Social influence and opinion diffusion

The research on social influence and opinion diffusion in the fields of sociology and computational science has been lasting for a long time. For example, there have been numerous studies on opinion formation [74,75] and influence maximization models [76]. In this subsection, we review the related studies about the spread of fake news.

(1) Social influence and contagion The concept of social contagion has expanded from the initial epidemic transmission to information dissemination in social networks, such as political views [77], emotional changes [78], fashion trends [79], and financial decisions [80]. In addition, Morone and Makse [81] introduce the percolation theory [82] to the influential nodes discovery, and they find that a large number of weakly-connected (low-degree) nodes can be optimal influencers. Amati et al. [83] utilize the degree, closeness, betweenness, and PageRank-centrality of nodes in Dynamic Retweet Graph [84] to find the most influential users on Twitter. The work of Qiu et al. [85] proposes *DeepInf*, a deep learning-based social influence prediction framework, which learns users' latent social representation to evaluate their social influence by incorporating network embedding, graph convolution, and the attention mechanism.

Some studies concentrate on the information contagion and persuasion mechanisms in social networks. For instance, Ugander et al. [86] find that whether users will be infected depends on the structure of their interrelated components, rather than the actual size of the community. Therefore, different social environment represented by target users' neighbors can be considered as a driving mechanism of social contagion. Kramer et al. [87] prove that each user's mood can be affected by other users on Facebook, which provides an experimental basis for the study of massive-scale social influence and contagion mechanism. Abebe et al. [88] study information contagion from the perspective of people's psychological sensitivity to persuasion. They further propose a dynamic model of social opinions that comprehensively utilizes the maximization and minimization of crowd opinions to affect social opinions.

(2) Spreading models/mechanisms As Ratkiewicz et al. [89] once stated, the early stage of false information dissemination tends to show pathological patterns. Several works have studied the spreading mechanisms of online social posts, which guide the CogSec protection. For example, Friggeri et al. [90]track the propagation of thousands of fake posts on Facebook. They find that low-credibility information cascades run deeper into the social network than normal information sharing cascades. Peng et al. [91] find that users are more delight to hear the gossip about celebrities or their friends.

Several works have been conducted on opinion dynamics based on the influence mechanisms in social networks, which can be divided into discrete models [92,93] and continuous models [94]. For understanding the vulnerability of social networks and increasing users' resilience to fake news, Wang et al. [95] propose a multivariable jump-diffusion guidance framework, which models the dynamics of opinions and guides public opinions to the desired state. Martins et al. [96] propose an opinion diffusion model, *CODA*, in which users with different opinions are regarded as discrete variables and each opinion is modeled as a continuous opinion function. Target users decide whether to change their own opinions or not based on the Bayesian descriptions of their neighbors' opinions. Furthermore, Yang et al. [97] design a role-aware information diffusion model named *RAIN*, which characterizes the interaction between users' social roles and their influence on information dissemination.

3.3 Fake news detection

Due to the great impact of fake news on social stability, economic development, and political democracy, it is imperative to study practical automatic fake news detection methods. Recently, there have been several efforts on fake news detection, which can be divided into *content-based*, *social context-based*, and *deep learning-based* methods, as presented in Table 3.

(1) Content-based methods they often rely on distinct writing styles or textualfeatures in news content (e.g., lexical features, syntactic features, and semantic features). For example, Castillo et al. [98] calculate a series of linguistic features to evaluate the credibilityof tweets, including the average number of words, URL links, the number of positive words, etc. Potthast et al. [99] propose a meta-learning model to detect fake news, which takes advantage of the differences in the writing styles of real and fake news. Moreover, Hu et al. [100] also propose a spammer detection method based on sentiment analysis.

The content-based detection methods usually utilize extracted content features as the input of machine learning algorithms (such as SVM, RF) to train fake news classifiers. However, this type of methods has obvious shortcomings:

- It takes time and effort to extract a large number of textual features manually.
- It is vulnerable to deception because well-manipulated news is difficult to be recognized just by textual features.
- It analyzes the credibility of each event/post in isolation while it ignores the relevance of different events/posts.

Table 3 A summary of fake news detection methods

Work	Method type	Detection model	Model input			
			News content	Propagation	User Response	User/Website profiles
Castillo et al. [98]	Content-based	J48 Decision Tree etc.	√			√
Potthast et al. [99]		Meta-learning Approach	√			
Hu et al. [100]		Matrix Factorization	√	√		
Qazvinian et al. [101]		Rumor Retrieval, Belief Classification	√	√		
Kwon et al. [102]		Decision Tree, Random Forest, SVM	√	√		
Horne et al. [103]		SVM etc.	√			
Tacchini et al. [104]	Social Context-based	Logistic Regression, Boolean Crowdsourcing Algorithms		√	√	
Ma et al. [105]		Dynamic Series-Time Structure Model	√	√		√
Jin et al. [106]		Credibility Propagation Network	√		√	
Yang et al. [107]		Probabilistic Graphical Model	√		√	√
Gupta et al. [108]		Credibility Propagation Network	√	√		
Jin et al. [109]		Hierarchical Content Network	√	√		
Shu et al. [110]		——				√
Wu et al. [111]		Hybrid SVM	√	√		√
Jin et al. [112]		Epidemiological Model		√		
Liu et al. [113]		Information Propagation Model		√		√
Kim et al. [114]		Homogeneity-Based Transmissive Process		√		√
Ma et al. [115]		Propagation Tree via Kernel Learning		√		
Yu et al. [116]	Deep learning-based	CNN	√		√	
Ma et al. [117]		RNN	√		√	
Li et al. [118]		GRU	√		√	
Liu et al. [119]		CNN+GRU		√		
Runchansky et al. [120]		RNN	√		√	√
Jin et al. [121]		RNN	√		√	
Liu et al. [122]		Attention	√		√	
Guo et al. [123]		LSTM+Attention	√		√	√
Popat et al. [124]		LSTM+Attention	√			√

(2) Social context-based methods they mainly focus on the features of human-content interactions, such as user profiles, reposts, comments, stances, and likes. For example, Tacchini et al. [104] estimate that social media posts can be detected as false information utilizing users' *like* behaviors. Ma et al. [105] make use of the temporal patterns of social context features to detect online low-credibility posts. Moreover, Jin et al. [106] propose a credibility propagation network model for false information detection by mining supporting and opposing opinions in microblogs. Yang et al. [107] propose an unsupervised fake news detection model, which incorporates the authenticity of news, reputation of publishers, and viewpoints of users on the target news event.

