APSIPA Transactions on Signal and Information Processing, 2022, 11, e42

Overview Paper

Combating Misinformation/ Disinformation in Online Social Media: A Multidisciplinary View

Mauro Barni[1], Yi Fang[2], Yuhong Liu[2*], Laura Robinson[3], Kazutoshi Sasahara[4], Subramaniam Vincent[5], Xinchao Wang[6] and Zhizheng Wu[7]

[1]*Department of Information Engineering and Mathematics, University of Siena, Italy*

[2]*Department of Computer Science and Engineering, Santa Clara University, USA*

[3]*Department of Sociology, Santa Clara University, USA*

[4]*Department of Innovation Science, Tokyo Institute of Technology, Japan*

[5]*Director of Journalism and Media Ethics, Markkula Center for Applied Ethics, Santa Clara University, USA*

[6]*Department of Electrical and Computer Engineering, National University of Singapore, Singapore*

[7]*School of Data Science, Chinese University of Hong Kong, Shenzhen, China*

ABSTRACT

Recently, the viral propagation of mis/disinformation has raised significant concerns from both academia and industry. This problem is particularly difficult because on the one hand, rapidly evolving technology makes it much cheaper and easier to manipulate and propagate social media information. On the other hand, the complexity of human psychology and sociology makes the understanding, prediction and prevention of users' involvement in mis/disinformation propagation very difficult. This themed series on "Multi-Disciplinary Dis/Misinformation Analysis and Countermeasures" aims to bring the attention and efforts from researchers in relevant disciplines together to tackle this challenging

*Corresponding author: Yuhong Liu, yhliu@scu.edu.

Received 07 May 2022; Revised 24 July 2022

ISSN 2048-7703; DOI 10.1561/116.0000127

problem. In addition, on October 20th, 2021, and March 7th 2022, some of the guest editorial team members organized two panel discussions on "Social Media Disinformation and its Impact on Public Health During the COVID-19 Pandemic," and on "Dis/Misinformation Analysis and Countermeasures – A Computational Viewpoint." This article summarizes the key discussion items at these two panels and hopes to shed light on the future directions.

1 Introduction

With the pervasiveness of social networks and media, digital information (health, climate, political, news articles, etc.) can be easily created and shared online by individuals, including people and bots. This significantly changes how humans access, search and perceive information. More people are making their economic, political, health and daily life decisions by referring to online information due to its convenience and low cost. Yet, online social media has become a battleground for malicious attackers to fabricate and propagate massive amounts of disinformation [45], with the participation of massive groups of people online (leader-follower feedback loops). This is often referred to as the "info-demic." The uncontrolled rapid propagation of disinformation can lead to severe consequences such as financial losses, hostile online environments, damaging people's confidence in trusting online information, and even endangering people's lives.

Existing data confirms that almost all major online social networks have severely suffered from massive propagation of disinformation that is deliberately fabricated without support of fact [39]. For example, during the COVID-19 pandemic, massive misinformation and fake news have generated confusion and influenced the public's perception of risks [1, 12, 40]. Deepfake technologies, which manipulate the digital media by replacing a person's face or body with another person's likeness, have emerged as a powerful tool to spread disinformation. For example, a recent deep-fake was posted on a Ukrainian news site of Volodymyr Zelensky telling soldiers to surrender during the war with Russia [21]. Even worse, a recent study found that disinformation spreads significantly farther, faster, deeper, and more broadly than the truth in all categories of information citebib43.

Dis/misinformation is a complex problem which cannot be well addressed in one traditional discipline. There is an emerging need for researchers of multiple disciplines (e.g., computing, communication, journalism, social psychology, law, etc.) to have a joint forum to understand the disinformation propagation mechanism, how people evaluate the authenticity of online information, and investigate potential solutions to combat info-demic.

The guest editorial team, with eight world leading experts working on related fields, has proposed a themed series at APSIPA Transactions of Signal and Information Processing on "Multi-Disciplinary Dis/Misinformation Analysis and Countermeasures." On October 20th, 2021, some of the guest editorial team members organized a panel discussion on "Social Media Disinformation and its Impact on Public Health During the COVID-19 Pandemic," at the IEEE Global Humanitarian Technology Conference (GHTC). Furthermore, on March 7th, 2022, the U.S. local chapter of APSIPA brought together several members from the guest editorial team to discuss "Dis/Misinformation Analysis and Countermeasures – A Computational Viewpoint."

We envision that the promotion of a broader understanding of the problem from different disciplines can facilitate conversations and collaborations, and integrate wisdom from the communities towards potential countermeasure strategies. Therefore, we summarize the opinions of experts from multiple disciplines, with both the panelists and the guest editorial team members as the authors of this paper. The rest of the paper is organized based on the key questions discussed in the two panels.

2 What Do Your Respective Disciplines Bring to the Social Media Dis/Misinformation Issue?

The manipulation of digital information and media is not a new problem. However, the recent advancement in computing technologies and the popularity of social media make such manipulations much easier and more influential than ever before, which poses great challenges to defense solutions. The involvement of human factors also makes the social media dis/misinformation issue more complex and dynamic. On the other hand, the same technologies used by the manipulation side can also be leveraged by the defense side. The massive data available on online social media also provides great opportunities to drive the advancement of relevant disciplines. In this section, we aim to provide readers with a holistic view of the problem by summarizing the opinions of researchers from different disciplines.

2.1 Multimedia Forensics (MMF) Techniques

[**Barni**] Fabrication of fake images and even videos is not a new problem by itself, given that the use of photomontages to propagate fake information and propaganda is as old as the history of photo cameras [8]. However, with the advent and diffusion of digital media, the problem has reached a new level, and researchers have started studying possible countermeasures [7].

Early works date back to about 15 years ago and were relying on subtle statistical traces to expose evidence of tampering. These include the detection

of multiple JPEG compression artifacts as an indirect proof of tampering, the detection of the traces left within the images by the camera that was used to create them, the exploitation of interpolation traces left within the images whenever an image region is resized, rotated and so on. Works based on geometric and semantic evidence have also been developed since the very beginning. These works include the analysis of image shadows and reflections, the adherence of the structural elements of the images to prospective geometry, the analysis of the coherence of ambient light, anomalous body motions like absence of eye blinking or unnatural head poses, etc.

The alarm, and the consequent search for remedies, made a further huge leap in 2017, when the news that a Reddit community used an open-source, AI-based, face-swapping application to digitally insert classmates, friends and celebrities into pornographic videos has travelled all around the world (https://tinyurl.com/ybnymyqj). This practice has immediately been banned, however a new path was opened and in only 5 years the production of fake media by means of modern AI tools has spread at an unstoppable pace. Not only have AI tools made media tampering accessible to the wide public, they also raised the quality of fake media up to a level never seen before. If the progress of AI has contributed to raise the alarm about media trustworthiness to a level never seen before, researchers have started looking at AI as a possible solution to restore the credibility of digital media. In the last few years, a new class of MMF tools relying on Convolutional Neural Networks (CNN) and other Deep Learning (DL) architectures has appeared [4, 5, 30]. Nowadays, AI-based MMF tools have reached a performance level that makes it possible to identify fake media contents with very high accuracy in the highly controlled conditions typical of laboratory experiments. Yet, the detection of fake media in the wild, when nothing or very little is known about the tampering techniques used by the forger, is still an open problem, whose solution will keep researchers busy for many years to come.

