On the design of a misinformation widget (Ms.W) against cloaked science [UNDER REVIEW]

David Arroyo, Alberto Gómez-Espés and Santiago Palmero-Muñoz
ITEFI-CSIC

Sara Degli-Esposti*
IFS-CSIC

***Abstract*—Amongst all types of fabricated information travelling on open social networks (OSN), scientific misinformation, or *cloaked science*, is particularly dangerous. Here we present the design of the TRESCA misinformation widget (Ms.W), which is both a methodology and a toolbox for investigating disinformation operations leveraging scientific communications. Ms.W follows a man-in-the-loop approach: the methodology takes into consideration ideological and psychological biases, while the toolbox integrates open source intelligence solutions for verifying the accuracy of claims and the credibility of sources. Overall, Ms.W. is a flexible investigative tool offering a REST API for advanced users, who can create and label datasets and add new functionalities to the toolbox.**

1. Introduction

The pandemic has exacerbated the risk that misleading scientific communications can harm public and individual health. The hoax about the benefits of bleach-based alcohol against SARS-COV-2 caused the hospitalisation of hundreds of people and deaths in some countries [1]. Emphasis about the origin of the pandemic in Wuan (China) and claims that SARS-COV-2 was human-made triggered hate speech against Chinese people on Twitter [2]. This last example is associated with a preprint of two scientific reports sponsored by the Rule of Law Society in September 2020 and authored by Dr Li-Meng Yan. The reports claimed to provide evidence that SARS-CoV-2 was deliberately engineered as a bioweapon in a Chinese lab. Despite being discredited by the scientific community [3], Dr Yan has been named whistleblower by some US news media and her theory has been endorsed by the Trump administration [4]. Media expert Dr. Joan Donovan considers the Yan reports an example of *cloaked science*. Cloaked science refers to the use of scientific jargon and procedures to cloak or hide a political, ideological, or financial agenda within the appearance of legitimate scientific research.

Cloaked science is a form of information operation or disinformation campaign that falls into the broader category of *hybrid threats*. Disinformation is deliberately false or misleading information that spreads for political gain or profit, or to discredit a target individual, group, movement, or political party. Misinformation refers to information whose inaccuracy is unintentional and spreads unknowingly. Hybrid threats can include disinformation campaigns, cyber-attacks, inducing political or economic corruption, infiltrating agents of influence, pressuring independent media and buying up critical infrastructures.

Clearly, in the investigation of information operations and hybrid threats we need to dig into the motivations and reasoning of adversaries. Thus, as in the case of infiltrators and sockpuppets [5], in order to disentangle dis- from mis-information we need a blend of human and machine intelligence in order to assess information veracity. This task is especially challenging when it comes to scientific information.

The complexity of science and its numerous controversies create the right breeding ground for fabricating scientific falsehoods starting from half-true arguments. The rise in the number of articles published before peer-review (pre-prints), the presence of retracted scientific articles, and the proliferation of predatory scientific journals also contribute to the weaponisation of science.

Self-proclaimed experts can reach large audiences by mixing pseudo-scientific myths and conspiracy theories. For example, the *Center for Countering Digital Hate* identified in in March 2021 twelve influential anti-vaxxers with large numbers of followers and producing high volume of content against COVID19 vaccines [6].

Despite the fact that fighting digital deception is a corporate and societal priority, there is still a lack of ”solutions (especially automated ones) that can mitigate the ease that existing online infrastructures allow adversaries to engage in deceptive content creation and dissemination” [7]. We respond to this call with a procedural contribution in the form of a methodology and with a technical contribution, which is the architecture of a toolbox called TRESCA misinformation widget, or Ms.W. After introducing the methodology, we present Ms.W API REST and functionalities and demonstrate their usefulness through a use case. Finally, we discuss our contributions and future directions for research.