Compared with the content-based methods, this type of method can extract more efficient features due to the integrated user attributes and interaction information. Besides, many works identify the dissemination patterns of fake news by building the spread process into specific network or tree structures [108,109]. The social context-based detection methods have relatively strong interpretability (compared with deep learning-based methods). Nevertheless, this kind of methods is difficult to prevent the dissemination of fake news due to the need for the whole news propagation path.

(3) Deep learning-based methods they mainly aim to learn latent representations of fake and real news for further detection. Existing deep learning-based detection methods mainly utilize convolutional neural networks (CNNs) [116] and recurrent neural networks (RNNs) [117]. For example, Li et al. [118] utilize the Bidirectional GRU model to detect online rumors, based on the observation that both the forward and backward sequences of social posts contain abundant interactive information. Liu et al. [119] find that there are obvious differences between the propagation patterns of true news and fake news, and they combine GRU (extracting global features) and CNN (extracting local features) to detect fake news. Ruchansky et al. [120] propose the RNN-based fake news detection model, which incorporates textual features of news, user response characteristics, and the credibility of source users.

Several works show that the performance (recall, precision, F1, etc.) of the deep learning-based methods is better than that of other methods in open datasets. This type of method can also achieve the early detection of fake news, but the biggest drawback of them is the poor interpretability of the deep learning models.

3.4 Malicious bot detection

The popularity and openness of social networks promote the emergence of social bots with certain autonomous decision-making abilities [125]. Like legitimate users, social bots can *make friends*, *post tweets*, *thumb up*, *chat* and so on through program control. de Lima Salgc and Berente [126] point out that about 8.5% of Twitter accounts are social bots, engaged in news, events, business communication, and other tasks. Most social bots provide convenience for users to exchange informa-

Table 4 A summary of bot detection methods

Work	Method type	Features
Boshmaf et al. [130]	Behavior-based	User profiles, Posting information
Haustein et al. [131]		Retweeting information
Gilani et al. [132]		Posting and Retweeting information
Yu et al. [133]		Second order statistical metrics
Varol et al. [134]		Social interactions
Thomas et al. [135]	Content-based	Features of related URL links
Egele et al. [136]		User profiles, message characteristics, behavioral profiles
Kudugunta and Ferrara [137]		User profiles, contextual features
Gao et al. [138]		Spam textual patterns
Messias et al. [139]	Influence-based	Posting information, influence scores
Abokhodair et al. [140]		Posting information, social structures, influence growth process, etc.
Freitas et al. [141]		Posting and retweeting information, social interactions

tion by automatically providing benign news and information, but there are also malicious social bots that can spread fake news and other harmful information [127–129]. Recently, a large number of malicious bot detection methods have been proposed, which can be categorized as *behavior-based*, *content-based*, and *influence-based* methods, as presented in Table 4.

(1) Behavior-based detection methods It is of great value to analyze and mine the behavior data of social bots in existing social networks. Boshmaf et al. [130]analyze the differences between social bots and human users in terms of the number of friends, posting interval, posting content and account attribute differences, and propose a random forest-based social bot detection method. Haustein et al. [131] analyze the differences between real Twitter users and social bots in retweeting scientific articles and find that social bots tend to be not selective in retweeting (involving topics, sources, etc.). Similarly, Gilani et al. [132] conduct a comparative study on the behaviors of human and social bots in posting and retweeting on Twitter, and find that social bots play a very important role in information transmission, despite their weak overall influence. Yu et al. [133] find that it is difficult to detect the behavior mimicking attacks of bots by using the traditional first-order (e.g., the mathematical expectation) and second-order (e.g., the variance) statistics, so they propose the new behavior statistical measurement method based on cross-entropy for bot detection. Besides, Varol et al. [134] find that compared with human users, the interaction object selection of social bots is more arbitrary and that there are fewer bidirectional connections between social bots and human users.

(2) Content-based detection methods This type of method concentrates on determining whether the message published by the user is malicious. Generally, whether the URL in the message content points to the malicious page can be used to determine whether the original publisher is a malicious social bot. For instance, Thomas et al. [135] propose a real-time URL detection scheme, which extracts features of related URL pages by visiting those published URLs. What's more, social bots can be detected through changes in the message content features. For example, Egele et al. [136] extract seven content features to model the message and then detect social bots by judging whether the message published later deviates from the created model. Kudugunta and Ferrara [137] propose an LSTM-based bot detection method, which incorporates contextual features and accounts' metadata for improving bot detection accuracy. Gao et al. [138] find that 63% of the text content of spam messages in Twitter are generated based on templates,and then they propose the social bot detection framework, *Tangram*, which divides malicious posts into fields, generates matching templates, and detects more malicious social bots.

(3) Influence-based detection methods This kind of method detects social bots from the perspective of social influence. For example, Messias et al. [139] conduct a comparative study on analyzing the influence of social bots, and propose their malicious behavior identification strategies, including regular posting tweets on a certain hot topic, different posting intervals, and attribute integrity. Similarly, Abokhodair et al. [140] analyze the posting behaviors, social structures, group behavior characteristics and influence growth processes of the social bot network. Finally, they find that more human-like behaviors can improve social bot influence. Freitas et al. [141] create 120 different attributes (gender, occupation, etc.) and behavior strategies (posting, reposting and interaction) of the social bots to characterize their infiltration process, and they find that about 20% of the social bots gain more than 100 followers by means of highly active interaction and posting behavior.

4 Open issues and future research directions

Although preliminary efforts have been made in the field of CogSec, there are still many research challenges that remain and need to be tackled in the future, some of which are discussed below.

