The Mass, Fake News, and Cognition Security

Bin Guo*¹, Yasan Ding¹, Yueheng Sun², Shuai Ma³, Ke Li¹

- 1. Northwestern Polytechnical University, China
 - 2. Tianjin University, China
 - 3. Beihang University, China guob@nwpu.edu.cn

Abstract—The wide spread of 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 to 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 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 early detection of fake news, explainable fake news debunking, social contagion and diffusion models of fake news, and so on.

Index Terms—Cyberspace; cognition security; fake news; crowd computing; human-content interaction.

I. 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 accessing news from traditional and curated mechanisms, such as news broadcast or daily news programs, people are turning to social media platforms which expose them to a broader range of opinions and information about the issues of the day. The growth of *social media* has changed patterns of consumption and exposure to a variety of news deliberately and incidentally, and social media platforms have become a major source of news¹, such as Facebook², Twitter³, YouTube⁴, Instagram⁵ and Snapchat⁶. Although social networks have accelerated the dissemination of information and promoted the communication of people, contemporary social

media platforms offer a hotbed of spreading fake news due to their low cost, easy access and high anonymity. A survey conducted by the Pew Research Center shows that nearly 23% of interviewed Americans have ever reposted and shared fake news on social networks⁷. In addition, the existence of social bots, botnets and trolls have also been a severe problem in social media platforms. It is reported that as many as 60 million trolls could be spreading fake news on Facebook [1]. Furthermore, the prevalence of fake news in social networks confuses the audience, creates panic, and seriously affects public safety and mass cognition security [2].

The spread of fake news is posing threats to diverse domains, such as vaccine safety, climate change, political elections, and stock stability [3]. For example, during the U.S. presidential election in 2016, PolitiFact, an independent fact checker of political statements, judged 70% of all statements about Donald Trump to be false or mostly false⁸ and Trump's supporters were far more likely to consume fake news than Clinton's supporters [4]. Consequently, 'fake news' was named the "word of the year" by Collins Dictionary in 2017 since it has aroused spread concern of 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 Turkish government's implementation of capital controls led to a 20% drop in the lira against the US dollar⁹, causing huge economic loss in Turkey. The fake news which claimed that the border between Greece and North Macedonia was open made hundreds of migrants and refugees pour across the Greek border¹⁰. It further results in the clash between Greek police and migrants. Thus, it can be seen that fake news is one of the current greatest threats to democracy, economy and journalism [5].

In 2018, the *Science* magazine launched a special issue about 'Fake News', where they discussed the conception, network propagation mechanism and social influence of fake news [2, 6]. In [7], Ruths divides the dissemination process of fake news into five key components, consisting of publishers, authors, articles, audience and rumors. Qiu *et al.* [8] find that both information overload and limited attention contribute to the degradation of human's ability to judge news whether fake or true. Lazer *et al.* [2] identify two categories of fake news interventions, including empowering individuals to evaluate the

¹ https://www.oberlo.com/blog/social-media-marketing-statistics

² https://www.facebook.com/

³ https://twitter.com/

⁴ https://www.youtube.com/

⁵ https://www.instagram.com/

⁶ https://www.snapchat.com/

⁷ https://www.journalism.org/2016/12/15/many-americans-believe-fake-ne ws-is-sowing-confusion/

⁸ https://www.politifact.com/personalities/donald-trump/

 $^{^9\,}https://www.theguardian.com/world/2018/aug/13/turkey-financial-crisis-l-ira-plunges-again-amid-contagion-fears$

¹⁰ https://www.dw.com/en/greek-police-clash-with-migrants-near-north-ma-cedonia-border/a-48240710

fake news and utilizing platform-based detection and algorithms.

An urgent concern is that the development of Artificial Intelligence (AI) technology puts forward higher requirements for fake news identification. The research of fake news will extend from text to high-quality, machine-generated and manipulated images, videos and audios on a massive scale [9]. For instance, Deepfakes [10, 11], creates audios or videos of real people they never said or did by neural networks, which has been widely used to forge politicians' speeches and illegal evidence [12], resulting in hurting public feelings and affecting the political situation seriously.

To summarize, fake news can influence the 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 dissemination mechanism of fake news before checking the fact. 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 investigates the successful and efficient ways to debunk fake news and maintain human cognition security.

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

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

- Characterizing the Cognition Security (CogSec) research area, ranging from its concept model and 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.

II. CHARACTERIZING COGNITION SECURITY

In addition to fake news, there are other types of information spreading on social media platforms that threaten 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 2.1. **Fake news**: A news article that is intentionally and verifiable false.

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

DEFINITION 2.2. **Human-content interaction**: *Publish, share, like, and comment of online contents (e.g., news, posts, photos, videos, etc).*

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

DEFINITION 2.3. **Cognition security**: CogSec refers to the potential impacts of fake news to human cognition, ranging from misperception, untrusted knowledge acquisition, targeted opinion/attitude formation, to biased decision making.

DEFINITION 2.4. **Cognition security protection**: CogSec protection is committed to effective intervention to ensure humans' CogSec, including the techniques of cognition mechanism investigation, diffusion pattern mining, early fake news detection, malicious bot detection, and so on.

Regarding the scale of human-beings the cognition security can affect, it can be categorized into the *individual* level, the *crowd* level, and the *society* level.

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 content of social media, expanding from the traditional "machine" security to "human-machine" fusion security, as presented in Table 2.

TABLE I. DEFINITIONS OF SOME TYPES OF MALICIOUS INFORMATION

Term	Definition	
Rumor	An item of circulating information whose veracity status is yet to be verified at the time of posting. [14]	
Ноах	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 verifiable false. [18]	

TABLE II. DIFFERENCES WITH OTHER CONCEPTS

Term	Research Focus	Security Paradigm
Network security	Data and content security	Machine security
Cognition security	The interaction and cognitive mechanism between human and contents in the cyberspace	Human-machine security

Recently, there have been several related studies and important findings regarding this research field, representative ones as presented below.