[Wu] Beyond video deepfake, audio files are also facing manipulations. Audio deepfake is usually combined with video deepfake, but producing fake speech to mimic a target speaker. Similar to face and fingerprint, voice contains a speaker's information and is naturally linked to a person's identity. Voice also contains messages that a speaker would like to communicate to others. Therefore, if a manipulated voice from a public figure spreads on the Internet, it could cause misinformation propagated to the public. For example, in the recent deepfake of Ukrainian President Volodymyr Zelensky asking Ukrainian soldiers to surrender during the war with Russia [21].

There are several approaches to create a fake speech for a target speaker, including audio splicing [26], and audio deepfake [20, 47]. Audio splicing is to copy, paste and delete audio segment to make up a new audio segment that the target speaker never speaks. Audio deepfake is usually created using

DL techniques. Audio deepfakes can be fully synthesis or partially synthesis. Audio splicing or audio deepfake is usually combined with video deepfake. In the deepfake of Ukrainian President Volodymyr Zelensky, the speech from Volodymyr Zelensky is fake and created mimic Volodymyr Zelensky.

Since audio can be mixed with background noise, music or applying different codec compression, it is challenging to detect. Most of the studies attempt to employ machine learning or deep learning algorithms to identify artifacts from the audio or to build a robust detector.

2.2 Natural Language Processing Related Techniques

[**Fang**] Prior research works have extensively studied the possibilities and limits of utilizing Natural Language Processing (NLP) for identifying misinformation. A typical NLP pipeline that tackles misinformation detection consists of three main stages: text preprocessing, feature extraction, and machine learning model training. Basic text preprocessing includes tokenization and stemming. In some cases, Part-of-Speech (PoS) tagging is also useful by tagging each token with its appropriate part of speech such as noun, pronoun, adjective, and so on. This will help the subsequent feature extraction component with understanding the context, since certain words may have different meanings in different contexts. After the preprocessing, the text needs to be converted to vectors of numbers to illustrate the linguistic features of the text, which is called feature extraction. There are a number of feature extraction methods, with the two most popular ones being bag of words (BoW) with Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings [22].

Many machine learning approaches to misinformation detection are to formulate the task as a supervised learning problem. The output target is either categorical (misinformation vs. authentic) or a numeric score of truthfulness. Training data can be collected from various sources including news websites, social media, and search engines. Data collection and annotation is one of the major challenges for automatic misinformation detection due to the cost of constructing a set of high-quality labeled data. It needs to annotate whether one piece of text, claim or statement is true or false according to the ground truth. In general, annotations can be conducted through expert journalists, crowd-sourcing workers, fact-checking websites, and industry detectors. The existing datasets for misinformation detection can be categorized as containing short statements such as LIAR [43] and FEVER [41], posts on social network sites (SNSs) such as BuzzFeedNews [37] and CREDBANK [25], and entire articles such as FakeNewsNet [36] and BS DETECTOR.

Despite the success of supervised models, the performance of these methods usually relies on having a large amount of labeled data for model training. However, to obtain reliable labels often requires much time and labor. In addition, news spreads on social media at very high speed when an event

happens, only very limited labeled data is available in practice. In consequence, unsupervised, semi-supervised [14] or weakly-supervised methods have been proposed [16].

2.3 Social Network Analysis

[**Liu**] One category of existing studies approach the misinformation issue from a social networking perspective [48, 49]. In particular, these studies consider the social network as a graph, with each user as a node and the interactions among them as edges.

Based on graph theory, some studies are conducted to analyze the node level behaviors, community level interactions (i.e., interactions among different nodes), and context information. For example, node level analysis [49] considers the indegree/outdegree of a node, the node's behavior patterns over time, the properties of its neighboring nodes (i.e., who the accounts are connected with). The community level analysis may focus on homophily and interactions among a group of nodes by assuming coordinated malicious nodes tend to be more densely linked groups. In addition, analysis on context information may include the timing and location of the posts and users involved. For example, as a large portion of the influential misinformation is produced in a highly coordinated manner, the initiator and early engagers accounts tend to demonstrate high homogeneity in their geographic location and posting/engaging timing. These features can then be fed into machine learning models to detect malicious users accounts.

Another aspect is to focus on the information propagation patterns, which aims to mainly address the who (i.e., who participates in the propagation) and how (i.e., how the propagation evolves) issues. Specifically, diverse information propagation models are proposed to understand how truthful contents and misinformation is propagated. One popular category of models is the compartmental models borrowed from epidemiology [50], which divides the population into compartments, such as susceptible (S), infectious (I), recovered (R), and exposed (E), with the assumption that individuals in the same compartment has the same transition probabilities/rates to other compartments. Examples include the SIR, SIS, SEIR, etc. Due to the simplicity, these models are often adopted to model social information propagation. Linear Threshold (i.e., IC) model assumes that a node is influenced by each of its neighbors, and if the number of its infected neighbors exceeds a certain threshold, this node will also be infected. The threshold intuitively represents the different latent tendencies of nodes to believe the information when their neighbors do. The Independent Cascading model [33] assumes that an infected node at each discrete step randomly chooses one of its neighbors (i.e., target neighbor) for propagation with a certain probability of success, which is independent of the propagation history. The process runs until no more infection is possible. The branching

process, originating in probability theory, has been extensively adopted to model information propagation that follows tree structures. Specifically, the branching process models a system of individuals which live for a random time and, at some point during their lifetime or at the moment of death, produce a random number of children as the next generation.

Last but not least, individual human behavior patterns have been verified by extensive studies as the driving forces for the dynamics of many social, technological and economic phenomena. In conventional studies, human activities are mainly modeled by the Poisson process. In 2005, Barabasi [3] first proposed that human behaviors followed non-Poisson statistics, characterized by bursts of rapidly occurring events separated by long periods of inactivity. Since then, many studies have adopted the heavy tailed bursty human dynamics [10, 17, 18, 23, 24] and found that such human dynamics can significantly influence the information propagation speed.

[**Sasahara**] Some studies suggest that the structure of social networks affects the spread of (mis)information. The frequency (popularity) of retweets of hashtags or memes on Twitter is known to be a heavy-tailed distribution, and the same distribution can be obtained by changing the time scale to daily, weekly, or monthly. To replicate these properties, the spread of memes was simulated in an artificial society of agents with finite attention using the actual social network structure of Twitter, and the aforementioned heavy-tailed distribution was reproduced [44]. Considering the fact in the context of the spread of fake news, it is suggested that one could spread fake news much more than others, regardless of the quality of the news. Moreover, it has been reported that even low-quality news can spread on a large scale under information overload in a similar network setting [29].

2.4 Journalism and Media Ethics

[**Vincent**] Journalism and Media Ethics (JME) is an applied ethics sphere scrutinizing decision making, design and norms in news publishing and news distribution. This area is complicated because "news" is hard to define, and news publisher behavior is hard to draw clear technical boundaries around. Definitions and categorization have already run into trade-offs. For example, narrow definitions of who is a news publisher or "what is news" can either exclude diverse voices at the cost of eliminating fake news sources, or broader definitions can bring everyone into news feeds, including counterfeit and junk publishers. Applied ethics in this area deploy normative analysis and ethical frameworks to drive design and decision making for both "journalism" standards in the social media era, and news distribution ethics, which is in the realm of social media and search products.

Figure 1 is a rough illustration of news flow in media ecosystem from origin to users (on news feeds), with decision-making vocabulary used both in journalistic work (publishing) and platform products (distribution).