(1) The cognition mechanism of fake news Regarding the CogSec protection, the first thing is to understand the human cognition mechanism of fake news. Acerbi [13] finds that the reason why fake news can be successfully spread is that it conforms to the general cognitive preference of the public, which provides the theoretical guidance for preventing the spread of fake news. With the rapid development of neuroscience, several studies have been conducted on the cognition mode of the human brain [142,143]. For example, Lewandowsky et al. [144] raise the problem of "technocognition" and summarize how fake news affect society negatively. The work of Arapakis et al. [145] proposes a measurement model for evaluating the change in users' interest when reading news, which is based on EEG recordings of human neural activities. All in all, the research on the cognition mechanism of fake news correlates to multiple disciplines such as psychology, neuroscience and cognitive science. We still need to explore specific cognition problems, partially summarized as follows:

- The influence of an individual's social cognition on large-scale social behaviors.
- The common features of fake news satisfying users' cognitive preferences.
- The effect of fake content stimulations (texts, pictures, audios or videos) on specific parts of the human brains.
- The impact of social interactions on an individual's cognition behaviors.
- The change of users' cognition characteristics with the dissemination of fake news.

(2) Explainable fake news debunking Existing automatic fake news detection models usually give the ultimate test results, without much credible decision-making explanations. However, the explanatory contents in fake news debunking and the transparency of detection models are essential, which contribute to enhancing the user trust in detection results and making effective use of human-machine intelligence. Some studies utilize the attention mechanism and graph models for explainable fake news detection. For instance, Chen et al. [146] find that users tend to make different comments in diverse rumor spreading processes, and thus they propose an RNN-based rumor detection model with attention mechanism for selecting key debunking information.Shu et al. [147] propose an interpretable detection framework called *dEFEND*, which utilizes a joint attention network of sentences and comments to capture top-k explainable comments for fake news detection. Moreover, Gad-Elrab et al. [148] propose a framework for generating explanations of candidate facts, incorporating knowledge graphs and texts, which provides reference for fake news detection. Nguyen et al. [149] utilize a probabilistic graphical model to represent the stance of news articles, reputation of news source, and the reliability of news annotators. Afterwards, they propose a variational inference algorithm to identify the credibility of the news event. In general, explainable fake news detection needs to explore more practical models with the development of interpretable machine learning (IML) [150,151], and other visualization techniques.

(3) Efficient and robust security countermeasures In addition to detecting fake news and malicious accounts in social networks, CogSec also pays attention to robust security protection measures to protect human cognition security and rapidly prevent the dissemination of fake news in the network communities:

- Information filtering based on online gatekeepers/guardians. As gatekeepers (or guardians) could improve the homogeneity of the communities, they can strengthen the dissemination of information. For example, Vo and Lee [152] study how guardians prevent users from continuing to share fake news after it was debunked. To enhance users' ability to distinguish fake news, they further propose a fact-checking based URL recommendation model to encourage users to actively participate in the fake news debunking.
- Fact-checking based on human-machine cooperation.In the future CogSec protection work, we need to make adequate use of crowd intelligence and hybrid human-machine intelligence. For instance, Kim et al. [153] propose an online fact-checking algorithm called *Curb*, which could decide when to send specific news content to human experts. Bhattacharjee et al. [154] put forward a news veracity detection model based on active learning, which utilizes human experts to label samples to improve the learning performance of the model. Therefore, human-machine cooperation will be a key and practical measure of CogSec protection.

(4) Early detection of fake news Information on social networks usually has a short life span, less than three days on average, and fake news always spreads like viruses in a few minutes [155]. However, detection methods based on aggregation features (e.g., propagation characteristics, etc.) are difficult to achieve satisfactory performance on early detection. Some works attempt to identify fake news at their early spreading stage by implicit signals, as shown in Table 5. For example, Zhao et al. [156] find that the questioning and opposing expressions in user reviews contribute to the early detection of rumors. Sampson et al. [157] utilize implicit linkages for acquiring additional information from several related events to deal with the problem that less data is available for early detection of fake news. Besides, some other techniques can be introduced to explore the early detection of fake news, such as transfer learning [161,162], zero-shot learning [163,164], and meta-learning [165–167].

Table 5 A comparison of early detection methods of fake news

Research work	Model/Approach	Model inputs
Kim et al. [153]	Stochastic differential equations to model user flagging procedure	Features from users' flags
Bhattacharjee et al. [154]	Active learning model+CNN	Features from the topics of news content
Zhao et al. [156]	Regular expressions matching and twitters clustering	Features from user comments (questions or inquiries)
Sampson et al. [157]	Conversation-centric approach based on machine learning classification algorithms	Features from hashtag linkages and web linkages, news content, user profiles, network propagation
Liu et al. [158]	Real-time algorithmic veracity prediction approach	Features from user comments, witness accounts, underlying belief, network propagation, etc.
Qian et al. [159]	Conditional Variational Autoencoder	Features from news content, user response knowledge
Tachiatschek et al. [160]	Bayesian inference for modeling user flagging procedure	Features from users' flags

In addition to the detection methods based on news content and user response knowledge, researchers are also actively exploring other methods to prevent the dissemination of fake news with the truth. For example, Ginsca and Weninger [168] review the evaluation methods of news credibility from the perspective of information retrieval (IR). Shi et al. [169] utilize the knowledge graph (KG) to automatically identify fake news. Concretely, the fact-checking framework understands the meaning of the statement by extracting discriminant meta paths from the knowledge graph, and then evaluates its credibility by combining the entity and predicate type information.

(5) Fast refutation of fake news For the crowd CogSec protection, in addition to detecting fake news as soon as possible and expanding the influence of fact-checking information, it also needs to utilize appropriate strategies to debunk the fake news. Nyhan and Reifler [170] find that the refutation of fake news may backfire, which will deepen users' trust in the low credibility information. Consequently, the strategies to rapidly prevent the dissemination of fake news with the truth require careful consideration of the time of intervention, the content of refutation information, and the selection of refuters. Besides, Bordia et al. [171] use the Elaboration Likelihood Model (ELM) to test the impact of the denial information quality, user relevance, and source reputation on the effectiveness of debunking information, which provides a theoretical basis for the fast refutation of fake news.

Some works study the ways of rapid fake news refutation. For example, Tanaka et al. [172] observe that fact-checking tweets with short URLs are easier to be shared. Ozturk et al. [173] find that displaying inaccurate information along with denial information on Twitter contributes to reducing the continued spread of fake news. Similarly, Alemanno [174] believes that the original fake information should not be deleted when debunking, but should be surrounded by several relevant articles. This method could provide users with more context and views, which enables users to automatically eliminate potential fake articles in their social media. However, what kind of content should be selected for refutation, and when to carry out fake news refutation need more in-depth researches.

