Fact-check Spreading Behavior in Twitter: a Qualitative Profile for False-claim News

Francisco S. Marcondes[1], José João Almeida[1], Dalila Durães[1,2], and
Paulo Novais[1]

[1] ALGORITMI Centre — Department of Informatics, University of Minho, Portugal
`francisco.marcondes@algoritmi.uminho.pt` and `{jj|pjon}@di.uminho.pt`
[2] CIICESI, ESTG, Polytechnic Institute of Porto, Felgueiras, Portugal
`dad@estg.ipp.pt`

Abstract. Fact-check spread is usually performed by a plain tweet with just the link. Since it is not proper human behavior, it may cause uncanny, hinder the reader's attention and harm the counter-propaganda influence. This paper presents a profile of fact-check link spread in Twitter (suiting for TRL-1) and, as an additional outcome, proposes a preliminary behavior design based on it (suiting for TRL-2). The underlying hypothesis is by simulating human-like behavior, a bot gets more attention and exerts more influence on its followers.

Keywords: chatbot, social agent, fake news, fact check, social media

1 Introduction

The spread of fake news on Twitter is commonplace [4]. Despite the existing efforts for identifying and debunking fake-news, initiatives for enlightening people about facts are still scarce in social-media [8]. This, however, is cornerstone since this is probably the best-suited counter-propaganda strategy [15]. In other words, spam-fighting strategies such as detecting and stopping bots suits for automatic spreading but does not refrain organic spreading. Therefore, efforts such as those of "guardians", people who fight against misinformation in social-media [13], are extremely valuable. However, good-will based guardians cannot face professional and high-technological propaganda structures [15]; for illustration, an estimate proportion for true-news and fake-news is of 1:17 [12].

This paper *objective* is to profile the fact-check spreading behavior in Twitter for the purpose of proposing a viable design hypothesis for an automatic fact-check spreading device (called automatic guardian or A-GUARDIAN) to be tested in near future. The key *contribution* is the raised profile as it clarifies this type of behavior and provides some background and insights for automation.

State of the Art. Fake-news spreading is based on computational propaganda. Propaganda is a fake-news with a political or financial end; computational propaganda is the propaganda carried out by bots and bot-nets [17]. The measures adopted for hinder propaganda is called counter-propaganda; carrying it out

using computer support is called computational counter-propaganda [15]. The most effective counter-propaganda strategy is enlightening people with the truth. There are two dimensions to be considered, the first is related to the content and the second to the spreading strategy.

Content-related counter-propaganda is performed through debunking false-claims with correct information (this is being held by fact-check agencies). These debunking information must reach people, then a spreading strategy must be devised. Nevertheless, there are constraints to be honored. Counter-propaganda cannot use deceptive strategies such as bot-nets with several fake-accounts, *i.e.* it must be ethical. This second dimension, however, is not being tackled [8].

An alternative to bot-nets is to consider "digital influencers" for reference. Digital influencers are people who succeed in influencing people through the content published in their communication channels . Anyway, such behavior is analogous to the "super-spreaders" find in the core of bot-nets [11]. Actually, know a profile to be a bot does not change its influence potential [6], unconsciously yet consistently humans interact with computers as social actors [10]. Yet this requires even simple relationship building features [2].

Methodology. This paper follows the Design Science approach adapted for researching programming. In short, the Ω-knowledge is developed following the TRL steps and the Λ-knowledge is driven by intuition [3]. This paper, as a profile, is aiming TRL-1 (basic principles or the Ω-knowledge). Different methods were applied according to the research needs. Section 2 uses systematic review, sections 2.2, 2.3 and 2.4 are based on case study strategy and section 2.5 is the result of a bibliography survey. Finally, section 3 presents preliminary yet coherent model and design hypothesis for TRL-2 (basic concept for the Λ-knowledge) for the gathered information.