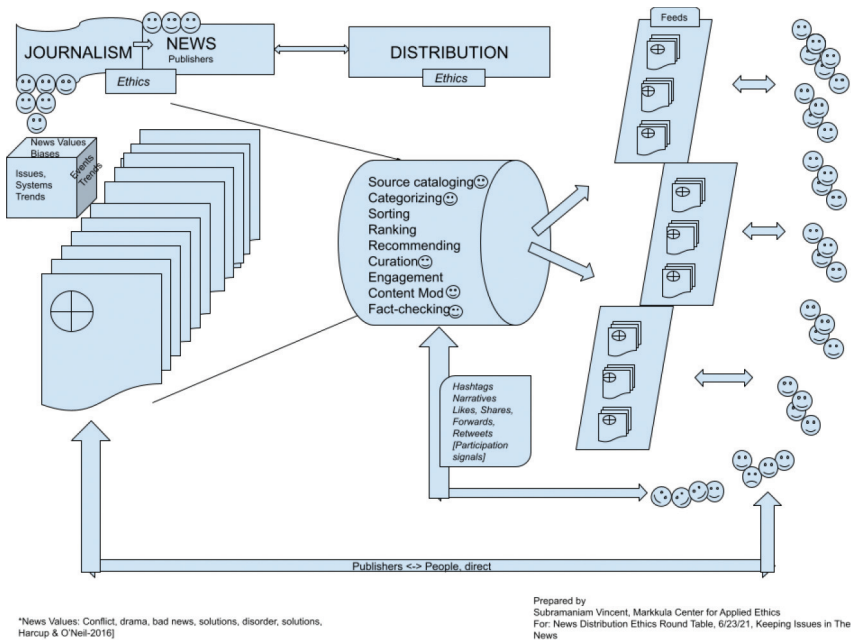

Figure 1: Illustration of news flow in media ecosystem.

JME brings the following to this area more specifically. A clearer articulation of the differences using normative vocabulary (rights, justice, dignity, power, democracy, transparency, and harm mitigation) between ethical and unethical storytelling. This includes accurate representations of people, realities and facts, in stories. It gets into inclusive sourcing, characterizations, anti-stereotyping, and holistic accuracy as opposed to atomic accuracy at the fact level, and so forth. This approach can be used to build new ethics-vocabulary based datasets that have features to explore for unique signals about producers of "news" – like content who do not use the latest journalistic standards or exploit well known weaknesses in journalistic routines and format to deceive the public. JME also helps identify top-down design questions in news publisher curation and algorithmic news feed design for products and offer framework analysis and justifications on how to decide to uprank, downrank, or eliminate particular types of content or sites from a news feed. Often the justification development is through a deliberative multi-stakeholder facilitation that JME can convene. JME can also validate existing best practices in select players in

the media industry – both publisher-producers and distributors – who may already be mitigating harm empirically, but whose operating principles can be called out with more universal clarification and appeal for others to use.

2.5 Social Sciences

[**Robinson**] Finally, taking a social sciences perspective to these issues allows us to consider factors implicated in information literacies. Most of the social science research around mitigating misinformation has centered on training schemes geared towards making information consumers less susceptible to disinformation before they encounter it [9]. A broad range of experimental interventions in psychology have tried out techniques for "inoculating" information consumers in various ways. For instance, in one such intervention subjects are confronted with a "game" in which they learn how purveyors of disinformation exploit manipulative techniques like fake experts and emotional language to seduce information consumers [2]. Another set of behavioral interventions attempts to more tightly couple information sharing decisions to individuals' perceptions of information accuracy. This type of intervention is based on the insight that information consumers often share information items they know to be false. Such interventions have shown that, when subjects are primed to condition sharing on their perception of information accuracy, they are much less likely to share information they know to be untrue or inaccurate with other individuals in their communication networks.

3 What are the Reasons for the Popularity of Dis/Misinformation in Online Social Media?

One critical step to fight against disinformation and misinformation in online social media is to understand the fundamental reasons for its popularity. As briefly touched on above, the involvement of human and social factors make today's dis/misinformation issue very unique and challenging. Therefore, in this section, we would like to particularly focus on the human, social and journalism ethics perspectives to understand the underlying reasons for the popularity of dis/misinformation in online social media.

3.1 Social Sciences

Perspectives drawn from across multiple social science fields are increasingly being brought to bear on the issue of disinformation and misinformation, particularly as they relate to patterns of social behavior. Among the most important approaches to issues of disinformation and misinformation concerns social patterns in the uptake of disinformation and misinformation items and

the propensity for differently situated individuals to produce, share, or consume such items. Existing research regarding disinformation and misinformation propagated through online sites point to important dynamics such as the outsize roles played by small minorities of "superproducers" and "superconsumers" of disinformation on the Internet [13], which contributes to the filter bubble discussed below (Section 3.2).

In addition to relative capacity to produce and/or consume misinformation, other important social science work has been conducted on both information overload and manipulation. Such work goes in-depth to study information uptake among younger consumers and suggests that the sheer volume of news and quasi-news items can overwhelm many news and media consumers. Once news and media consumers become overwhelmed, their capacity and/or willingness to engage becomes diminished. Further, they become unable or unwilling to exert the cognitive effort to distinguish more from less reliable sources of information [2].

Other social science approaches highlight the modes of reception of different forms of disinformation among differently situated information consumers. One of the key findings from this body of research has been that a large array of manipulation techniques can be employed to convince news consumers of the truth of even fabricated items. For example, "fact softening" procedures whereby disinformation creators undermine the believability of verifiable facts frequently help disinformation to gain traction among news consumers. Finally, from the perspective of psychology it is well-established that susceptibility to disinformation depends at least partially on the degree of the concordance between the consumer's preexisting beliefs and the informational items they encounter. Individuals will more readily believe untruths when they comport with preexisting beliefs than when they conflict with such beliefs [9].

3.2 Online Social Interactions

[Sasahara] From the online social interaction perspective, when people connect only with those who have similar values and interests on an SNS and repeat this process, a closed information environment is created, in which only similar information is circulated and the same opinions are heard over and over again. This type of closed information environment is called an "echo chamber." Echo chambers have been observed in various SNS platforms [6], which may be a factor in the spread of mis/dis-information. Sasahara *et al.* [35] built a computational model of opinion dynamics on an SNS to obtain mechanistic insights into the emergence of echo chambers. (A simplified version is available online [34]). The simulation results over time are shown in Figure 2, where t represents time steps of the simulation. Specifically, subplot A shows the average diversity of messages on the screen, measured using Shannon entropy with the opinion range divided into 10 bins; subplot B shows the temporal

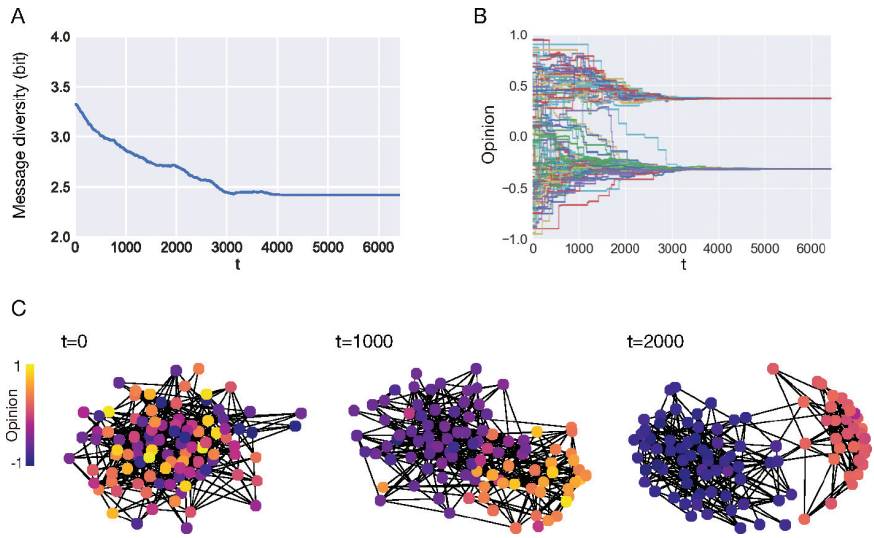

Figure 2: Emergence of echo chambers in an agent-based simulation [35].