(1) Echo chambers [23-25]. It traps users by only exposing them to opinions and beliefs they are already in agreement with [26]. Echo chambers is compounded by the rise of algorithmic news recommendation and content filtering [27], which makes

users always browse their favorite information and implicitly influences users' cognitive behaviors. For example, Barberá et al. [28] observe that information is mainly exchanged among users with similar ideological preferences in the case of political issues. Similarly, Quattrociocchi et al. [29] demonstrate that such echo chambers really reinforce selective exposure and group polarization. People tend to only concentrate on confirming claims and ignore obvious objections, because they focus on their preferred information. Moreover, Zajonc et al. [30] 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. Above all, highly homogeneous echo chambers in social networks can decrease people's ability to identify fake news and increase their misperceptions, contributing to spreading false information [31].

(2) Online gatekeepers [32, 33]. It refers to information controller (information selection, deletion, manipulation or integration etc.) in the process of information dissemination [34]. Xu et al. [35] observe that users in social networks are highly likely to become gatekeepers. In [36], Garimella et al. explore the role of gatekeepers in the creation of echo chambers in case of political news, and they find these gatekeepers usually have lower clustering coefficient. Although online gatekeepers consume information with different viewpoints, they tend to share only a certain viewpoint to strengthen the homogeneity of target community and form a closed field of public opinion, which contributes to the dissemination of fake news [37]. Therefore, effective use of gatekeepers to prevent the spread of fake news needs to be further studied.

(3) Media bias [38, 39]. It is one type of cognitive bias, which means that journalists are unable to report news events fairly and objectively due to their partial opinions [40]. As Jamieson et al. [41] recognize, the news media does not just report the facts, but is often affected by government influence, targeting at audiences' preference, sponsor pressure and so on. Under the comprehensive impact of various aspects as well as the purpose of chasing headlines, media outlets often release claims without thorough verification, which provides an opportunity for the spread of fake news. Puglisi [42] finds that the New York Times may lean democratic. Besides, Gerber et al. [43] estimate that voters who read the Washington Post regularly are 8% more likely to vote democratic candidate in the 2005 governor election in Virginia. Many media researchers fear that unregulated media will have a major impact on our society [44], but competition among different media outlets can eliminate ideological bias in some cases [45].

(4) The spread of fake news [46, 47]. There are many factors that contribute to the spread of fake news, such as cognitive limitation of readers [48], usability of social media platforms [49], and demographics of audiences [50]. Some studies have been carried out on the propagation characteristics and structures of fake news. For example, DiFonzo *et al.* [51] find that rumors containing negative emotions are more likely to be spread. Guess *et al.* [52] state that conservatives are more likely to share fake news and that Facebook accounts over 65 years

old spread about seven times as much fake news as the young during the 2016 US presidential election. Budak *et al.* [53] demonstrate that the popularity of fake news is the result of news production and consumption. They further find that male voters are more impressed by fake news publishers.

III. 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.

A. Human-Content Cognition Mechanism

Understanding the mechanism that people share, repost, and agree of online contents is critical to protect their cognitive security. A thorough understanding of the mechanisms should rely on knowledge from psychology, cognition science, and neuroscience [54].

(1) Personality, content sharing, and debunking. Interpersonal social interaction, centered on content sharing, enables information to spread efficiently [55]. Actually, content sharing behaviors among users in social networks, such as publish, repost, and like, will gradually affect the reach and influence of news [56]. There are several studies that aim to learn information sharing mechanism in social media. For instance, Scholz et al. [57] present a neurocognitive framework to understand mechanisms under information sharing. Based on the New York Times health news articles dataset, they find that the core functions of sharing relate to both self-expression and social bond strengthen. Hodas et al. [58] reveal a systematic link between personality type and mood, brain response, and the type of content people choose to share online. They observe that users' preferences might be predicted from both personality and transitory mood state. In [59], Falk et al. focus on neural responses of information consumers' brains. They find that individuals are more capable of spreading their opinions to others, thus generating greater mentalizing-system activity in the initial process of information sharing.

Some works predict content reposts in social networks. For example, Hu *et al.* [60] predict the popularity of pictures and their diffusion paths in social networks based on *Diffusion-LSTM*, a memory-based deep recurrent neural network model. A combination of user social features and image features is used to characterize individual reposting behaviors. Similarly, Zhang *et al.* [61] propose an attention-based deep neural network to combine contextual and social context information for *retweet* behavior prediction.

In [62], Lewandowsky *et al.* observe audiences' memories for misinformation and study the role of cognitive factors in misinformation debunking. They further divide human cognitive problems in the face of misinformation into four categories, including *continued influence effect*, *familiarity backfire effect*, overkill backfire effect, and worldview backfire effect, which provides the theoretical basis and suggestions for CogSec protection.

Neuroscience in human-content interaction. Neuroscience has also been widely used in many research areas (e.g., healthcare [63], intelligent control [64, 65], artificial intelligence [66], economics [67] etc.) related to humancomputer interaction. As presented by Poldrack et al. [68], the use of new tools, e.g., Electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), for Magnetoencephalography (MEG), imaging and manipulating the brain will continue to advance understanding of how the human brain gives rise to thought and action.

Regarding CogSec, neuroscience has been previously used for understanding human-content interaction. Some efforts have been conducted to understand/predict population-level behaviors/preferences (e.g., ratings and sharing in social media) based on small groups of individuals' neural responses. For example, researchers test the possibility of using fMRI to predict the relative popularity of music¹¹. Dmochowski *et al.* [69] find that naturalistic stimuli (viewing multimedia contents) evoke highly reliable brain activities across viewers. Falk et al. [70] further conclude that neural responses of a small group of individuals can be used to predict the behavior of large-scale populations. In particular, neural activities in a medial prefrontal region of interest which are previously associated with individual behavior change can predict the population response. Hasson et al. [71] report the unexpected finding that brains of different individuals show a highly significant tendency to act in unison during free viewing of a complex scene such as a movie sequence. In [72], Adolphs identifies a series of neural structures involved in users' perceptions and judgements of content stimuli, and analyzes humans' ways of reasoning and decision-making. In general, neuroscience provides the theoretical basis for understanding human-content interaction, and has practical significance for the protection of public CogSec.