2 Profiling the Fight against Fake-news

2.1 Virtuous-bots Landscape.

A survey evaluating the use of bots on fighting fake-news was presented in [8] within the time-frame between 2015-2017. For including the meantime until July 2019 a quick "re-survey" was carried on by repeating that paper parameters (the query submitted to Scopus was `TITLE-ABS-KEY((chat* OR social OR conversation* OR dialogue) W/0 (system OR agent OR bot) AND (fake-news OR fake news OR misinformation))`). The same results hold, there are few researches concerned with fact-checks spreading [8, 13].

Another survey, this time directed towards Twitter in August 2019, was for identifying profiles that identify themselves as bots and aimed to fight misinformation. The query `search?q=bot fact fake &src=typed_query &f=user` retrieved

---

[3] Intuition has been a controversial due to some authors opposed it to reason , however intuition is the highest skill level attained when the skill is internalized not requiring extensive councious reasoning . Applied to problem-solving, it the ability to see a solution beforehand

3 results of which 1 is currently active. The `@unfakingnews` is a cyborg working mostly by re-tweeting information from credited sources. In turn, the query `search?q=bot fact &src=typed_query &f=user` retrieved 49 results; 19 were classified as "not applicable" due to not being a bot or active. The remaining were classified as NLP-based bots (4); non-sense bots (9); link-spread or re-tweeting bots (5); and fun-facts bots (11). Therefore, 5 out of 49 accounts match the intended behavior of fighting misinformation by spreading factual information yet only two are active `@skeptics_bot` and `@TXLegeFactBot`; both for link spreading.

It was not possible to find virtuous-bots replying to other users' tweets for warning about a fake-news. This may be due the `@DroptheIBot` effect. Roughly, it was a bot when someone said something like "illegal immigrant" it replied with a politically correct suggestion for an alternative. This profile was reported many people, and in the end, suspended by Twitter [3]. Somehow it degenerated into a subverted yet virtuous spam engine.

2.2   A Fact-check Report Profile.

For this paper, a sample of 50 fact-check reports and 7360 related tweets within a time-frame ranging from September 2008 to July 2019 (excluded repetitions), was fetched on July 20, 2019. It was selected "trending" fact-check reports collected from the *Hot 50* page of Snopes' website (www.snopes.com) whose link was queried in Twitter for collecting fact-check link spread tweets *cf.* [13].

The Snopes uses *labels* for classifying its fact-check reports according to its subject and *ratings* for classifying the truth-level of an analyzed claim. This helps in understanding how tweets linked to fact-check reports concentrate. The reports classified as NONE, even being significant, were rejected for this paper.

Figure 1a depicts the most viewed reports are those for debunking false claims. It is depicted in figure 1c that tweets linked with false-claims are the most numerous yet well distributed between the sampled reports.In another hand, figure 1b reveals the most popular subjects being *junk-news*, *politics* and *photos*, however, as depicted in 1d the highest tweet concentration was about *politics*. These same data were also collected one month earlier and four months after the presented sample suggesting similar results.

Therefore, it is safe to suggest that false-claims debunking with political subjects is a popular topic for Snopes' readers both on the website and Twitter.

2.3   A Fact-check Spreading Profile.

False-claim debunking is the rating with better widespread interest distribution. Other ratings common behavior is one or two reports receiving massive attention (measured in the number of tweets) whereas the further being neglected. In this sense, false-claim reports are well-suited for relationship building and working on the counter-propaganda effort at once. Therefore, in order to increase effectiveness, this paper focuses on fact-check spread for false-claims.