(6) The influence maximization of fact-checking information When fake news is detected, the fact-checking information needs to be released to the social network as soon as possible. Therefore, the maximization of information influence is still noteworthy, and social contagion mechanism could provide reference for this issue. Social contagion is a common phenomenon in human society [175], which contributes to opinion dynamics, behavior shaping, and cognition biases in social networks. Some works pay attention to modeling the contagion and propagation of information in social networks, which provide a reference for maximizing the impact of facts verification information. For instance, Chang et al. [176] explore how social media marketing persuades users to share information to maximize information coverage. Huang et al. [177] propose a social contagion model, which introduces a persuasion mechanism into the threshold model. They then find that persuasion mechanism improves the impact of information cascade in social networks and that the effect of persuasion is often more significant in heterogenous social networks than in homogeneous networks.In the future, there are still some issues to be further studied:

- The research on novel information communication theories which considers users' cognition preferences, information timeliness, and users' social roles, etc.
- The timely discovery methods of influential users on social networks to maximize the impact of information dissemination.
- The mechanisms of maximizing the influence of the real information on the audience after the fake news is exposed.

5 Conclusion

In the context of the spread of fake news in social networks, we present a novel research issue, named Cognitive Security (CogSec). To characterize the CogSec, we propose some relevant definitions and review several related findings, including echo chambers, online gatekeepers, media bias, etc. We further investigate the key research challenges and techniques of CogSec, which can be categorized into human-content cognition mechanisms, social influence and opinion diffusion, fake news detection, and malicious bot detection. The study of CogSec is still at its early stage,and there are still numerous challenges and open issues to be addressed by AI researchers, social and neuroscience scientists, as well as security engineers.

Acknowledgements This work was partially supported by the National Key R&D Program of China (2019QY0600), the National Natural Science Foundation of China (Grant Nos. 61772428, 61725205, 61902320, 61925203, U1636210), and Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

1. Iyengar S, Massey D S. Scientific communication in a post-truth society. Proceedings of the National Academy of Sciences, 2019, 116(16): 7656–7661
2. Fernandez M, Alani H. Online misinformation: challenges and future directions. In: Proceedings of the Web Conference. 2018, 595–602
3. Guess A, Nyhan B, Reifler J. Selective exposure to misinformation: evidence from the consumption of fake news during the 2016 US presidential campaign. European Research Council, 2018, 9(3): 4–52
4. Zhou X, Zafarani R, Shu K, Liu H. Fake news: fundamental theories, detection strategies and challenges. In: Proceedings of the ACM International Conference on Web Search and Data Mining. 2019, 836–837
5. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science, 2018, 359(6380): 1146–1151
6. Lazer D M J, Baum M A, Benkler Y, Berinsky A J, Greenhill K M, Menczer F, Metzger M J, Nyhan B, Pennycook G, Rothschild D. The science of fake news. Science, 2018, 359(6380): 1094–1096
7. Ruths D. The misinformation machine. Science, 2019, 363(6425): 348
8. Qiu X, Oliveira D F M, Shirazi A S, Flammini A, Menczer F. Limited individual attention and online virality of low-quality information. Nature Human Behaviour, 2017, 1(7): 0132
9. Bakdash J, Sample C, Rankin M, Kantarcioglu M, Holmes J, Kase S, Zaroukian E, Szymanski B. The future of deception: machine-generated and manipulated images, video, and audio?. In: Proceedings of IEEE International Workshop on Social Sensing. 2018
10. Floridi L. Artificial intelligence, deepfakes and a future of ectypes. Philosophy & Technology, 2018, 31(3): 317–321

11. Yang X, LiY, Lyu S. Exposing deep fakes using inconsistent head poses. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2019, 8261–8265

12. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting world leaders against deep fakes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019, 38–45

13. Acerbi A. Cognitive attraction and online misinformation. Palgrave Communications, 2019, 5(1): 15–21

14. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R. Detection and resolution of rumours in social media: a survey. ACM Computing Surveys (CSUR), 2018, 51(2): 32–67

15. Kumar S, West R, Leskovec J. Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of International Conference on World Wide Web. 2016, 591–602

16. 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. 2017, 647–653

17. Wu L, Morstatter F, Hu X, Liu H. Big Data in Complex and Social Networks. 1st ed. London: Chapman and Hall/CRC, 2016

18. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: a data mining perspective. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22–36

19. Zhou X, Zafarani R. Fake news: a survey of research, detection methods, and opportunities. 2018, arXiv preprint arXiv:1812.00315

20. Jr S B, Campos G F, Tavares G M, Igawa R A, Jr M L P, Guido R C. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(1s): 26–42

21. Macskassy S A. On the study of social interactions in twitter. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. 2012, 226–233

22. Forouzan B A. Cryptography & Network Security. New York: McGraw-Hill, 2007

23. Greenstadt R, Beal J. Cognitive security for personal devices. In: Proceedings of ACM Workshop on AISec. 2008, 27–30

24. Kinsner W. Towards cognitive security systems. In: Proceedings of IEEE International Conference on Cognitive Informatics and Cognitive Computing. 2012, 539

25. DiFranzo D, Gloria M J K. Filter bubbles and fake news. ACM Crossroads, 2017, 23(3): 32–35

26. Vaccari C. From echo chamber to persuasive device? rethinking the role of the Internet in campaigns. New Media & Society, 2013, 15(1): 109–127

27. Flaxman S, Goel S, Rao J M. Filter bubbles, echo chambers, and online news consumption. Public Opinion Quarterly, 2016, 80(S1): 298–320

28. Flintham M, Karner C, Bachour K, Creswick H, Gupta N, Moran S. Falling for fake news: investigating the consumption of news via social media. In: Proceedings of CHI Conference on Human Factors in Computing Systems. 2018, 376–385

29. Barberá P, Jost J T, Nagler J, Tucker J A, Bonneau R. Tweeting from left to right: is online political communication more than an echo chamber?. Psychological Science, 2015, 26(10): 1531–1542