B. Social Influence and Opinion Diffusion

The study of social influence and opinion diffusion in social networks has a long tradition in the social, physical, and computational sciences. For example, there have been numerous studies on opinion formation [73, 74] and influence maximization models [75]. Here, 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 the process of information dissemination across social networks, such as political views [76], emotional changes [77], fashion trends [78], and financial decisions [79]. Some works measure the influence of opinions in social networks, aiming to make information far-reaching. For example, Morone et al. [80] introduce percolation theory [81] to social network influential node discovery and find that a large number of weakly-connected (low-degree) nodes can be optimal influencers. Amati et al. [82] utilize degree, closeness, betweenness and PageRank-centrality of nodes in Dynamic Retweet Graph [83]

to find the most influential users in Twitter. In [84], Qiu *et al.* propose *DeepInf*, a deep learning-based influence prediction framework, which learns users' latent social representation to evaluate their social influence by incorporating *network embedding*, *graph convolution*, and *attention* mechanism.

Some studies concentrate on the contagion and persuasion mechanisms of messages in social networks. For instance, Ugander et al. [85] find that whether social network users will be infected depends on the number and structure of their interrelated components, rather than the actual size of the community. Therefore, different social environments and influences represented by target users' neighbors can be considered as the driving mechanism of social contagion. In [86], Kramer et al. prove that each user's emotions can be affected by other users in Facebook, which provides an experimental basis for massive-scale social influence and contagion. Abebe et al. [87] study the process of information contagion from the perspective of changes in 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 for influencing social opinions.

(2) Spreading models/mechanisms. As Ratkiewicz et al. [88] state, the early stages of the diffusion of rumors tend to show pathological patterns. Thus, some work has studied the spreading mechanisms and modes of online information to provide guidance for CogSec protection. For example, Friggeri et al. [89] track the propagation of thousands of rumors appearing on Facebook. They find that rumor cascades run deeper in the social network than normal sharing cascades. Vosoughi et al. [6] report that fake news is more novel than real news, suggesting that people are more willing to spread novel information. Besides, the true information usually evokes the users' sadness, happiness, and trust, while fake news often triggers public surprise, fear, and disgust. Similarly, Peng et al. [90] find that users are more delight to hear positive gossip and more annoyed to hear negative gossip of themselves, compared with celebrities and their friends. Vicario et al. [31] find that misinformation in social networks often leads to homogeneous and polarized communities and propose a data-driven percolation model of misinformation spreading, which demonstrates that homogeneity and polarization are the determinants of predicting the size of information cascade.

Several works have been carried out on the opinion dynamics based on influence mechanism in social networks, which can be divided into discrete models [91, 92] and continuous models [93]. For instance, aiming at understanding the vulnerability of social networks and increasing users' resilience to fake news, Wang *et al.* [94] propose a multivariable jump diffusion guidance framework, which models the dynamics of opinions and guides public opinions to the desired state. Martins *et al.* [95] propose an opinion diffusion model, *CODA*, in which different opinions of users are regarded as discrete variables and each opinion is modeled

¹¹ https://www.sciencemag.org/news/2011/06/can-brain-scans-predict-musi c-sales

as continuous opinion function. Target users decide whether to change their own opinions or not based on Bayesian descriptions of their neighbor opinions. In [96], Yang et al. design a role-aware information diffusion model (RAIN), which characterizes the interaction between users' social roles and their influence on the spreading of information.

C. Fake News Detection

Since fake news has a great impact on social stability, economic development, and political democracy, it is imperative to study efficiently automatic fake news detection technology [19]. 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.

- (1) Content-based methods, which often rely on unique writing styles or language features in news content (e.g., lexical features, syntactic features, and topic features) [97, 98]. For example, Castillo et al. [99] calculate a series of linguistic features to evaluate Twitters' credibility, including the average number of words, URL links, the number of positive words etc. Potthast et al. [100] propose a meta-learning model to detect fake news, which utilizes differences in writing styles between the truth and fake news. Hu et al. [101] propose a spammer detection method based on sentiment information.
- (2) Social context-based methods, which mainly focus on the characteristics of human-content interactions, such as user profiling, reposts, comments, stances, and likes etc. For example, Tacchini et al. [102] estimate that social media platform posts can be detected as hoax utilizing netizens' like behaviors. Ma et al. [103] make use of the temporal patterns of social context features to detect online rumors. In [104], Jin et al. propose a credibility propagation network model for rumor detection by mining supporting or opposing opinions in microblogs. Yang et al. [105] propose an unsupervised fake news detection model, incorporating the authenticity of news, users' reputation, and users' viewpoints on target news event.
- (3) Deep learning-based methods, which aim to learn latent representations of fake and real news accurately for further detection. Existing deep learning-based detection methods mainly apply convolution neural network (CNN) [106] and recurrent neural network (RNN) [107] models. For example, Li et al. [108] 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. [109] 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. [110] propose the RNN-based fake news detection model, incorporating textual features of news, user response, and the source users. Similarly, Shu et al. [111] further explore the social relations among publishers, news and online users [112, 113], and then propose a tri-relationship embedding network, TriFN, which models the human-content interactions for fake news detection.

D. Malicious Bot Detection

The popularity and openness of social network promote the emergence of social bots with certain autonomous decision-making ability [114]. Like legitimate users, social bots can make friends, post tweets, thumb up, chat and so on through program control. Salge et al. [115] 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 information by automatically providing benign news and information, but there are also malicious social bots that can spread rumors and harmful information [114, 116, 117]. 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.