The sample for false-claims is composed of 1987 tweets, almost one-third of total sample, scattered through 1483 profiles, a proportion of 1:1.4, suggesting
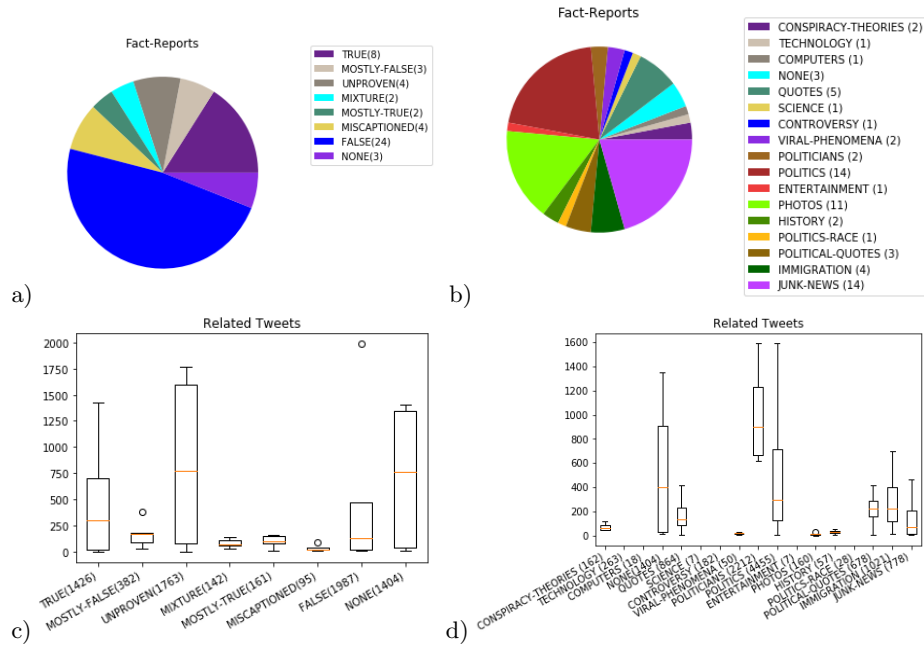
Fig. 1. a) Fact-check reports classified by the claim *rating*; b) Fact-check reports classified by the its subject *labels*; and c) Tweets spreading fact-check reports links classified by the analyzed claim *rating*; and d) Tweets spreading fact-check reports links classified by the report's subject *labels*. For a remark, a fact-check report may receive several subject labels, therefore the total size presented in figure (b) is greater than the size of the sample. The none rating and label are used for uncategorized reports from Snopes.

an absence of bots. These tweets were split into two groups, one for tweets sent in a profile's time-line (872, 44%) another for tweets sent in replying for another profile tweet (1115, 56%). Each group was once again splat (Jaccard > 0.5) in tweets composed almost by the fact-check link (respectively 420 and 544, 48% and 49%) and other in tweets using the fact-check for supporting an argumentation (respectively 452 and 571, 52% and 51%). Highlight that few fake-news links are tweeted for replying [11]. Refer to table 1 for a summary.

It was identified two main behaviors for direct link spreading, one is to insert a link within a tweet and another is by sharing a fact-check from an external source (such as the agency website). External source sharing is usually followed by common words such "fact-check" and "via <source>", *etc.* whereas directly inserting a link is actually just the link supported by the Twitter's thumbnail feature. For this sample, prevailed external source sharing for direct tweets (respectively 119 and 301, 14% and 36%) and link insertion for replying (539 and 5, 48% and 1%). Opinion tweets, in turn, can be divided into four major groups *warning*, *informing*, *explaining*, *commenting* and *judging*. The *warning* includes

	Total	Only Link	Almost Title	Opinion	Replied	Not Replied
Direct	872 (44%)	119 (13.6%)	301 (34.5%)	452 (51.8%)	-	-
Reply	1115 (56%)	539 (48.3%)	5 (0.4%)	571 (51.2%)	-	-
Total	1987	658 (33.1%)	306 (15.4%)	1023 (51.5%)	602 (30.3%)	1385 (69.7%)

Table 1. Tweets spreading fact-checks for false-claims tweets.

tweets like `This is false, debunked a long while ago. Please see <link>` and *judging* includes moral opinion such `you people are absolute garbage <link>`.