changes in population opinions; and subplot C shows the temporal changes in the social network structure. From these three subplots, we observe that over time, (1) the diversity of information that users see in their environment decreased (i.e., they came into contact with only similar information); (2) opinions among users became polarized; and (3) social networks among users were divided into two groups with different opinions. These results suggest that an echo chamber spontaneously emerges on an SNS. It was also found that the speed of its formation is accelerated under conditions of strong social influence and high unfollowing frequency. It is possible that our behavioral tendencies to be socially influenced by similar others and to socially disconnect from dissimilar others combined with the existence of SNS platforms that facilitate these behaviors, thereby promoting the formation of an echo chamber in the information environment.

Another problem in the information environment is the "filter bubble," in which mis/dis-information is likely to circulate. A filter bubble is also a closed information environment formed by algorithms that have learned the user's personal information and only filter information that may be of interest to the user [28]. Many online services, such as search, advertisement, and news feeds, have such information filters embedded in them. Not only do these filters hide information about opinions and values that differ from their own, but they are also personalized for each user, creating a situation (bubble) where it is difficult for everyone to access common facts and truth. Since the prediction accuracy in the information filter is expected to increase in the future, there is an increasing

risk that personalized mis/dis-information could be created from the estimated personal information and misused to guide one's thoughts and behaviors.

It is also clear that bots (i.e., automated accounts controlled by a set of algorithms) influence the spread and amplification of mis/dis-information. In the early stages of the COVID-19 pandemic, numerous coronavirus-related disinformation circulated [46]. Figure 3 is a retweet network, where the links represent retweets and the nodes represent bots, with green, in particular, representing normal bots and red representing malicious bots that frequently spread mis/dis-information. This figure shows an echo chamber situation, with little information flow between clusters of conservative Trump (@realDonaldTrump) and liberals Joe Biden (@JoeBiden) and Hillary Clinton (@HillaryClinton).

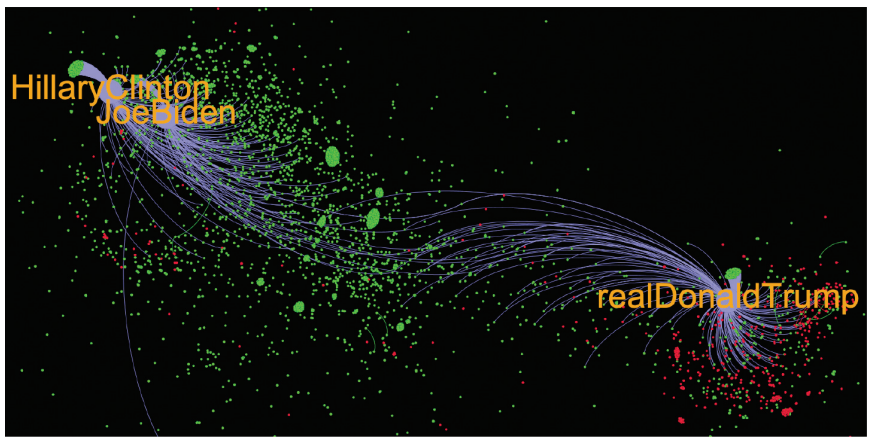

Figure 3: Retweet network consisting of social bots and echo chamber.

3.3 Journalism Ethics

There are multiple reasons and they are all interconnected together to amplify each other in particular news cycles and scenarios.

3.3.1 Anti-democratic Leader-follower Agency

In America, some major producers of disinformation, especially political actors in democracies, see disinformation as a means to end. Stem the loss of cultural power. They see that democratic power-sharing around equality is causing the kind of change in society that they term radical, subversive to authority, and are remaking the image of the country that places some groups in a first-amongst-equals status. They prefer, instead, to return back to traditional norms that are psychologically more satisfying or at least slow the process

down. These actors know the psychological stresses, anxieties, and most importantly fear of "the other" that sections of the public carry. They also know that the enhanced virality mechanisms of social media run frictionlessly on leader-follower feedback loops around specific narratives. They combine the emotive pull of the message for their audiences with social media design to use participative engagement to bring reach. It is a win-win for such politicians and their followers to build a kind of walled narrative morality that is hard for democratic discourse to penetrate. They are able to intentionally release disinformation over long periods of time through narratives that people feel they belong in and can identify with. They can "flood the zone" until their accounts are taken down. Even people who know that stories are disinformation or are suspicious about its veracity acknowledge that they still share it because they involve signaling identity and group membership belonging, or they involve hate-based rejection of the other tribe, so it gives people satisfaction to say "the other" is "wrong" and "evil."

3.3.2 Social Media Design Offering Virality for Outrage and Disinformation

The very design of the powerful like, retweet, share and forward features as universal affordances on all ports for all accounts since 2009 has allowed all actors the same access to social media's reach infrastructure, especially the virality features introduced after 2009 [SV 1]. Under such circumstances, tribal messaging – whether it is based on falsity or truth is incidental – was simply more likely to spread faster because it plays to emotions and identity (the fast thinking part of the brain) and now the deliberative-slower part of the brain. [SV 3]

3.3.3 Profit-making Possibilities Ensure Supply is Paid for

False news spreads quickly and also faster than the truth. Provocateurs, amateurs, and propaganda operations can all make money on the spread of disinformation or a mix of non-facts and facts in outrage inducing media, using factors 1 and 2. There has been mushrooming of digital campaign expertise on the supply side that can ensure that content is continually produced and inserted into social media along appealing false narratives, and such campaigns may pay for themselves until content moderation policies take down such accounts. After 2018–19, many platforms have taken stronger action, but the mere fact that Facebook took down over 70,000 QAnon accounts only after the January 6th 2021 insurrection is proof about how pervasive the supply side is.

3.3.4 Content Overload/Infinite Supply

There is a massive overload of newslike content on digital surfaces and attention is scarce. This compounds the problem. Fast scrolling makes it easier to

discover the content themes and narratives that best fit our brain's satisfaction-and-reward seeking tendencies.

3.3.5 Journalism's Weaknesses

The journalistic news ecosystem has its own weaknesses. When major news websites (nodes on the social media networks) broadcast newsworthy actors' unfounded claims (even if to critique or to cover as part of impartiality or both-sides-ism), it brings legitimacy through amplification to fringe followers. They then are able to create clips and out-of-context nuggets that further feed back into the social media ecosystem.

4 What are the Major Challenges Facing Your Respective Disciplines?

Addressing the dis/misinformation issue is not a trivial task. There are various challenges in different disciplines. In this section, we aim to discuss the major challenges facing each specific discipline.