- (1) Behavior-based detection methods. It is of great value to analyze and mine the behavior data of social bots in existing social networks [118]. Boshmaf et al. [119] analyze the differences between social bots and human users in terms of the number of friends, post time interval, post content and account attribute differences, and propose a random forest based social bot detection method. Haustein et al. [120] analyze the differences between real Twitter users and social bots in retweeting scientific articles, and find that social bots tend not to be selective in retweeting (involving topics, sources, etc.). In [121], Gilani et al. 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. Besides, Varol et al. [122] find that compared with human users, the interaction selection of social bots is more arbitrary and that there are fewer bidirectional connections between them and human users.
- (2) Content-based detection methods, which focus on determining whether a message posted by a user is a malicious message. Generally, whether the URL in the message content points to the malicious page can be used to determine whether the account that published the message is malicious social bot. For instance, Thomas et al. [123] propose a real-time URL detection scheme, which extracts features of related URL pages by visiting each published URLs. What's more, social bots can be detected through changes in the message content features. For example, Egele et al. [124] extract 7 content features, model the messages, and then judge whether the messages published later deviate from the created model to detect social bots. In [125], Kudugunta et al. propose a LSTM-based bot detection method, incorporating contextual features and accounts' metadata for improving bot detection accuracy. Gao et al. [126] find that 63% of the text content of spam messages in Twitter is generated based on templates, and 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, which detect social bots on the perspective of social influence. For example, Messias et al. [127] conduct comparative studies on analyzing the influence of social bots, and propose their malicious

behavior strategies, including regular posting tweets on a certain hot topic, different posting intervals, and attribute integrity. Similarly, Abokhodair *et al.* [128] analyze the posting behavior, social structure, group behavior characteristics and influence growth process of social bot network. Finally, they find that more human-like behaviors can improve social bot influence. Freitas *et al.* [129] create 120 different attributes (sex, occupation, etc.) and behavior strategies (active, posting action and interaction) of the social bots for characterizing their infiltration process, and they find that about 20% of the social bots gain more than 100 followers by means of high active interaction and posting behavior.

IV. OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

Though there have been initial efforts in the field of CogSec, there are still numerous research challenges to be tackled in the future, some of which are discussed below.

- (1) The human cognition mechanism of fake news. Regarding to CogSec protection, the first thing is to understand the human cognition mechanism of fake news. Acerbi [13] estimates that fake news can be successfully disseminated because it meets the general cognitive preferences of the public, which provides theoretical guidance for preventing the spread of fake news. With the rapid development of neuroscience, several studies have investigated the cognitive patterns of the human brain [57, 130, 131]. For example, Lewandowsky et al. [132] raise the problem of "technocognition" and summarize the ways in which fake news affect the society negatively. In [133], Arapakis et al. propose a measurement model for evaluating the interest changes of users in reading news, which is based on EEG registration of people's neural activity. 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 cognitive problems, partially summarized as follows.
 - The influence of individual's social cognition on largescale social behaviors.
 - The common features of fake news satisfying users' cognitive preferences.
 - The effect of fake news content stimulations (text, pictures, audios or videos) on specific parts of the human brains.
 - The impact of social interactions on individual's cognition.
 - The change of users' cognitive characteristics with the dissemination of fake news.

(2) The social contagion and diffusion models of fake news. Social contagion is a common phenomenon in human society [134], which contributes to opinion dynamics, behavior shaping, and cognitive preferences in social networks. Some works pay attention to modeling the contagion and propagation of information in social networks. For instance, Chang et al. [135] explore how social media marketing persuades users to share information with the purpose of achieving mass cohesion and information diffusion. Huang et al. [136] propose a social contagion model based on introducing a persuasion mechanism

into the threshold model. They then estimate that persuasion mechanism improves the influence 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 numerous issues to be further studied:

- The study of novel information dissemination theories which introduce the users' cognition preferences, timeliness of information, and social roles of individuals, etc.
- The evaluation of influential users on social networks for maximizing the impact of information dissemination.
- The fast influence maximization mechanisms of true information to the mass after fake news debunking.
- (3) Early detection of fake news. Information on social networks usually has a short life span, averaging less than three days, and fake news always spread like viruses with a few minutes [89, 137]. Actually, detection methods based on aggregation features (e.g., propagation characteristics, etc.) are difficult to achieve better performance on early detection [138]. Therefore, the early detection of fake news is an important issue. Some works attempt to identify fake news at their early spreading stage. For example, Zhao et al. [139] find that queries and objections in users' comments contribute to early detection of rumors. Chen et al. [140] find that users tend to comment differently in different rumors' spreading process and propose an RNN-based rumor detection model with attention mechanism for early detection. In [141], Sampson et al. utilize implicit linkages for acquiring additional information from several related events to deal with the problem that less data is available in the early detection of fake news. Although several studies have been conducted on the early detection of fake news, the performances of them still need to be improved.
- (4) Explainable fake news debunking. Existing automatic fake news detection models [99, 107] usually just give the testing results, with little decision-making basic explanations. However, the explanation in fake news debunking or the transparency of detection models is essential, which contributes to users' trust in detection results, fusion of human-machine intelligence, and further prevention of the spread of fake news. Some studies utilize the attention mechanism [140, 142] and graph models [143, 144] for explainable fake news debunking. For instance, Popat et al. [145] propose an automatic end-toend fake news detection model combined with external evidence articles, DeClarE, based on Bidirectional LSTM with attention. Similarly, Guo et al. [146] introduce social contexts into rumor detection via attention mechanism to enhance the interpretability of detection models, based on hierarchical LSTM. Gad-Elrab et al. [147] propose a framework for generating explanations of candidate facts, incorporating knowledge graphs and texts, which provides reference for fake news detection. In general, explainable fake news debunking needs to explore more practical models with the development of interpretable machine learning (IML) [148, 149], such as probabilistic graphical model (PGM) [150], knowledge graph based on complex rules [151], and other mechanisms.