There is however a quality difference between direct and reply tweets. Usually, direct tweets are sent directed towards an "abstract entity", then the tone sounds conciliatory. Replying tweets, on another hand, may sound quite aggressive as critics are usually directed to actual people such as who authored or shared the false-claim. For instance, an impersonal `stop lying` is not shocking but by mentioning or replying to someone it may be considered hostile. Therefore, even possessing similar structures, text generation for opinion tweets is not straightforward. A caution is not to reproduce `TAY`'s misbehavior (become an offender shortly after start learning from social media [16]).

Fake-news is quite a sensitive matter, even a deterministic `ALICE`-bot for commenting fake-news, and eventually mentioning or replying people, may easily become troublesome. For this sample 602 (30%) tweets received a reply and 1385 (70%) did not; most of the replies did not receive any further reply neither from the original author nor from a third-person. In addition, non-opinion tweets are responsible for about 50% of link spreading activity. Therefore, considering the risks shown by `TAY` and `@DroptheIBot`, further research on commenting and replying is required and therefore these subjects were postponed.

2.4   A Profile for the Human Behavior.

Observing the sampled bots and some fact-check agencies it can be realized them tweet fact-checks as plain tweets whereas human usually includes around four words for expressing an emotion . Within Internet text messages, emoji, letter capitalization and the number of dots are usually related on expressing emotions.

By simulating human behavior, it is more likely for a bot to exert influence within its followers [6], therefore an "asceptical" false-claim fact-check spread may sound artificial and falls into the uncanny valley [9]. In this situation it is more likely for a tweet create rejection than calling attention and invite for a careful reading. Therefore, properly simulating human behavior may be a strategic cornerstone on fighting fake-news. For profiling and enabling proper human-like text generation, the sampled phrases of direct only-link tweets, excluded repetitions, presented in figure 2.

2.5   The Enemy Behavior.

Fake-news spreading bots are designed with two major concerns. One is to efficiently spread fake-news and the other is to avoid social-media surveillance.

'#FakeNews', '#repeal2A', '#stopsharingfakenews', '6-29-19', 'Actually no', 'Another hoax...', 'COPY AND PASTE!', 'DEBUNKED!!!!!', 'FACT CHECK ', 'FAKE QUOTE', 'FALSE', 'FALSE!!!!!', 'False', 'False quote', 'False...', 'False/hoax', 'Gone viral yet again:', 'HOAX!!!', 'Hoax!', 'Hooks! Sorry.', 'Its Fake ☺ ', 'Just an fyi', 'Just. Stop.', 'Kayleena: not true', 'Listen Up!', 'NOT DEAD!', 'NOT TRUE ', 'Nah ', 'No one got herpes', 'No.', 'PSA', 'y'all:', 'Please read...', 'Please... Just stop...', 'Saw that coming...', 'Stop falling for hoaxes', 'Stop lying.', 'Terry Crowl', 'The Fonz is ok.', 'The answer is no', 'This is 2019 af', 'Viga Hall Please note', 'Worth reading and watching:', '', '× False.', '☺'.

Fig. 2. Text snippets used within the tweet data-set.

Therefore, the presented behavior for these bots overlaps those concerns into a coherent behavior. For instance, following circadian rhythms helps in avoiding surveillance [4] whereas choosing a centered or inflammatory attitude is related to a bot's effectiveness [7].

The strategy is often based on "super-spreaders" within the core of a bot-net who posts the same fake-news up to thousands of times by targeting, through replies and mentions, users with many followers (this is called amplification) [11]. This is especially common for the early spread (before organic sharing becomes prevalent). The expectation is to trigger a rumor cascade (people retweeting or re-claiming a fake-news) [14] by exploring some cognitive biases such as information overload, confirmation bias, motivated reasoning, social trust, *etc.* Highlight, it is more likely to a human spread a fake-news than a true-news [14].