4.1 Journalism Ethics

[Vincent] There are challenges to both journalism and journalism ethics. One challenge is deeply intrinsic to journalism. Despite the technical nature to some of the practice, journalism is a largely cultural occupation, very unlike technology and the sciences. In culture, norms are often set and exemplified by the political, social, and professional elite. Journalism has always had an elitist bias (favoring the elite in quoting, carrying views, expertise, giving framing power and narrative setting to the elite in stories). Recently, in the American call to norms, post #Metoo and #GeorgeFloyd and #BLM, journalism ethics' primary advocacy to journalism is to pull the practice away from elitism and democratize it by directing toward inclusion and diversifying sourcing. However, this is easier said than done. At the grassroots level today, there is substantial distrust, distaste and cultural (moral values-based) backlash against working journalists [15]. Also, people who deeply believe in disinformation laced narratives are not suitable for being sources to journalists merely in the name of diverse sourcing. That leads to more amplification risks. This places an extra ethical burden on journalists on how to diversify and democratize their sourcing paradigms, under daily and weekly story deadlines. This is particularly hard because two things are going on. One, a culture war is being fought upstream of media on liberal democracy "having gone too far." Two, disinformation narratives are a tool in that culture war and hence all of the false narratives that show downstream on social media surfaces have

tremendous agency, unending force, and continuous supply. It is not possible to shut down cultural conflicts overnight because the culture war is manifesting in political, election-rules, redistricting and voter eligibility battles on the ground. The former CEO of Reddit said poignantly that "the Internet is the MAIN battleground for our culture wars" [19]. Journalists are having to cope with this situation, which demands not mere devotion to factual accuracy, but also other types of accuracy. Accuracy of representation and narrative. This calls for truth-based-framing and commitment to inclusion of the lived experiences of people, which is a kind of non-expert non-elite truth too, particularly of those who are recently economically marginalized by globalization (rural white people in America) and those who are historically marginalized (Black, Native American and people of color).

An underlying systemic issue that complicates ethics is that professional boundaries in journalism are self-drawn by industry and not demarcated by the technical aspects of the work of licensing of the profession itself. Anyone can be a journalist with some training and that has for long been a strength. Lawyers, doctors, architects, engineers, and others can often bring domain experience into journalism and do excellent reporting, editing and Pulitzer prizewinning impactful investigations. So this is both a strength and a weakness. The Internet era has exposed this weakness much more because of the relative boundarylessness of journalistic work and the straightforward mimicability of the news article format (headline, picture, text, and video). This has made it easy for anyone to post newslike content and gain circulation amongst followers, create impressions and outrage. This places a deeper and self-imposed burden on AI-based media products such as social media news feeds, search news surfaces, where there is often a combination of algorithmic sorting and human curation and publisher registration processes at work. Media product design in the disinformation and culture-war era has ethical burdens that did not exist pre-2009 when enhanced virality mechanisms (later introduced by Facebook, Twitter, others) did not exist. Gray area websites that do not peddle only disinformation but often mix factual and non-factual stories can use outrage to drive up traffic and trends that social media news feed recommender systems are contending with. Strongly human-curated news products such as Apple News do not have the same set of news quality and false narrative propagation problems as social products such as Facebook, Twitter, and Whatsapp.

The third deeper change for journalism and media ethics is not new to this area, but is particularly felt here in hard and painful ways. Applied ethics operates at the normative end of conversations, deliberation, and decision making. Technology that builds media products operates largely with empirical strength. It serves business purposes primarily and the public square (or democracy) incidentally or at most a second priority. The private technology oligopolies (Google literally commands 85–90% of the search market outside of China; Facebook, Twitter, Snapchat, and Pinterest are a handful of companies

dominating the global market on social media) have created a tension between the normative advocacy vs proprietary empirical knowledge that unfairly favors the latter. Social media and search companies have amassed massive amounts of data on content and behavior that is unavailable (or difficult to get easily) to academic researchers. To allow deeper normative questions to be answered by public academics in the media space rigorously, subsidiary empirical questions often need to be framed and answered for example to understand the costs vs benefits or harm mitigation effects on a utilitarian scale for a new feature. The empirical questions are themselves interdisciplinary and require social psychologists, engineers, UX designers, computer scientists and journalists to work together to define a problem narrowly, propose a hypothesis and objectively observe any phenomenon or effects of tweak carefully, and systematically and replicably. Interdisciplinary science is complex in its own right. But much empirical knowledge about the lessons from how technology firms are already fighting disinformation (despite the culture wars upstream) remains trapped under the hood in tech companies. As a result of pressure from academia for transparency, there have been recent moves to release more data (using differential privacy) and annual reporting. But the suggestive and prototyping power of research and innovation in media technology still favors "big tech" and that in itself is a challenge for democracy.

4.2 Natural Language Processing

[**Fang**] One of the main challenges in NLP for misinformation detection is the lack of high-quality annotated misinformation datasets. The existing labeled datasets are often in a relatively small scale and domain dependent. Consequently, unsupervised and semi-supervised methods become important for misinformation detection by leveraging unlabeled data. Transfer learning and meta learning can also be explored to transfer knowledge from the domains with more abundant data to the low-resource domains with scarce data.

Misinformation can be intentionally created to mimic the language style of authentic information to mislead readers to believe false information, which makes it difficult to detect based on the content itself. Some studies rely on auxiliary information. For example, to identify whether a given topic or content is worth checking can be a potential research direction for facilitating mechanisms to help detect misinformation. In addition, some other related tasks such as document summarization and stance detection can be used to enhance misinformation detection. Text summarization can be applied to identify the main theme of the information and stance detection can recognize the argument of the information creator, which could be useful signals for the classification of misinformation. Another challenge in misinformation detection is to generate an alarm for misinformation at the early stage before it spreads widely, which can significantly help intervene misinformation.

4.3 Multimedia Based Dis/Misinfo

[**Barni**] Multimedia forensics investigation techniques based on AI technology have reached a maturity level allowing almost perfect detection of fake media, including deepfakes, under strictly controlled conditions. When the goal of the manipulations is given, the techniques possibly used to create the forgery are at least partially known, when the characteristics of genuine images and videos are also known, including the source of the images/videos, the compression algorithm possibly used to encode them, and when the quality and size of the analyzed contents is good enough, fake media can be spotted with an accuracy that approaches 100%. What a pity, then, that these conditions hardly ever occur. Detecting forgeries in the wild, taking into account the huge variability of the operating conditions characterizing the forensic analysis, is a largely unsolved problem, whose solution is going to require a long lasting effort. All the more that, as of now, the effort put in this struggle by MMF researchers is surpassed by far by the opposite effort of researchers and media companies aiming at developing more and more accurate and efficient tools for the generation of synthetic media. Other problems making the battle against the diffusion of fake media hard are:

Necessity of huge and representative training datasets: AI architectures based on Deep Learning (DL) must be trained on huge amounts of labeled data which is often difficult to gather in MMF applications. For instance, it is difficult to build a dataset with tens of thousands of photomontages of good visual quality, let alone videos and audios. This problem is exacerbated by the availability of a wide number of tools the forger may use and by the lack of knowledge typical of forensic scenarios (e.g., lack of information about media source and processing history). The necessity of dealing with situations that could not be foreseen at training time is also a typical problem of MMF.

Need for interpretable and accountable tools: The black-box nature of AI techniques (CNN in primis) makes it difficult to interpret the results of the analysis and understand why a certain decision is made. While this is a general and recognized problem, in most MMF applications the accountability of the results provided by the forensic tools is simply vital (think about the use of the results of the analysis in a court). The possibility of interpreting the results of AI-based-techniques would also avoid that the forgery detectors make their decision based on so-called confounding factors that are not directly related to the problem at hand, thus opening a security breach that can be exploited by counterfeiters to avoid that their manipulations are detected.