V. CONCLUSION

In the context of the spread of fake news in social networks, we present a novel research issue, named Cognitive Security (CogSec). In order 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 mechanism, 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.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China (No. 61772428,61725205).

REFERENCES

- S. Iyengar, and D. S. Massey, "Scientific communication in a post-truth society," *Proceedings of the National Academy of Sciences*, vol.116, no.16, pp.7656-7661, 2019.
- [2] D. M. J. Lazer, et al., "The science of fake news," Science, vol.359, no.6380, pp.1094-1096, 2018.
- [3] M. Fernandez, and H. Alani, "Online misinformation: Challenges and future directions," in *Companion Proceedings of the The Web Conference* 2018. International World Wide Web Conferences Steering Committee, 2018, pp. 595-602.
- [4] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign," *European Research Council*, vol.9, 2018.
- [5] X. Zhou, et al., "Fake news: Fundamental theories, detection strategies and challenges," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019, pp.836-837.
- [6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol.359, no.6380, pp.1146-1151, 2018.
- [7] D. Ruths, "The misinformation machine," Science, vol.363, no.6425, pp.348-348, 2019.
- [8] X. Qiu, et al., "Limited individual attention and online virality of low-quality information," Nature Human Behaviour, vol.1, no.7, pp.0132, 2017
- [9] J. Bakdash, et al., "The Future of Deception: Machine-Generated and Manipulated Images, Video, and Audio?," in 2018 International Workshop on Social Sensing (SocialSens). IEEE, 2018, pp.2-2.
- [10] L. Floridi, "Artificial intelligence, deepfakes and a future of ectypes," Philosophy & Technology, vol.31, no.3, pp.317-321, 2018.
- [11] P. Korshunov, and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," arXiv preprint arXiv:1812.08685, 2018.
- [12] S. Agarwal, et al., "Protecting World Leaders Against Deep Fakes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2019, pp.38-45.
- [13] A. Acerbi, "Cognitive attraction and online misinformation," *Palgrave Communications*, vol.5, no.1, pp.15-21, 2019.
- [14] A. Zubiaga, et al., "Detection and resolution of rumours in social media: A survey," ACM Computing Surveys (CSUR), vol.51, no.2, pp.32-67, 2018.
- [15] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp.591-602.
- [16] S. Volkova, et al., "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017, pp.647-653.

- [17] L. Wu, et al., "Mining misinformation in social media," in Big Data in Complex and Social Networks, 1st ed., London, UK: Chapman and Hall/CRC, 2016, pp.135-162.
- [18] K. Shu, et al., "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vo.19, no.1, pp.22-36, 2017.
- [19] X. Zhou, and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," arXiv preprint arXiv:1812.00315, 2018.
- [20] G. F. C. Campos, et al., "Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol.14, no.1s, pp.26-42, 2018.
- [21] S. A. Macskassy, "On the study of social interactions in twitter," in Sixth International AAAI Conference on Weblogs and Social Media. 2012.
- [22] B. A. Forouzan, Cryptography & network security, New York, NY, USA: McGraw-Hill, Inc., 2007.
- [23] D. DiFranzo, and M. J. K. Gloria, "Filter bubbles and fake news," ACM Crossroads, vol.23, no.3, pp.32-35, 2017.
- [24] C. Vaccari, "From echo chamber to persuasive device? Rethinking the role of the Internet in campaigns," *New Media & Society*, vol.15, no.1, pp.109-127, 2013.
- [25] A. Tsang, and K. Larson, "The echo chamber: Strategic voting and homophily in social networks," in *Proceedings of the 2016 international* conference on autonomous agents & multiagent systems. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp.368-375.
- [26] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public opinion quarterly*, vol.80, no.S1, pp.298-320, 2016.
- [27] M. Flintham, et al., "Falling for fake news: investigating the consumption of news via social media," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, pp.376-385.
- [28] P. Barberá, et al., "Tweeting from left to right: Is online political communication more than an echo chamber?," Psychological science, vol. 26, no.10, pp.1531-1542, 2015.
- [29] W. Quattrociocchi, A. Scala, and C. R. Sunstein, "Echo chambers on Facebook," Available at SSRN 2795110, 2016.
- [30] R. B. Zajonc, "Attitudinal effects of mere exposure," *Journal of personality and social psychology*, vol.9, no.2p2, pp.1-27, 1968.
- [31] M. Del Vicario, et al., "The spreading of misinformation online," Proceedings of the National Academy of Sciences, vol.113, no.3, pp.554-559, 2016.
- [32] J. B. Singer, "Online journalists: Foundations for research into their changing roles," *Journal of computer-mediated communication*, vol.4, no.1, pp.JCMC412, 1998.
- [33] R. K. Nielsen, "News media, search engines and social networking sites as varieties of online gatekeepers," in *Rethinking journalism again*, London, UK: Routledge, 2016, pp.93-108.
- [34] C. L. Bui, "HOW ONLINE GATEKEEPERS GUARD OUR VIEW-NEWS PORTALS'INCLUSION AND RANKING OF MEDIA AND EVENTS," *Global Media Journal*, vol.9, no.16, 2010.
- [35] W. W. Xu, and M. Feng, "Talking to the broadcasters on Twitter: Networked gatekeeping in Twitter conversations with journalists," *Journal of Broadcasting & Electronic Media*, vol.58, no.3, pp.420-437, 2014.
- [36] K. Garimella, et al., "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018, pp.913-922.
- [37] N. DiFonzo, "Ferreting facts or fashioning fallacies? Factors in rumor accuracy," *Social and Personality Psychology Compass*, vol.4, no.11, pp.1124-1137, 2010.
- [38] R. M. Entman, "Framing bias: Media in the distribution of power," *Journal of communication*, vol.57, no.1, pp.163-173, 2007.
- [39] C. F. Chiang, and B. Knight, "Media bias and influence: Evidence from newspaper endorsements," *The Review of Economic Studies*, vol.78, no.3, pp.795-820, 2011.
- [40] S. Iyengar, and D. R. Kinder, News that matters: Television and American opinion. Palo Alto, CA, USA: University of Chicago Press, 2010, pp.6-34.