Since social media profiles become more influencing after a popularity threshold [1], another common strategy is to start deception only after a threshold. For this strategy proper cultural and psycho-biological behavior plays a major [6]. Bots are tailored for targeting these triggers and exert influence [11].

3   A Feature Design for the A-GUARDIAN

The basic principles considered for this paper were discussed in section 2. For a summary, section 2.1 found that by one side there are few computational counter-propaganda research and applications for social media. Also, the few instances found are only plain link spreading bots probably due to the @DroptheIBot effect).

The result for section 2.2 is that false-claims debunking with a political subject is a popular topic for the Snopes' readers. For section 2.3, since reports about false-claims have a better interest distribution than other ratings, this paper focused on this rating. Due to difficulties in text generation and replying to sensitive matters such as fake-news as those presented by @DroptheIBot and TAY, opinion and reply text generation require further research being postponed for another paper. For direct tweet spread, it prevailed sharing from external sources behavior. Section 2.4 discusses the need for a bot to earn trust and popularity to exert influence, which requires relationship building and to avoid the uncanny valley. It also presented a set of human tweet samples for discussion.

The two major concerns for fake-news bots discussed in 2.5, is to avoid social media surveillance and effectively spread a fake-news. This is achieved by simulating human behavior, setting text generation towards a target audience,

sending the same fake-news several times for amplification aiming to trigger rumor cascades (the @DroptheIBot effect is not a concern for fake-news bots).

3.1 Basic Concepts (TRL-2)

The surveyed and sampled data presented in section 2 was summarized as the requirement list presented in table 2. Requirements are the way that programming science distinguishes a research design from a regular design. A regular design requirement is the result of a negotiation whereas the research design requirement is not. The result then is evaluated according to how far the design is from an ideal requirement (or design description).

Therefore, this proposal is not for a full social-bot but just to one feature: *direct almost-link fact-check sharing for debunking reports about false-claims*. Although it requires a logged-in Twitter account for running, this feature is not for a Twitter-bot. This is due to data shown most of the human direct link spread is performed through external sources; therefore, *like*, *reply* and *retweet* behaviors are out of scope for this feature.

FEATURE: **Direct almost-link false-claims debunking fact-check sharing**	
# **Requirement**	**Line**
1 Given a list of fact-check reports to share (feed-list), the A-GUARDIAN shares each report through the agency's sharing feature.	1.4
2 Given a data-set with a desired tone, the A-GUARDIAN generates a text snippet for to be inserted with the link in the tweet.	1.5
3 Given a list of target profiles and a feed-list, the A-GUARDIAN selects who to mention within a tweet (it could be none).	1.6
4 Given a fact-check reports, the A-GUARDIAN select those to compose the feed-list for the next iteration.	1.2
5 After a share-event, the A-GUARDIAN sleeps for a certain time	1.7
6 After a sharing iteration, the A-GUARDIAN sleeps for a certain time	1.9

Table 2. Feature and associated requirement list, extracted from TRL-1 (section **??**)

The sharing feature for the requirements in table 2 is presented in the algorithm 1. This, as any bot, is a thread within a continuous loop that eventually is put to sleep. The proposed algorithm, given the provided profile, is quite straightforward however, these simple measures draw the bot nearer to human behavior, compared to average bots as shown [4]. The preconditions that the snippetDataset and targetProfiles cannot be empty means that it is not possible to this bot for spread information without generating snippets nor mentioning people. In short, this strategy is embedded in this bot's essence which in turn is based on the raised profile. The proposed feature is composed of four operations to fulfill its responsibility placed into a generic Handler class:

– chooseFeed(factCheckList) – This function aims to accomplish requirement #4 however, for understanding about repetition and spacing data must be collected. For a proportion, *politics* topic composes at least 25% of a feed-list.

Algorithm 1 The A-GUARDIAN Prototype Algorithm Description.