Lack of security: Many works have shown how easy it is to generate adversarial examples capable of deceiving pattern recognition techniques based on CNN [27]. As a matter of fact, the basic technology behind the use of AI to create forged media [11] stages a race of arms between two networks, a discriminating network (somewhat playing the role of the forensic analyst) and a generator network (playing the role of the counterfeiter), and ultimately relies on the capability of the generating network to evade the detection capability of the discriminator. Understanding and ensuring the security of AI tools is a crucial problem, if AI has to be used under the intrinsically adversarial conditions typical of MMF applications.

4.4 Information Propagation on Social Media

[**Liu**] Although extensive studies have been proposed to model information propagation in social media, due to the heterogeneity and complexity of human psychology and behaviors, there are many remaining challenges. Some examples are as follows.

From the individual user aspect, many existing studies over-simplify human users' status as two discrete values (i.e., either infected or uninfected) and ignore users' continuous psychology and behavior pattern changes.

From the users' social interaction aspect, many existing studies model the propagation from one user to another as a random variable. The lack of understanding of the underlying physical mechanisms makes it very hard to customize models for individual level characteristics and interactions. More importantly, when a user's decision is influenced by multiple connections and factors at the same time, it becomes even more challenging to integrate these multiple random processes and retrieve accurate results.

From the propagation process, existing models often reflect the dynamic process as a series of snapshots (i.e., discrete time) or a stochastic process (e.g., heavy-tail human dynamics). There is a lack of mechanisms to understand the process as a continuous time process which is more natural to model human decision makings/changes.

From the information input aspect, most current propagation studies ignore the propagated contents, e.g. how different it is from each individual's belief, how it is propagated (i.e., the timing and frequency), etc. More importantly, these aspects may influence each other, making the propagation process very complex. For example, the propagation process and users' social connections may influence each other. On the one hand, how close two users interact may influence the propagation speed between them. On the other hand, what has been propagated between two users will also influence their interactions in the future (increase or decrease in their trust).

From the intervention aspect, most existing studies limit the intervention problem as given a limited budget, how to find the most influential nodes in the network to initiate the "truthful" information propagation. They can neither consider real-time propagation patterns, nor dynamically change defense solutions based on the rapidly evolving attack progress.

Therefore, some major questions remain open. (1) Can we mathematically model the real-time propagation patterns? If yes, can we change the propagation speed, for example, to slow down the propagation process? (2) Will the propagation speed change during the propagation process? If so, under what conditions can we dissipate it and eventually eliminate it? (3) Will the uninfected nodes, who receive the propagated information, generate any feedback to influence the infected nodes, leading to backward propagation? (4) Will the input patterns, such as strength or frequency, influence the propagation? If so, can we propagate "truthful" information with appropriately designed input patterns to interfere with the propagation of false information?

[**Sasahara**] The background of high information uncertainty and social unrest, such as the COVID-19 and the Ukraine situation, and the development of information technology that allows for the advancement and mass production of "fakes," have led to the current infodemics. The increase in the amount of information created with a high percentage of uncertain information (i.e., low signal-to-noise ratio) will have a negative impact on daily life, economic activities, and even democracy if false information is frequently misinterpreted as fact or even if facts are not believed to be true. In such information overload, our limited attention and the structure of social networks make it probable that even low-quality information, regardless of its truth or falsity, can spread on a large scale. In addition, there are multiple factors that amplify non-credible information, such as human factors—emotions, cognitive biases, and social influences, as well as system-level effects arising from human-platform interactions, such as echo chambers and filter bubbles, which are complex and intertwined.

Our information ecosystem continues to evolve. If AI and other technologies are misused in the future for economic and political agendas and used as weapons for hacking elections, guiding public opinion, manipulating impressions, and information warfare, it could spiral out of control. To counter the trend toward the advancement and mass production of "fake" that is currently occurring, it is necessary to introduce a mechanism to deliver reliable information transparently and a mechanism to encourage slow information sharing into the information ecosystem. In order for social media to become a true public space that embraces diverse voices in the future while avoiding a flood of fake news, information technology needs to be restructured in a human-centered form to build trust.

5 What do You Envision as the Future Directions or Promising Solutions?

Moving forward, some promising models, technologies and solutions have been proposed and are currently under investigation. We would like to share the visions of the authors from both technology and human perspectives.

5.1 Technology Perspective

[**Barni**] Active authentication techniques open some promising directions for the future. With these techniques, the devices or the AI models used to create a synthetic media, or to manipulate an existing document, are engineered in such a way to ease the subsequent identification of the generated media document as synthetic or fake. The quality of the document would be good enough to fool a human observer, thus maintaining the effectiveness of the tools that produced it, however, the media manipulation techniques would be accompanied by the release of additional tools to distinguish the contents they generate from natural media, and, possibly, to trace them back to the specific device or AI model which generated them. This resembles a classical watermarking scenario, wherein a watermark is indissolubly embedded within the media at creation time, and possibly used afterwards to authenticate the media, verify its integrity and trace it back to its origin. This approach would mark a drastic paradigm change with respect to current media authentication solutions based on passive MMF techniques, since authentication would be achieved with the active help of the party which trained the media-generation, or media-processing, network. As a matter of fact, training a model for high-quality synthetic media generation (or processing) requires a huge investment that only big players can afford. To avoid that such an effort is exploited by third parties for malicious purposes, like opinion manipulation, disinformation campaigns, defamation, it is in the interest of those who incur the burden of the training to design their models in such a way to ease the identification of the origin of synthetic media, thus avoiding that they are used for illegal purposes.

[**Barni and Wu**] The role of deep learning (DL) in the never ending struggle between counterfeiters and media analysts is twofold. On one hand, DL continuously enriches the toolbox available to counterfeiters with new, more efficient and higher quality techniques, allowing the generation of better forgeries at a lower price. On the other hand, DL models provide better analysis capabilities, allowing to identify the traces left by the counterfeiters into the tampered media with great accuracy. What used to be a battle of witness is now becoming a battle between AI and AI, with the part that can rely on larger resources (economical, but not only) that is likely to win the race of arms.

Let's take audio DeepFake detection as an example. Figure 4 illustrates the process to modify a source speech to sound like a target speaker, which can be used to generate audio DeepFakes. DL can be used to implement the conversion function for more realistic voices. On the other hand, the process of generating realistic voices is not perfect, and each step will introduce some artifacts, deep learning can be used to learn the differences between real speech and synthetic speech. This is how deep learning helps to detect DeepFakes.

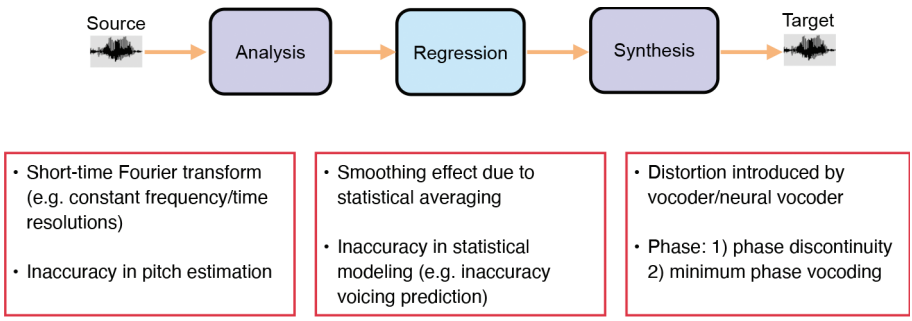

Figure 4: Deep learning as a tool to generate audio Deepfakes.