- [41] K. H. Jamieson, and K. K. Campbell, *Interplay of Influence: News, Advertising, Politics and the Internet Age (with InfoTrac)*. Belmont, CA, USA: Wadsworth Publishing, 2005, pp.22-46.
- [42] R. Puglisi, "Being the New York Times: the political behaviour of a newspaper," *The BE Journal of Economic Analysis & Policy*, vol.11, no.1, 2011.
- [43] A. S. Gerber, D. Karlan, and D. Bergan, "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions," *American Economic Journal: Applied Economics*, vol.1, no.2, pp.35-52, 2009.
- [44] F. N. Ribeiro, et al., "Media bias monitor: Quantifying biases of social media news outlets at large-scale," in Twelfth International AAAI Conference on Web and Social Media. 2018.
- [45] C. Budak, S. Goel, and J. M. Rao, "Fair and balanced? quantifying media bias through crowdsourced content analysis," *Public Opinion Quarterly*, vol.80, no.S1, pp.250-271, 2016.
- [46] A. Bovet, and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature communications*, vol.10, no.1, pp.7-20, 2019.
- [47] A. Kucharski, "Post-truth: Study epidemiology of fake news," *Nature*, vol.540, no.7634, pp.525-525, 2016.
- [48] N. DiFonzo, et al., "Validity judgments of rumors heard multiple times: the shape of the truth effect," Social Influence, vol.11, no.1, pp.22-39, 2016.
- [49] E. W. T. Ngai, S. S. C. Tao, and K. K. L. Moon, "Social media research: Theories, constructs, and conceptual frameworks," *International Journal of Information Management*, vol.35, no.1, pp.33-44, 2015.
- [50] H. Allcott, and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol.31, no.2, pp.211-236, 2017
- [51] N. DiFonzo, et al., "Rumor clustering, consensus, and polarization: Dynamic social impact and self-organization of hearsay," Journal of Experimental Social Psychology, vol.49, no.3, pp.378-399, 2013.
- [52] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science advances*, vol.5, no.1, pp.eaau4586, 2019.
- [53] C. Budak, "What happened? The Spread of Fake News Publisher Content During the 2016 US Presidential Election," in *The World Wide Web Conference*. ACM, 2019, pp:139-150.
- [54] R. A. Poldrack, and M. J. Farah, "Progress and challenges in probing the human brain," *Nature*, vol.526, no.7573, pp.371-382, 2015.
- [55] G. Csibra, and G. Gergely, "Natural pedagogy as evolutionary adaptation," Philosophical Transactions of the Royal Society B: Biological Sciences, vol.366, no.1567, pp.1149-1157, 2011.
- [56] J. N. Cappella, H. S. Kim, and D. Albarracín, "Selection and transmission processes for information in the emerging media environment: Psychological motives and message characteristics," *Media psychology*, vol.18, no.3, pp.396-424, 2015.
- [57] C. Scholz, et al., "A neural model of valuation and information virality," Proceedings of the National Academy of Sciences, vol.114, no.11, pp.2881-2886, 2017.
- [58] N. O. Hodas, and R. Butner, "How a user's personality influences content engagement in social media," in *International Conference on Social Informatics*. Springer, Cham, 2016, pp.481-493.
- [59] E. B. Falk, et al., "Creating buzz: the neural correlates of effective message propagation," Psychological Science, vol.24, no.7, pp.1234-1242, 2013.
- [60] W. Hu, et al., "Who will share my image?: Predicting the content diffusion path in online social networks," in Proceedings of the eleventh ACM international conference on web search and data mining. ACM, 2018, pp.252-260.
- [61] Q. Zhang, et al., "Retweet prediction with attention-based deep neural network," in Proceedings of the 25th ACM international on conference on information and knowledge management. ACM, 2016, pp.75-84.
- [62] S. Lewandowsky, et al., "Misinformation and its correction: Continued influence and successful debiasing," Psychological Science in the Public Interest, vol.13, no.3, pp.106-131, 2012.
- [63] R. J. Davidson, et al., "Depression: perspectives from affective neuroscience," Annual review of psychology, vol.53, no.1, pp.545-574, 2002.