Require: assertTrue[(social_media, logged_in), (factCheckList, updated)];
Require: [factCheckList ∧ snippetDataset ∧ targetProfiles] ≠ ∅;
 1: **loop**
 2:       feed ← handler.CHOOSEFEED(factCheckList)
 3:       **for all** report **in** feed **do**
 4:              media.SHARE(report,   /
 5:                            handler.BUILDSNIPPET(snippetDataset),   /
 6:                            handler.SORTMENTIONS(targetProfiles))
 7:            thread.SLEEP(handler.SLEEPINGTIMEFOR('share'))
 8:       **end for**
 9:       thread.SLEEP(handler.SLEEPINGTIMEFOR('iteration'))
10: **end loop**
Ensure: ∀ *feed* ∈ *feedList* ⊂ *bot's timeline*
Ensure: assertTrue(social_media, logged_out)

- `buildSnippet(snippetDataset)` – see *Snippet Generation*.
- `sortMentions(targetProfiles)` – This function aims to accomplish requirement #3 however, there are still data to be collected for understanding the mentioning patterns. The target profiles are those with a larger number of followers.
- `sleepingTimeFor(evt=['share', 'iteration'])` – This function aims to accomplish requirements #5 and #6 however, more data should be gathered for understand the *mean time between shares* and the *sharing rhythm* in a circadian sense.

Snippet Generation. This discusses the `buildSnippet(snippetDataset)` function for accomplish the requirement #2. The aim is to exert influence, which requires relationship building as a target. Therefore, a proper human-like snippet generation strategy is a cornerstone. As presented in the figure 2 the sampled tweets are loose phrases yet loaded with sentiments. Handling short texts is a difficult task due to the short amount of information for working with [5], this is even worst for snippets since they are even more reduced texts that may or not be within proper lexical-syntactic structures and often filled with onomatopoeia. Given these constraints, a caution is to avoid *overtechnology* [4].

For a design approach, a text generator may follow either a deterministic or a statistical approach. Based on the figure 2 the statistical approach would favor textual patterns, requiring a supervised learning procedure to keep making sense (and avoiding a `TAY`-like misbehavior); the deterministic approach would favor diversity by randomizing template structures yet requiring a CBR procedure for evolving. Also, the snippet generator can either simulate an individual profile or a general profile; the statistical approach would suit for the first and the deterministic for the second. This paper aims the second since the sampled snippets were not linked with specific profiles.

---

[4] Overtechnology is an anti-pattern similar to *overengineering* or *overuseofpatterns* and related to *bleedingedge* and *goldplating*. It is the act of designing an artifact to be more "technological" than the necessary for its intended use, often due to marketing purpose or technological obfuscation

1. [{☺ } | it's | the answer is | again | actually | another] ({false | hoax | not true | no | fake | debunked | lie | fact-check} | nah);
2. [in | this is] <date>;
3. [{please} | just | y'all | worth|{.}|{!}] ({read | stop | listen up} | copy and paste);
4. and [{.}|{!}|{☺ }| a{f} | #FakeNews | #StopSharingFakenews] for tweet ending.

Fig. 3. Text snippet grammar derived from the snippet data present in figure 2.

The grammar presented in figure 3 was generated based on patterns extracted from figure 2. A suitable grammar-based mechanism for generating these snippets is to raffle the number of words or punctuation marks to be included before and after the link (up to four), then, for each word, it raffles the word's case and for each punctuation mark the amount to be used. Some tweet instances: `NAH` `<link>` `afff` ☺, `Hoax` `<link>` `#StopSharingFakenews`, `.... !!` `<link>`, `<link>` `Please please !!!!` and `DEBUNKED` `<link>`. The sentiment interpretation for a snippet is left for the tweet reader.

4   Conclusion and Research Hypothesis

Fake-news spreading bots have been using several approaches for deception in social-media for propaganda effort. The counter-propaganda cannot use most of these tactics in order to remain ethical. Despite the existence of some shared approaches, due to their diverse aim, computational propaganda and counter-propaganda must have distinctive strategies for achieving their goal.