5.2 Human Perspective

[**Sasahara**] Closed information environments on SNSs, such as echo chambers and filter bubbles, facilitate the spread of misinformation. To fix this, it is important to develop information technologies that mitigate our innate tendency to see only what one wants to see and connect only with those one wants to connect with. In addition, there are issues to be addressed as a society. The rumor intensity formula known in social psychology suggests that the strength of a rumor is the arithmetical product of the significance of the subject matter and the level of uncertainty about the available information (Allport & Postman, 1965). According to this formula, for important topics to the public, such as the Ukraine situation and the COVID-19, it is necessary to reduce the ambiguity of the situation to reduce the spread of misinformation. To this end, scientists and research institutions should continue to present accurate information along with evidence against unscientific information. Finally, information literacy is a critical skill for every one of us. Fact-checking organizations around the world are actively verifying doubtful news stories and making their findings open to the public. Although the information on these sites is not real-time, it is helpful to understand the characteristics of misinformation that occurred in the past. Acquiring information literacy by

learning from these sites is equivalent to taking a "vaccine" against info-demic and essential to avoid easily sharing information.

[**Robinson**] Finally, future directions for research in the social sciences include many of the solutions indicated by scholars of digital inequality and digital literacies stemming from unequal access to resources, education, and IT training [38]. Many social scientists bridge larger concerns with inequalities and social reproduction to these processes. For these scholars, issues of digital inclusion, divides, and inequalities must be considered to fully understand the agency with which individuals and groups encounter online mis- and dis-information [31, 32]. An important area of digital divide studies targets its gaze on information literacy as both the cause and result of larger social inequities. Scholars working in this vein distinguish relationships between digital literacies and digital competencies stemming from unequal access to digital skills. For them, digital literacy and information literacy, conceptualized as facility with procedures for finding and understanding online information, is the mechanism associated with the capacity to discern "true" news items from "false" news items. Both general digital literacies – defined as familiarity with the Internet – and specific digital literacies – defined as the level of comprehension of newsfeed algorithms

are associated with the rate at which individuals can discern true from false factual items. As both kinds of digital literacies are linked to what digital inequality researchers such as Hargittai, Ragnedda, and others have termed "second-level" digital divides. On top of first-level access divides, these second-level divides are conceptualized as skill divides critical to informational literacies. These divides in digital skills are correlated with the capacity for truth discernment and necessarily reflect foundational divides in digital training, skill acquisition, and confidence. Indeed, social science studies of the dynamics of disinformation and misinformation diffusion increasingly suggest that the lack of digital skills among disadvantaged users may make them unwittingly spread information which they may not realize is unreliable or false.

6 Suggestions

Last but not least, defending against mis/disinformation is a long term fight. Regardless of how advanced technologies can be developed, eventually the general public and our younger generations are the major parties getting involved in the battle. It is essential for them to be well educated and protected. Therefore, we would like to provide them with some practical suggestions.

6.1 Any Suggestions to the General Public?

[**Liu**] With the massive propagation of misinformation, it is challenging for individual users to make accurate differentiation. Fortunately, the community has taken active efforts and made some helpful resources and tools available to the general public. For example, there are some fact check websites available, e.g., International Fact-Checking Network (IFCN), Factcheck.org, Snopes, PolitiFact. In addition, major social media platforms also take their actions to actively check the factual aspect of social media posts, label misinformation and inform users about it, and suppress mis/disinformation propagation. With the availability of these resources and tools, however, at the end of the day, it still relies on individual users being more responsible and thinking twice before they like, retweet, and comment on a random social media post.

6.2 Any suggestions to the younger generation?

[**Robinson**] As the authors here have shown, the causes and consequences of misinformation and disinformation are both vast and heterogeneous. While there are no easy solutions to the many facets of the problem, one solution is clear to many: investing in our children. A number of scholars and organizations are taking up the cause to create and disseminate earlier and better educational tools to teach children how to better navigate the challenges of mis- and disinformation. For example, in a recent piece in *The New York Times* [42], Tugend draws connections between digital and analogue media to argue that while news or media literacy are not new, they are increasingly urgent in the United States given our rapid polarization and political divides. The author, Tugend, highlights a number of issues to be surmounted including the lack of commonly applied core standards adopted across the United States given the right of states to determine their own curriculum. At the time of writing only fourteen states required media literacy for K-12 students. Nonetheless, Tugend also draws attention to the efforts of organizations such as the News Literacy Project to develop educational tools such as the "News Lit Quiz: How newsliterate are you" (see https://newslit.org/tips-tools/how-news-literate-are-you-quiz/ and https://cor.stanford.edu/curriculum/). According to Tugend's account:

> "*Researchers focused on two major skills. The first is lateral reading. It encourages readers who come to an unfamiliar website to refrain from exploring the site more deeply until they have opened other tabs and found other websites to help them determine the authenticity or reliability of the newly discovered site. The other skill is click restraint. Ideally, users would resist the impulse to click on the first results that appear in say, a Google search, until they have scanned the full list for credibility and then click selectively.*"

Unfortunately, such training is more likely to be made available to resourced students, thus bringing us back to the importance of digital inclusion as perhaps our best chance to combat informational inequalities leading to the creation, dissemination, and uptake of misinformation. From this angle of vision, skills deficits have been linked to socio-economic disadvantage may need to be addressed both for children and remediated for adults whose education may have omitted them. From this perspective, while these skills are highlighted for K-12 students there is no reason not to believe that older learners could benefit as well from lifelong learning taking similar approaches to community and senior centers, community college programs for adults, and others committed to the cause.

7 Conclusion

Through the discussion with experts from different disciplines, we can conclude that combating mis- and disinformation is a very challenging and long term task. Solutions from one discipline alone cannot completely address this issue. Collaborations from different disciplines are critical. Furthermore, there is an arm race between the malicious users and the defender. The rapid development of technology itself may provide helpful tools to detect/suppress mis/disinformation propagation, but the same technology can also be used by malicious users to facilitate faster propagation of mis/disinformation. The tools are neutral, but depending on who is using them, the consequences can be very different. The battle may last long. Persistent efforts are critical.

Acknowledgment

The authors would like to thank APSIPA Transactions on Signal and Information Processing for approving this themed series on "Multi-Disciplinary Dis/Misinformation Analysis and Countermeasures." The authors also acknowledge the support of IEEE GHTC 2021 conference to host the panel discussion on "Social Media Disinformation and its Impact on Public Health During the COVID-19 Pandemic," and the APSIPA U.S. Local Chapter to host the panel discussion on "Dis/Misinformation Analysis and Countermeasures – A Computational Viewpoint," the active participation of the audience, APSIPA and many individuals/groups who have helped to promote the themed series and the panels.

References

[1] D. Baines, R. Elliott, and Others., "Defining Misinformation, Disinformation and Malinformation: An Urgent Need for Clarity During the COVID-19 Infodemic," *Discussion Papers*, 20(06), 2020, 20–6.

[2] J. P. Baptista and A. Gradim, "Understanding Fake News Consumption: A Review," *Social Sciences*, 9(10), 2020, 185.

[3] A.-L. Barabasi, "The Origin of Bursts and Heavy Tails in Human Dynamics," *Nature*, 435(7039), 2005, 207.