- [64] K. S. LaBar, and R. Cabeza, "Cognitive neuroscience of emotional memory," *Nature Reviews Neuroscience*, vol.7, no.1, pp.54-64, 2006.
- [65] P. A. Howard-Jones, "Neuroscience and education: myths and messages," Nature Reviews Neuroscience, vol.15, no.12, pp.817-824, 2014.
- [66] D. Hassabis, et al., "Neuroscience-inspired artificial intelligence," Neuron, vol.95, no.2, pp.245-258, 2017.
- [67] C. Camerer, G. Loewenstein, and D. Prelec, "Neuroeconomics: How neuroscience can inform economics," *Journal of economic Literature*, vol.43, no.1, pp.9-64, 2005.
- [68] R. A. Poldrack, and M. J. Farah, "Progress and challenges in probing the human brain," *Nature*, vol.526, no.7573, pp.371-382, 2015.
- [69] J. P. Dmochowski, et al., "Audience preferences are predicted by temporal reliability of neural processing," *Nature communications*, vol.5, pp.4567-4575, 2014.
- [70] E. B. Falk, E. T. Berkman, and M. D. Lieberman, "From neural responses to population behavior: Neural focus group predicts population-level media effects," *Psychological science*, vol.23, no.5, pp.439-445, 2012.
- [71] U. Hasson, et al., "Intersubject synchronization of cortical activity during natural vision," science, vol.303, no.5664, pp.1634-1640, 2004.
- [72] R. Adlolphs, "Cognitive neuroscience of human social behavior," *Nature Reviews Neuroscience*, vol.4, pp.165-178, 2003.
- [73] M. H. DeGroot, "Reaching a consensus," Journal of the American Statistical Association, vol.69, no.345, pp.118-121, 1974.
- [74] R. B. Cialdini, R. E. Petty, and J. T. Cacioppo, "Attitude and attitude change," *Annual review of psychology*, vol.32, no.1, pp.357-404, 1981.
- [75] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp.137-146.
- [76] P. Rozin, and E. B. Royzman, "Negativity bias, negativity dominance, and contagion," *Personality and social psychology review*, vol.5, no.4, pp.296-320, 2001.
- [77] E. Hatfield, J. T. Cacioppo, and R. L. Rapson, "Emotional contagion," Current directions in psychological science, vol.2, no.3, pp.96-100, 1993.
- [78] J. J. Argo, D. W. Dahl, and A. C. Morales, "Positive consumer contagion: Responses to attractive others in a retail context," *Journal of Marketing Research*, vol.45, no.6, pp.690-701, 2008.
- [79] F. Allen, and D. Gale, "Financial contagion," *Journal of political economy*, vol.108, no.1, pp.1-33, 2000.
- [80] F. Morone, and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol.524, no.7563, pp.65-147, 2015.
- [81] C. Moore, and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Physical Review E*, vol.61, no.5, pp.5678-5683, 2000.
- [82] G. Amati, et al., "Influential users in Twitter: detection and evolution analysis," Multimedia Tools and Applications, vol.78, no.3, pp.3395-3407, 2019
- [83] G. Amati, and et al., "TWITTER TEMPORAL EVOLUTION ANALYSIS: COMPARING EVENT AND TOPIC DRIVEN RETWEET GRAPHS," IADIS International Journal on Computer Science & Information Systems, vol.11, no.2, 2016.
- [84] J. Qiu, et al., "Deepinf: Social influence prediction with deep learning," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp.2110-2119.
- [85] J. Ugander, et al., "Structural diversity in social contagion," Proceedings of the National Academy of Sciences, vol.109, no.16, pp.5962-5966, 2012.
- [86] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol.111, no.24, pp.8788-8790, 2014.
- [87] R. Abebe, et al., "Opinion dynamics with varying susceptibility to persuasion," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp.1089-1098.
- [88] J. Ratkiewicz, et al., "Truthy: mapping the spread of astroturf in microblog streams," in Proceedings of the 20th international conference companion on World wide web. ACM, 2011, pp.249-252.
- [89] A. Friggeri, et al., "Rumor cascades," in Eighth International AAAI Conference on Weblogs and Social Media. 2014.

- [90] X. Peng, et al., "The ugly truth: negative gossip about celebrities and positive gossip about self entertain people in different ways," Social neuroscience, vol.10, no.3, pp.320-336, 2015.
- [91] M. Granovetter, "Threshold models of collective behavior," American journal of sociology, vol.83, no.6, pp.1420-1443, 1978.
- [92] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *International Colloquium on Automata*, *Languages, and Programming*. Springer, 2005, pp.1127-1138.
- [93] S. Chatterjee, and E. Seneta, "Towards consensus: Some convergence theorems on repeated averaging," *Journal of Applied Probability*, vol.14, no.1, pp.89-97, 1977.
- [94] Y. Wang, et al., "Steering opinion dynamics in information diffusion networks," arXiv preprint arXiv:1603.09021, 2016.
- [95] A. C. R. Martins, "Continuous opinions and discrete actions in opinion dynamics problems," *International Journal of Modern Physics C*, vol.19, no.04, pp.617-624, 2008.
- [96] Y. Yang, et al., "RAIN: Social Role-Aware Information Diffusion," in AAAI. 2015, vol.15, pp.367-373.
- [97] V. L. Rubin, and T. Lukoianova, "Truth and deception at the rhetorical structure level," *Journal of the Association for Information Science and Technology*, vol.66, no.5, pp.905-917, 2015.
- [98] S. Kumar, and N. Shah, "False information on web and social media: A survey," arXiv preprint arXiv:1804.08559, 2018.
- [99] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp.675-684.
- [100] M. Potthast, et al., "A stylometric inquiry into hyperpartisan and fake news," arXiv preprint arXiv:1702.05638, 2017.
- [101] X. Hu, et al., "Social spammer detection with sentiment information," in 2014 IEEE International Conference on Data Mining. IEEE, 2014, pp.180-189.
- [102] E. Tacchini, et al., "Some like it hoax: Automated fake news detection in social networks," arXiv preprint arXiv:1704.07506, 2017.
- [103] J. Ma, et al., "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp.1751-1754.
- [104]Z. Jin, et al., "News verification by exploiting conflicting social viewpoints in microblogs," in Thirtieth AAAI Conference on Artificial Intelligence. 2016.
- [105]S. Yang, et al., "Unsupervised fake news detection on social media: A generative approach," in Proceedings of 33rd AAAI Conference on Artificial Intelligence. 2019.
- [106] F. Yu, et al., "A convolutional approach for misinformation identification," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.
- [107] J. Ma, et al., "Detecting rumors from microblogs with recurrent neural networks," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016, pp.3818-3824.
- [108] L. Li, G. Cai, and N. Chen, "A Rumor Events Detection Method Based on Deep Bidirectional GRU Neural Network," in 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2018, pp.755-759.
- [109]Y. Liu, Y. F. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [110] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp.797-806.
- [111] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp.312-320.
- [112] K. Shu, H. R. Bernard, and H. Liu, "Studying fake news via network analysis: detection and mitigation," in *Emerging Research Challenges* and Opportunities in Computational Social Network Analysis and Mining. Springer, 2019, pp.43-65.
- [113] K. Shu, et al., "The Role of User Profile for Fake News Detection," arXiv preprint arXiv:1904.13355, 2019.