The design proposal presented in algorithm 1 is based on the profile presented in section 2 and, as a result, it draws nearer to human behavior considering the properties presented in [4]. In short, the expectation is that algorithm 1 by simulating human behavior allows people to "recognizes" it as "peer" creating a social connection enhancing the counter-propaganda effort.

The hypothesis is that *a fact-check spread bot gets more attention and exerts more influence on its followers by simulating human-like behavior compared to plain "only link" tweeting* currently widespread in fact-check agencies and true-news spreading bots. The algorithm 1 summarizes these paper findings.

For future works, it is expected to gather the missing information in order to design the lacking operation for providing a complete description of such behavior. In addition, it is expected to implement the proposed design in order to collect data to better understand the dynamics of fact-check spreading in Twitter and eventually improve the bot effectiveness.

Acknowledgments

References

1. AIELLO, L. M., DEPLANO, M., SCHIFANELLA, R., AND RUFFO, G. People are strange when you're a stranger: Impact and influence of bots on social networks. In *Sixth International AAAI Conference on Weblogs and Social Media* (2012).

2. BICKMORE, T. W., AND PICARD, R. W. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact. 12*, 2 (June 2005), 293327.

3. BROOKER, P. My unexpectedly militant bots: A case for programming-as-social-science. *The Sociological Review* (2019), 0038026119840988.

4. FERRARA, E., VAROL, O., DAVIS, C., ET AL. The rise of social bots. *CACM 59*, 7 (2016).

5. JINGLING, Z., HUIYUN, Z., AND BAOJIANG, C. Sentence similarity based on semantic vector model. In *Proceedings of the 2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (Washington, DC, USA, 2014), 3PGCIC '14, IEEE Computer Society, pp. 499–503.

6. LUCAS, G. M., BOBERG, J., TRAUM, D., ARTSTEIN, R., GRATCH, J., GAINER, A., JOHNSON, E., LEUSKI, A., AND NAKANO, M. Culture, errors, and rapport-building dialogue in social agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (New York, NY, USA, 2018), IVA 18, Association for Computing Machinery, p. 5158.

7. LUCERI, L., DEB, A., BADAWY, A., AND FERRARA, E. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of The 2019 World Wide Web Conference* (2019), ACM, pp. 1007–1012.

8. MARCONDES, F. S., ALMEIDA, J. J., AND NOVAIS, P. A short survey on chatbot technology: Failure in raising the state of the art. In *International Symposium on Distributed Computing and Artificial Intelligence* (2019), Springer, pp. 28–36.

9. MORI, M., MACDORMAN, K. F., AND KAGEKI, N. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine 19*, 2 (2012), 98–100.

10. NASS, C., STEUER, J., AND TAUBER, E. R. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), ACM, pp. 72–78.

11. SHAO, C., CIAMPAGLIA, G. L., VAROL, O., YANG, K.-C., FLAMMINI, A., AND MENCZER, F. The spread of low-credibility content by social bots. *Nature communications 9*, 1 (2018), 4787.

12. SHAO, C., HUI, P., WANG, L., JIANG, X., FLAMMINI, A., MENCZER, F., AND CIAMPAGLIA, G. Anatomy of an online misinformation network. *PLoS ONE 13* (2018).

13. VO, N., AND LEE, K. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research 38* (New York, NY, USA, 2018), SIGIR '18, ACM, pp. 275–284.

14. VOSOUGHI, S., ROY, D., AND ARAL, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.

15. WALLER, J. *Strategic Influence: Public Diplomacy, Counterpropaganda, and Political Warfare*. Institute of World Politics Press, 2009.

16. WOLF, M. J., MILLER, K., AND GRODZINSKY, F. S. Why we should have seen that coming: Comments on microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc. 47*, 3 (Sept. 2017), 54–64.

17. WOOLLEY, S., AND HOWARD, P. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford Studies in Digital Politics. Oxford University Press, 2018.