[4] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and Non-aligned Double JPEG Detection Using Convolutional Neural Networks," *Journal of Visual Communication and Image Representation*, 49, 2017, 153–63.

[5] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 2016, 5–10.

[6] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The Echo Chamber Effect on Social Media," *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[7] E. Delp, N. Memon, and M. Wu, *Special Issue on Digital Forensics*, No. 2, 2009.

[8] H. Farid, "Seeing is Not Believing," *IEEE Spectrum*, 46(8), 2009, 44–51.

[9] A. Gampa, S. Wojcik, M. Motyl, B. Nosek, and P. Ditto, "logical Reasoning: Ideology Impairs Sound Reasoning," *Social Psychological and Personality Science*, 10(8), 2019, 1075–83.

[10] K.-I. Goh and A.-L. Barabási, "Burstiness and Memory in Complex Systems," *EPL (Europhysics Letters)*, 81(4), 2008, 48002.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 2014, 27.

[12] M. Gottlieb and S. Dyer, "Information and Disinformation: Social Media in the COVID-19 Crisis," *Academic Emergency Medicine*, 2020.

[13] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake News on Twitter During the 2016 US Presidential Election," *Science*, 363(6425), 2019, 374–8.

[14] G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis, "Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2018, 322–5.

[15] *Hard News: Journalists and the Threat of Disinformation*, 2022, https://pen.org/report/hard-news-journalists-and-the-threat-of-disinformation/.

[16] S. Helmstetter and H. Paulheim, "Weakly Supervised Learning for Fake News Detection on Twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2018, 274–7.

[17] J. L. Iribarren and E. Moro, "Impact of Human Activity Patterns on the Dynamics of Information Diffusion," *Physical Review Letters*, 103(3), 2009, 038702.

[18] C. Lee and D. J. Wilkinson, "A Hierarchical Model of Non-homogeneous Poisson Processes for Twitter Retweets," *Journal of the American Statistical Association*, 2019, 1–22.

[19] E. Lopatto, *The Former CEO of Reddit Would Like You All to Stop Bickering Online*, 2022, https://www.theverge.com/2022/4/15/23026971/reddit-elon-musk-moderation-yishan-wong.

[20] H. Malik and R. Changalvala, "Fighting AI with AI: Fake Speech Detection Using Deep Learning," in *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*, Audio Engineering Society, 2019.

[21] A. Mathews, "Preservation of the Evidentiary Moving Image," 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, 2013, 26.

[23] B. Min and K.-I. Goh, "Burstiness: Measures, Models, and Dynamic Consequences," in *Temporal Networks*, Springer, 2013, 41–64.

[24] G. Miritello, E. Moro, and R. Lara, "Dynamical Strength of Social Ties in Information Spreading," *Physical Review E*, 83(4), 2011, 045102.

[25] T. Mitra and E. Gilbert, "Credbank: A Large-scale Social Media Corpus with Associated Credibility Annotations," in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9, No. 1, 2015, 258–67.

[26] X. Pan, X. Zhang, and S. Lyu, "Detecting Splicing in Digital Audios Using Local Noise Level Estimation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, 1841–4.

[27] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-box Attacks Against Deep Learning Systems Using Adversarial Examples," *arXiv preprint arXiv:1602.02697*, 1(2), 2016, 3.

[28] E. Pariser, *The Filter Bubble: What the Internet is Hiding from You*, Penguin UK, 2011.

[29] X. Qiu, D. FM Oliveira, A. S. Shirazi, A. Flammini, and F. Menczer, "Limited Individual Attention and Online Virality of Low-quality Information," *Nature Human Behaviour*, 1(7), 2017, 1–7.

[30] Y. Rao and J. Ni, "A Deep Learning Approach to Detection of Splicing and Copy-move Forgeries in Images," in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2016, 1–6.

[31] L. Robinson, S. R. Cotten, H. Ono, A. Quan-Haase, G. Mesch, W. Chen, J. Schulz, T. M. Hale, and M. J. Stern, "Digital Inequalities and Why They Matter," *Information, Communication & Society*, 18(5), 2015, 569–82.

[32] L. Robinson, J. Schulz, G. Blank, M. Ragnedda, H. Ono, B. Hogan, G. Mesch, S. R. Cotten, S. B. Kretchmer, T. M. Hale, *et al.*, "Digital Inequalities 2.0: Legacy Inequalities in the Information Age," 2020.

[33] K. Saito, R. Nakano, and M. Kimura, "Prediction of Information Diffusion Probabilities for Independent Cascade Model," in *International Conference on Knowledge-based and Intelligent Information and Engineering Systems*, Springer, 2008, 67–75.

[34] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, "On the Inevitability of Online Echo Chambers," *arXiv preprint arXiv:1905.03919*, 2019.

[35] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Social Influence and Unfollowing Accelerate the Emergence of Echo Chambers," *Journal of Computational Social Science*, 4(1), 2021, 381–402.

[36] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media," *arXiv preprint arXiv:1809.01286*, 2018.

[37] C. Silverman, L. Strapagiel, H. Shaban, E. Hall, and J. Singer-Vine, "Hyperpartisan Facebook Pages are Publishing False and Misleading Information at an Alarming Rate," *Buzzfeed News*, 20(1), 2016.

[38] N. Sirlin, Z. Epstein, A. A. Arechar, and D. G. Rand, "Digital Literacy is Associated with More Discerning Accuracy Judgments But Not Sharing Intentions," *Harvard Kennedy School Misinformation Review*, 2021.

[39] K. Starbird, *Information Wars: A Window into the Alternative Media Ecosystem*, 2017, https://medium.com/hci-design-at-uw/information-wars-a-window-into-the-alternative-media-ecosystem-a1347f32fd8f.

[40] F. Tagliabue, L. Galassi, and P. Mariani, "The "Pandemic" of Disinformation in COVID-19," *SN Comprehensive Clinical Medicine*, 2(9), 2020, 1287–9.

[41] J. Thorne, A. Vlachos, C. Christodoulopoulos, and M. A., "Fever: A Large-scale Dataset for Fact Extraction and Verification," *arXiv preprint arXiv:1803.05355*, 2018.

[42] A. Tugend, *These Students Are Learning About Fake News and How to Spot It*, https://www.nytimes.com/2020/02/20/education/learning/news-literacy-2016-election.html.

[43] W. Y. Wang, ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2017.

[44] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition Among Memes in a World with Limited Attention," *Scientific Reports*, 2(1), 2012, 1–9.

[45] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," *ACM SIGKDD Explorations Newsletter*, 21(2), 2019, 80–90.

[46] W. Xu and K. Sasahara, "Characterizing the Roles of Bots on Twitter During the COVID-19 Infodemic," *Journal of Computational Social Science*, 2021, 1–19.

[47] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, *et al.*, "Add 2022: the First Audio Deep Synthesis Detection Challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 9216–20.

[48] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin, "Fake News Propagates Differently from Real News Even at Early Stages of Spreading," *EPJ Data Science*, 9(1), 2020, 7.

[49] X. Zhou and R. Zafarani, "Network-based Fake News Detection: A Pattern-driven Approach," *ACM SIGKDD Explorations Newsletter*, 21(2), 2019, 48–60.

[50] Y. Zhuang and O. Yağan, "Information Propagation in Clustered Multilayer Networks," *IEEE Transactions on Network Science and Engineering*, 3(4), 2016, 211–24.