30. Bessi A. Personality traits and echo chambers on facebook. Computers in Human Behavior, 2016, 65: 319–324

31. Zajonc R B. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology, 1968, 9(2p2): 1–27

32. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley H E, Quattrociocchi W. The spreading of misinformation online. Proceedings of the National Academy of Sciences, 2016, 113(3): 554–559

33. Singer J B. Online journalists: foundations for research into their changing roles. Journal of Computer-Mediated Communication, 1998, 4(1): JCMC412

34. Nielsen R K. News Media, Search Engines and Social Networking Sites as Varieties of Online Gatekeepers. Rethinking Journalism Again. London: Routledge, 2016

35. Bui C L. How online gatekeepers guard our view-news portals' inclusion and ranking of media and events. Global Media Journal, 2010, 9(16): N_A

36. Xu W, Feng M. Talking to the broadcasters on twitter: networked gatekeeping in twitter conversations with journalists. Journal of Broadcasting & Electronic Media, 2014, 58(3): 420–437

37. Garimella K, De Francisci Morales G, Gionis A, Mathioudakis M. Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. In: Proceedings of the World Wide Web Conference. 2018, 913–922

38. DiFonzo N. Ferreting facts or fashioning fallacies? factors in rumor accuracy. Social and Personality Psychology Compass, 2010, 4(11): 1124–1137

39. Entman R M. Framing bias: media in the distribution of power. Journal of Communication, 2007, 57(1): 163–173

40. Chiang C F, Knight B. Media bias and influence: evidence from newspaper endorsements. The Review of Economic Studies, 2011, 78(3): 795–820

41. Iyengar S, Kinder D R. News That Matters: Television and American opinion. Palo Alto: University of Chicago Press, 2010

42. Jamieson K H, Campbell K K. Interplay of Influence: News, Advertising, Politics and the Internet Age (with InfoTrac). Belmont: Wadsworth Publishing, 2005

43. Puglisi R. Being the new york times: the political behaviour of a newspaper. The BE Journal of Economic Analysis & Policy, 2011, 11(1): 1–48

44. Gerber A S, Karlan D, Bergan D. Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. American Economic Journal: Applied Economics, 2009, 1(2): 35–52

45. Ribeiro F N, Henrique L, Benevenuto F, Chakraborty A, Kulshrestha J, Babaei M, Gummadi K P. Media bias monitor: quantifying biases of social media news outlets at large-scale. In: Proceedings of the 12th International AAAI Conference on Web and Social Media. 2018, 290–299

46. Budak C, Goel S, Rao J M. Fair and balanced? quantifying media bias through crowdsourced content analysis. Public Opinion Quarterly, 2016, 80(S1): 250–271

47. Bovet A, Makse H A. Influence of fake news in twitter during the 2016 US presidential election. Nature Communications, 2019, 10(1): 7–20

48. Kucharski A. Post-truth: study epidemiology of fake news. Nature, 2016, 540(7634): 525

49. DiFonzo N, Beckstead J W, Stupak N, Walders K. Validity judgments of rumors heard multiple times: the shape of the truth effect. Social Influence, 2016, 11(1): 22–39

50. Ngai E W T, Tao S S C, Moon K K L. Social media research: theories, constructs, and conceptual frameworks. International Journal of Information Management, 2015, 35(1): 33–44

51. Allcott H, Gentzkow M. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017, 31(2): 211–236

52. DiFonzo N, Bourgeois M J, Suls J, Homan C, et al. Rumor clustering, consensus, and polarization: dynamic social impact and self-organization of hearsay. Journal of Experimental Social Psychology, 2013, 49(3): 378–399

53. Guess A, Nagler J, Tucker J. Less than you think: prevalence and predictors of fake news dissemination on facebook. Science Advances, 2019, 5(1): eaau4586