- [114]E. Ferrara, et al., "The rise of social bots," Communications of the ACM, vol.59, no.7, pp.96-104, 2016.
- [115] C. A. de Lima Salge, and N. Berente, "Is that social bot behaving unethically?," Communications of the ACM, vol.60, no.9, pp.29-31, 2017.
- [116] Z. Chu, et al., "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," *IEEE Transactions on Dependable and Secure Computing*, vol.9, no.6, pp.811-824, 2012.
- [117] Y. Boshmaf, et al., "Design and analysis of a social botnet," Computer Networks, vol.57, no.2, pp.556-578, 2013.
- [118] Z. Chu, et al., "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," *IEEE Transactions on Dependable and Secure Computing*, vol.9, no.6, pp.811-824, 2012.
- [119] Y. Boshmaf, et al., "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th annual computer security applications conference*. ACM, 2011, pp.93-102.
- [120]S. Haustein, *et al.*, "Tweets as impact indicators: Examining the implications of automated "bot" accounts on T witter," *Journal of the Association for Information Science and Technology*, vol.67, no.1, pp.232-238, 2016.
- [121]Z. Gilani, et al., "An in-depth characterisation of Bots and Humans on Twitter," in arXiv preprint arXiv:1704.01508, 2017.
- [122]O. Varol, et al., "Online human-bot interactions: Detection, estimation, and characterization," in Eleventh international AAAI conference on web and social media. 2017.
- [123] K. Thomas, et al., "Design and evaluation of a real-time url spam filtering service," in 2011 IEEE symposium on security and privacy. IEEE, 2011, pp.447-462.
- [124] M. Egele, et al., "Towards detecting compromised accounts on social networks," *IEEE Transactions on Dependable and Secure Computing*, vol.14, no.4, pp.447-460, 2015.
- [125]S. Kudugunta, and E. Ferrara, "Deep neural networks for bot detection," Information Sciences, vol.467, pp.312-322, 2018.
- [126] H. Gao, et al., "Spam ain't as diverse as it seems: throttling OSN spam with templates underneath," in *Proceedings of the 30th Annual Computer Security Applications Conference*, ACM, 2014, pp.76-85.
- [127] J. Messias, et al., "You followed my bot! Transforming robots into influential users in Twitter," *Peer-reviewed Journal on the Internet*, vol.18, no.7-1, 2013.
- [128] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in Twitter," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp.839-851.
- [129] C. Freitas, et al., "Reverse engineering socialbot infiltration strategies in twitter," in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2015, pp.25-32.
- [130] J. Guixeres, et al., "Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising," Frontiers in psychology, vol.8, pp.1808-1821, 2017.
- [131]B. Yılmaz, et al., "Like/dislike analysis using EEG: determination of most discriminative channels and frequencies," *Computer methods and programs in biomedicine*, vol.113, no.2, pp.705-713, 2014.
- [132]S. Lewandowsky, U. K. H. Ecker, and J. Cook, "Beyond misinformation: Understanding and coping with the "post-truth" era," *Journal of Applied Research in Memory and Cognition*, vol.6, no.4, pp.353-369, 2017.
- [133]I. Arapakis, M. Barreda-Angeles, and A. Pereda-Baños, "Interest as a proxy of engagement in news reading: Spectral and entropy analyses of EEG activity patterns," *IEEE Transactions on Affective Computing*, vol.10, no.1, pp.100-114, 2017.
- [134] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*, Paris, France: Cambridge university press, 2008, pp.50-85.
- [135] Y. T. Chang, H. Yu, and H. P. Lu, "Persuasive messages, popularity cohesion, and message diffusion in social media marketing," *Journal of Business Research*, vol.68, no.4, pp.777-782, 2015.
- [136] W. M. Huang, et al., "Contagion on complex networks with persuasion," Scientific reports, vol.6, pp.23766-23773, 2016.
- [137] J. Cao, et al., "Automatic rumor detection on microblogs: A survey," arXiv preprint arXiv:1807.03505, 2018.
- [138] T. N. Nguyen, C. Li, and C. Niederée, "On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners," in *International Conference on Social Informatics*. Springer, 2017, pp.141-158.

- [139] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp.1395-1405.
- [140] T. Chen, et al., "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2018, pp.40-52.
- [141]J. Sampson, et al., "Leveraging the implicit structure within social media for emergent rumor detection," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016, pp.2377-2382.
- [142] Q. Liu, et al., "Mining significant microblogs for misinformation identification: an attention-based approach," ACM Transactions on Intelligent Systems and Technology (TIST), vol.9, no.5, pp.50-67, 2018.
- [143] N. Nakashole, and T. M. Mitchell, "Language-aware truth assessment of fact candidates," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp.1009-1019.
- [144] F. Ma, et al., "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in Proceedings of the 21th ACM SIGKDD

- International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp.745-754.
- [145] K. Popat, et al., "DeClarE: Debunking fake news and false claims using evidence-aware deep learning," arXiv preprint arXiv:1809.06416, 2018.
- [146] Guo H, et al., "Rumor detection with hierarchical social attention network," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp.943-951.
- [147]M. H. Gad-Elrab, et al., "ExFaKT: a framework for explaining facts over knowledge graphs and text," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019, pp.87-95.
- [148] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," arXiv preprint arXiv:1808.00033, 2018.
- [149]F. Doshi-Velez, and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
- [150] A. T. Nguyen, et al., "An interpretable joint graphical model for fact-checking from crowds," in *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [151] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol.8, no.3, pp.489-508, 2017.