54. Budak C. What happened? the spread of fake news publisher content during the 2016 US presidential election. In: Proceedings of the World Wide Web Conference. 2019, 139–150
55. Poldrack R A, Farah M J. Progress and challenges in probing the human brain. Nature, 2015, 526(7573): 371–382
56. Csibra G, Gergely G. Natural pedagogy as evolutionary adaptation. Philosophical Transactions of the Royal Society B: Biological Sciences, 2011, 366(1567): 1149–1157
57. Cappella J N, Kim H S, Albarracín D. Selection and transmission processes for information in the emerging media environment: psychological motives and message characteristics. Media Psychology, 2015, 18(3): 396–424
58. Scholz C, Baek E C, O'Donnell M B, Kim H S, Cappella J N, Falk E B. A neural model of valuation and information virality. Proceedings of the National Academy of Sciences, 2017, 114(11): 2881–2886
59. Hodas N O, Butner R. How a user's personality influences content engagement in social media. In: Proceedings of International Conference on Social Informatics. 2016, 481–493
60. Falk E B, Morelli S A, Welborn B L, Dambacher K, Lieberman M D. Creating buzz: the neural correlates of effective message propagation. Psychological Science, 2013, 24(7): 1234–1242
61. Hu W, Singh K K, Xiao F, Han J, Chuah C N, Lee Y J. Who will share my image?: predicting the content diffusion path in online social networks. In: Proceedings of ACM International Conference on Web Search and Data Mining. 2018, 252–260
62. Zhang Q, Gong Y, Wu J, Huang H, Huang X. Retweet prediction with attention-based deep neural network. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016, 75–84
63. Lewandowsky S, Ecker U K, Seifert C M, Schwarz N, Cook J. Misinformation and its correction: continued influence and successful debiasing. Psychological Science in the Public Interest, 2012, 13(3): 106–131
64. Davidson R J, Pizzagalli D, Nitschke J B, Putnam K. Depression: perspectives from affective neuroscience. Annual Review of Psychology, 2002, 53(1): 545–574
65. LaBar K S, Cabeza R. Cognitive neuroscience of emotional memory. Nature Reviews Neuroscience, 2006, 7(1): 54–64
66. Howard-Jones P A. Neuroscience and education: myths and messages. Nature Reviews Neuroscience, 2014, 15(12): 817–824
67. Hassabis D, Kumaran D, Summerfield C, Botvinick M. Neuroscience-inspired artificial intelligence. Neuron, 2017, 95(2): 245–258
68. Camerer C, Loewenstein G, Prelec D. Neuroeconomics: how neuroscience can inform economics. Journal of Economic Literature, 2005, 43(1): 9–64
69. Poldrack R A, Farah M J. Progress and challenges in probing the human brain. Nature, 2015, 526(7573): 371–382
70. Dmochowski J P, Bezdek M A, Abelson B P, Johnson J S, Schumacher E H, Parra L C. Audience preferences are predicted by temporal reliability of neural processing. Nature Communications, 2014, 5(1): 1–9
71. Falk E B, Berkman E T, Lieberman M D. From neural responses to population behavior: neural focus group predicts population-level media effects. Psychological Science, 2012, 23(5): 439–445
72. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. Science, 2004, 303(5664): 1634–1640
73. Adlolphs R. Cognitive neuroscience of human social behavior. Nature Reviews Neuroscience, 2003, 4: 165–178
74. DeGroot M H. Reaching a consensus. Journal of the American Statistical Association, 1974, 69(345): 118–121
75. Cialdini R B, Petty R E, Cacioppo J T. Attitude and attitude change. Annual Review of Psychology, 1981, 32(1): 357–404
76. Kempe D, Kleinberg J, TardosÉ. Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003, 137–146
77. Rozin P, Royzman E B. Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 2001, 5(4): 296–320
78. Hatfield E, Cacioppo J T, Rapson R L. Emotional contagion. Current Directions in Psychological Science, 1993, 2(3): 96–100
79. Argo J J, Dahl D W, Morales A C. Positive consumer contagion: responses to attractive others in a retail context. Journal of Marketing Research, 2008, 45(6): 690–701
80. Allen F, Gale D. Financial contagion. Journal of Political Economy, 2000, 108(1): 1–33
81. Morone F, Makse H A. Influence maximization in complex networks through optimal percolation. Nature, 2015, 524(7563): 65–147
82. Moore C, Newman M E J. Epidemics and percolation in small-world networks. Physical Review E, 2000, 61(5): 5678–5683
83. Amati G, Angelini S, Gambosi G, Rossi G, Vocca P. Influential users in Twitter: detection and evolution analysis. Multimedia Tools and Applications, 2019, 78(3): 3395–3407
84. Amati G, Angelini S, Capri F, Gambosi G, Rossi G, Vocca P. Twitter temporal evolution analysis: comparing event and topic driven retweet graphs. IADIS International Journal on Computer Science & Information Systems, 2016, 11(2): 155–162
85. Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J. Deepinf: social influence prediction with deep learning. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018, 2110–2119
86. Ugander J, Backstrom L, Marlow C, Kleinberg J. Structural diversity in social contagion. Proceedings of the National Academy of Sciences, 2012, 109(16): 5962–5966
87. Kramer A D I, Guillory J E, Hancock J T. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 2014, 111(24): 8788–8790
88. Abebe R, Kleinberg J, Parkes D, Tsourakakis C E. Opinion dynamics with varying susceptibility to persuasion. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018, 1089–1098
89. Ratkiewicz J, Conover M, Meiss M, Goncalves B, Patil S, Flammini A, Menczer F. Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th International Conference Companion on World Wide Web. 2011, 249–252
90. Friggeri A, Adamic L, Eckles D, Cheng J. Rumor cascades. In: Proceedings of International AAAI Conference on Weblogs and Social Media. 2014, 101–110
91. Peng X, Li Y, Wang P, Mo L, Chen Q. The ugly truth: negative gossip about celebrities and positive gossip about self entertain people in different ways. Social Neuroscience, 2015, 10(3): 320–336
92. Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978, 83(6): 1420–1443
93. Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks. In: Proceedings of International Colloquium on Automata, Languages, and Programming. 2005, 1127–1138
94. Chatterjee S, Seneta E. Towards consensus: some convergence theorems on repeated averaging. Journal of Applied Probability, 1977, 14(1): 89–97
95. Wang Y, Theodorou E, Verma A, Song L. Steering opinion dynamics in information diffusion networks. 2016, arXiv preprint arXiv:1603.09021
96. Martins A C R. Continuous opinions and discrete actions in opinion dynamics problems. International Journal of Modern Physics C, 2008, 19(4): 617–624

97. Yang Y, Tang J, Leung C W K, Sun Y, Chen Q, Li J, Yang Q. RAIN: social role-aware information diffusion. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. 2015, 367–373

98. Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of International Conference on World Wide Web. 2011, 675–684

99. Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B. A stylometric inquiry into hyperpartisan and fake news. 2017, arXiv preprint arXiv:1702.05638

100. Hu X, Tang J, Gao H, Liu H. Social spammer detection with sentiment information. In: Proceedings of IEEE International Conference on Data Mining. 2014, 180–189

101. Qazvinian V, Rosengren E, Radev D R, Mei Q. Rumor has it: identifying misinformation in microblog. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011, 1589–1599

102. Kwon S, Cha M, Jung K, Chen W, Wang Y. Prominent features of rumor propagation in online social media. In: Proceedings of IEEE International Conference on Data Mining. 2013, 1103–1108

103. Horne B D, Adali S. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017, 759–766

104. Tacchini E, Ballarin G, Della Vedova M L, Moret S, de Alfaro L. Some like it hoax: automated fake news detection in social networks. 2017, arXiv preprint arXiv:1704.07506

105. Ma J, Gao W, Wei Z, Lu Y, Wong K F. Detect rumors using time series of social context information on microblogging websites. In: Proceedings of ACM International on Conference on Information and Knowledge Management. 2015, 1751–1754

106. Jin Z, Cao J, Zhang Y, Luo J. News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016, 2972–2978

107. Yang S, Shu K, Wang S, Gu R, Wu F, Liu H. Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. 2019, 5644–5651

108. Gupta M, Zhao P, Han J. Evaluating event credibility on twitter. In: Proceedings of the SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics. 2012, 153–164

109. Jin Z, Cao J, Jiang Y G, Zhang Y. News credibility evaluation on microblog with a hierarchical propagation model. In: Proceedings of IEEE International Conference on Data Mining. 2014, 230–239

110. Shu K, Wang S, Liu H. Understanding user profiles on social media for fake news detection. In: Proceedings of IEEE Conference on Multimedia Information Processing and Retrieval. 2018, 430–435

111. Wu K, Yang S, Zhu K Q. False rumors detection on sinaweibo by propagation structures. In: Proceedings of the 31st IEEE International Conference on Data Engineering. 2015, 651–662

112. Jin F, Dougherty E, Saraf P, Cao Y, Ramakrishnan N. Epidemiological modeling of news and rumors on twitter. In: Proceedings of the 7th Workshop on Social Network Mining and Analysis. 2013, 1–9

113. Liu Y, Xu S. Detecting rumors through modeling information propagation networks in a social media environment. IEEE Transactions on Computational Social Systems, 2016, 3(2): 46–62

114. Kim J, Kim D, Oh A. Homogeneity-based transmissive process to model true and false news in social networks. In: Proceedings of ACM International Conference on Web Search and Data Mining. 2019, 348–356

115. Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017, 708–717

116. Yu F, Liu Q, Wu S, Wang L, Tan T. A convolutional approach for misinformation identification. In: Proceedings of International Joint Conference on Artificial Intelligence. 2017, 3901–3907

117. Ma J, Gao W, Mitra P, Kwon S, Jansen B J, Wong K F, Cha M. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of International Joint Conference on Artificial Intelligence. 2016, 3818–3824

118. Li L, Cai G, Chen N. A rumor events detection method based on deep bidirectional GRU neural network. In: Proceedings of the 3rd IEEE International Conference on Image, Vision and Computing. 2018, 755–759

119. Liu Y, Wu Y F B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018, 354–361

120. Ruchansky N, Seo S, Liu Y. CSI: a hybrid deep model for fake news detection. In: Proceedings of ACM on Conference on Information and Knowledge Management. 2017, 797–806

121. Jin Z, Cao J, Guo H, Zhang Y, Luo J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of ACM International Conference on Multimedia. 2017, 795–816

122. Liu Q, Yu F, Wu S, Wang L. Mining significant microblogs for misinformation identification: an attention-based approach. ACM Transactions on Intelligent Systems and Technology, 2018, 9(5): 50–67

123. Guo H, Cao J, Zhang Y, Guo J, Li J. Rumor detection with hierarchical social attention network. In: Proceedings of ACM International Conference on Information and Knowledge Management. 2018, 943–951

124. Popat K, Mukherjee S, Yates A, Weikum G. DeClarE: debunking fake news and false claims using evidence-aware deep learning. 2018, arXiv preprint arXiv:1809.06416

125. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. Communications of the ACM, 2016, 59(7): 96–104

126. de Lima Salge C A, Berente N. Is that social bot behaving unethically?. Communications of the ACM, 2017, 60(9): 29–31

127. Chu Z, Gianvecchio S, Wang H, Jajodia S. Detecting automation of twitter accounts: are you a human, bot, or cyborg?. IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 811–824

128. Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M. Design and analysis of a social botnet. Computer Networks, 2013, 57(2): 556–578

129. Yu S, Gu G, Barnawi A, Guo S, Stojmenovic I. Malware propagation in large-scale networks. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(1): 170–179

130. Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M. The socialbot network: when bots socialize for fame and money. In: Proceedings of the 27th Annual Computer Security Applications Conference. 2011, 93–102

131. Haustein S, Bowman T D, Holmberg K, Tsou A, Sugimoto C R, Larivière V. Tweets as impact indicators: examining the implications of automated “bot” accounts on twitter. Journal of the Association for Information Science and Technology, 2016, 67(1): 232–238

132. Gilani Z, Farahbakhsh R, Tyson G, Wang L, Crowcroft J. An in-depth characterisation of bots and humans on Twitter. 2017, arXiv preprint arXiv:1704.01508

133. Yu S, Guo S, Stojmenovic I. Fool me if you can: mimicking attacks and anti-attacks in cyberspace. IEEE Transactions on Computers, 2013, 64(1): 139–151

134. Varol O, Ferrara E, Davis C A, Menczer F, Flammini A. Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017, 280–289

135. Thomas K, Grier C, Ma J, Paxson V, Song D. Design and evaluation of a real-time url spam filtering service. In: Proceedings of IEEE Symposium on Security and Privacy. 2011, 447–462

136. Egele M, Stringhini G, Kruegel C, Vigna G. Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable and Secure Computing, 2015, 14(4): 447–460
137. Kudugunta S, Ferrara E. Deep neural networks for bot detection. Information Sciences, 2018, 467: 312–322
138. Gao H, Yang Y, Bu K, Chen Y, Downey D, Lee K, Choudhary A. Spam ain't as diverse as it seems: throttling OSN spam with templates underneath. In: Proceedings of the 30th Annual Computer Security Applications Conference. 2014, 76–85
139. Messias J, Schmidt L, Oliveira R A R D, Souza F B D. You followed my bot! transforming robots into influential users in twitter. Peer-Reviewed Journal on the Internet, 2013, 18(7–1): 1–14
140. Abokhodair N, Yoo D, McDonald D M. Dissecting a social botnet: growth, content and influence in twitter. In: Proceedings of ACM Conference on Computer Supported Cooperative Work & Social Computing. 2015, 839–851
141. Freitas C, Benevenuto F, Ghosh S, Veloso A. Reverse engineering socialbot infiltration strategies in twitter. In: Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2015, 25–32
142. Guixeres J, Bigné E, AusínAzofra J M, Alcañiz Raya M, Colomer Granero A, Fuentes Hurtado F, Naranjo Ornedo V. Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. Frontiers in Psychology, 2017, 8: 1808–1821
143. Yılmaz B, Korkmaz S, Arslan D B, Güngör E, Asyalı M H. Like/dislike analysis using EEG: determination of most discriminative channels and frequencies. Computer Methods and Programs in Biomedicine, 2014, 113(2): 705–713
144. Lewandowsky S, Ecker U K H, Cook J. Beyond misinformation: understanding and coping with the "post-truth" era. Journal of Applied Research in Memory and Cognition, 2017, 6(4): 353–369
145. Arapakis I, Barreda-Angeles M, Pereda-Baños A. Interest as a proxy of engagement in news reading: spectral and entropy analyses of EEG activity patterns. IEEE Transactions on Affective Computing, 2017, 10(1): 100–114
146. Chen T, Li X, Yin H, Zhang J. Call attention to rumors: deep attention based recurrent neural networks for early rumor detection. In: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018, 40–52
147. Shu K, Cui L, Wang S, Lee D, Liu H. dEFEND: explainable fake news detection. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019, 395–405
148. Gad-Elrab M H, Stepanova D, Urbani J, Weikum G. ExFaKT: a framework for explaining facts over knowledge graphs and text. In: Proceedings of ACM International Conference on Web Search and Data Mining. 2019, 87–95
149. Nguyen A T, Kharosekar A, Lease M, Wallace B. An interpretable joint graphical model for fact-checking from crowds. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018, 1511–1518
150. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Communications of the ACM, 2019, 63(1): 68–77
151. MurdochW J, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 2019, 116(44): 22071–22080
152. Vo N, Lee K. The rise of guardians: fact-checking url recommendation to combat fake news. In: Proceedings of ACM SIGIR Conference on Research & Development in Information Retrieval. 2018, 275–284
153. Kim J, Tabibian B, Oh A, Schölkopf B, Gomez-Rodriguez M. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In: Proceedings of ACM International Conference on Web Search and Data Mining. 2018, 324–332
154. Bhattacharjee S D, Talukder A, Balantrapu B V. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In: Proceedings of IEEE International Conference on Big Data. 2017, 556–565
155. Cao J, Guo J, Li X, Jin Z, Guo H, Li J. Automatic rumor detection on microblogs: a survey. 2018, arXiv preprint arXiv:1807.03505
156. Zhao Z, Resnick P, Mei Q. Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of International Conference on World Wide Web. 2015, 1395–1405
157. Sampson J, Morstatter F, Wu L, Liu H. Leveraging the implicit structure within social media for emergent rumor detection. In: Proceedings of ACM International on Conference on Information and Knowledge Management. 2016, 2377–2382
158. Liu X, Nourbakhsh A, Li Q, Fang R, Shah S. Real-time rumor debunking on twitter. In: Proceedings of ACM International on Conference on Information and Knowledge Management. 2015, 1867–1870
159. Qian F, Gong C, Sharma K, Liu Y. Neural user response generator: fake news detection with collective user intelligence. In: Proceedings of International Joint Conference on Artificial Intelligence. 2018, 3834–3840
160. Tschiatschek S, Singla A, Gomez Rodriguez M, Merchant A, Krause A. Fake news detection in social networks via crowd signals. In: Proceedings of the Web Conference. 2018, 517–524
161. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Proceedings of International Conference on Artificial Neural Networks. 2018, 270–279
162. Li Z, Wei Y, Zhang Y, Yang Q. Hierarchical attention transfer network for cross-domain sentiment classification. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018, 5852–5859
163. Wang W, Zheng V W, Yu H, Miao C. A survey of zero-shot learning: settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1–37
164. Socher R, Ganjoo M, Manning C D, Ng A. Zero-shot learning through cross-modal transfer. In: Proceedings of Advances in Neural Information Processing Systems. 2013, 935–943
165. Yao H, Liu Y, Wei Y, Tang X, Li Z. Learning from multiple cities: a meta-learning approach for spatial-temporal prediction. In: Proceedings of The World Wide Web Conference. 2019, 2181–2191
166. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of International Conference on Machine Learning-Volume 70. 2017, 1126–1135
167. Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. In: Proceedings of International Conference on Machine Learning. 2016, 1842–1850
168. Ginsca A L, Popescu A, Lupu M. Credibility in information retrieval. Foundations and Trends in Information Retrieval, 2015, 9(5): 355–475
169. Shi B, Weninger T. Fact checking in heterogeneous information networks. In: Proceedings of the 25th International Conference Companion on World Wide Web. 2016, 101–102
170. Nyhan B, Reifler J. When corrections fail: the persistence of political misperceptions. Political Behavior, 2010, 32(2): 303–330
171. Bordia P, DiFonzo N, Haines R, Chaseling E. Rumors denials as persuasive messages: effects of personal relevance, source, and message characteristics. Journal of Applied Social Psychology, 2005, 35(6): 1301–1331
172. Tanaka Y, Sakamoto Y, Honda H. The impact of posting urls in disaster-related tweets on rumor spreading behavior. In: Proceedings of the 47th Hawaii International Conference on System Sciences. 2014, 520–529
173. Ozturk P, Li H, Sakamoto Y. Combating rumor spread on social media: the effectiveness of refutation and warning. In: Proceedings of the 48th Hawaii International Conference on System Sciences. 2015, 2406–2414
174. Alemanno A. How to counter fake news? a taxonomy of anti-fake news

approaches. European Journal of Risk Regulation, 2018, 9(1): 1–5

175. Barrat A, Barthelemy M, Vespignani A. Dynamical Processes on Complex Networks. Paris: Cambridge University Press, 2008

176. Chang Y T, Yu H, Lu H P. Persuasive messages, popularity cohesion, and message diffusion in social media marketing. Journal of Business Research, 2015, 68(4): 777–782

177. Huang W M, Zhang L J, Xu X J, Fu X. Contagion on complex networks with persuasion. Scientific Reports, 2016, 6: 23766–23773

Bin Guo, PhD, professor, PhD supervisor of Northwestern Polytechnical University, China. He is a senior member of CCF. His main research interests include: ubiquitous computing, social and community intelligence, urban big data mining, mobile crowd sensing and human-computer interaction.

Yasan Ding is a PhD candidate at Northwestern Polytechnical University, China. His main research interest is social media data mining.

Yueheng Sun, PhD, associate professor of Tianjin University, China. His main research interests include: social network analysis, social media data processing and their applications in social management.

Shuai Ma, PhD, professor, PhD supervisor of Beihang University, China. He is a senior member of CCF. His main research interests include: big data, database theory and systems, graph and social data analysis, data cleaning and data quality.

Ke Li is a PhD candidate at Northwestern Polytechnical University, China. His main research interests include: probabilistic graphical model and social media mining.

Zhiwen Yu, PhD, professor, PhD supervisor of Northwestern Polytechnical University, China. He is a senior member of CCF. His main research interests include: pervasive computing, context aware systems, personalization, recommendation technology, mobile social networks and multimedia intelligent service.