2. Blending human and machine intelligence: partial automation of fact-checking in the investigation of cloaked science

On 19 May 2021 the Spanish fact-checking agency Newtral disputed the accuracy of a text circulating on WhatsApp saying that in India only people who were vaccinated were getting infected. The text misquoted an excerpt from an interview with Spanish doctor Amaia Foces, who lives in New Delhi. Was this story the product of poor quality journalism or was it part of an orchestrated cloaked science operation? How can we know it?

In the investigation of cloaked science we need tools for assessing authors’ credibility and reputations besides claim accuracy. To the best of our knowledge, the only solution that focuses on fact-checking scientific claims is *CORD-19 Claim Verification demo* [8]. This solution can be used to assess the veracity of claims in scientific articles, but not the trustworthiness or rationale of the source. As disinformation investigation requires human intelligence for cyberattribution and the evaluation of users’ intentionality, tasks can only be partially automated.

We argue that in investigating disinformation, we need to treat the veracity of claims and the credibility of sources as interrelated elements. Assessing authors’ reputation and credentials can be applied as a predictor of content veracity and can also be used to establish authors’ motivations and worldview. The methodology presented below and the logic of the toolbox are based on this basic assumption: the recursive relationship between claim veracity and source credibility and the need for users’ constant engagement and critical thinking to draw a conclusion.

3. Ms.W methodology

The construction of truth through language is a social practice: as such, its automation can only be partial. Ms.W methodology underlines the active role users need to play in distinguishing true from false information by applying technical tools and critical thinking.

The basic assumption behind Ms.W methodology is the interconnection between claims and sources. By ‘claim’ we mean a statement about reality that can appear on a post anywhere, from a newspaper article to a meme or a video. By ‘source’ we mean the author of a claim, the publisher of the newspaper where the claim appears, or any other individual or relevant entities spreading the message.

The methodology includes various steps for verifying the accuracy of claims and the credibility of sources, while also taking into account ideological and psychological biases. These steps include actions that should be performed with the support of Ms.W toolbox and psychological considerations that should help users leverage their critical thinking ability.

1) **Assessing the credibility of the source.** Where does the post come from? Is the author or source real and credible? What is the motivation/worldview of the author or source?

   a) Verify if the accounts authoring the post are real persons or bots.

      - Ms.W toolbox relies on the Botometer to assess the likelihood of a Twitter account being a bot [9].

   b) If the author claim to be an expert, for example a scientist, verify their credentials.

      - Ms.W toolbox integrates functionalities to search authors' social profiles in OSN and also scientists' profiles on Google Scholar or e-thesis online services (such as EthOS or Teseo).

   c) Check for partisan bias, which is the worldview the source might be implicitly reproducing.

      - Ms.W toolbox helps users assess outlet ideological bias based on mediabiasfactcheck.com categorisation.

   d) Be aware of your own psychological biases. If the person who shared the post is not the author, but a family member or friend, do not trust their judgement simply because you know them well.

2) **Verifying the veracity of the claim.** Do I know enough to judge? What research exists or evidence supports the claim? Was the information or image taken out of context? Does the headline match the content? Does the post or article make you feel really excited or angry? Does the message create distrust or division?

   a) See whether a reputable fact-checking organisation has verified the claim.

      - From Ms.W toolbox send a query to Google Fact Check and skeptics.stackexchange.com

   b) Verify that the source is not reposting old news stories claiming they are timely and relevant.

      - From Ms.W toolbox perform a reverse image search from Google find out the true origin of the picture.
      - From Ms.W toolbox check whether a quote has been misreported based on Wikiquote API and other similar resources such as quotationspage.com.

   c) Determine if the title of the article reflects the story written in it.

      - Use the clickbait functionality available in Ms.W toolbox if you are dealing with multiple articles or read the article and revise its content.

   d) If the post triggers a strong negative or positive emotional reaction, it might be disinformation and an attempt to increase polarisation, division, and distrust among groups.

      - After ensuring that the post is not satirical, be suspicious of messages that try to: undermine the integrity of election systems; spread hate and division based on misogyny, racism, anti-Semitism, Islamophobia, and homophobia; denigrate immigrants; promote conspiracies about global networks of power.

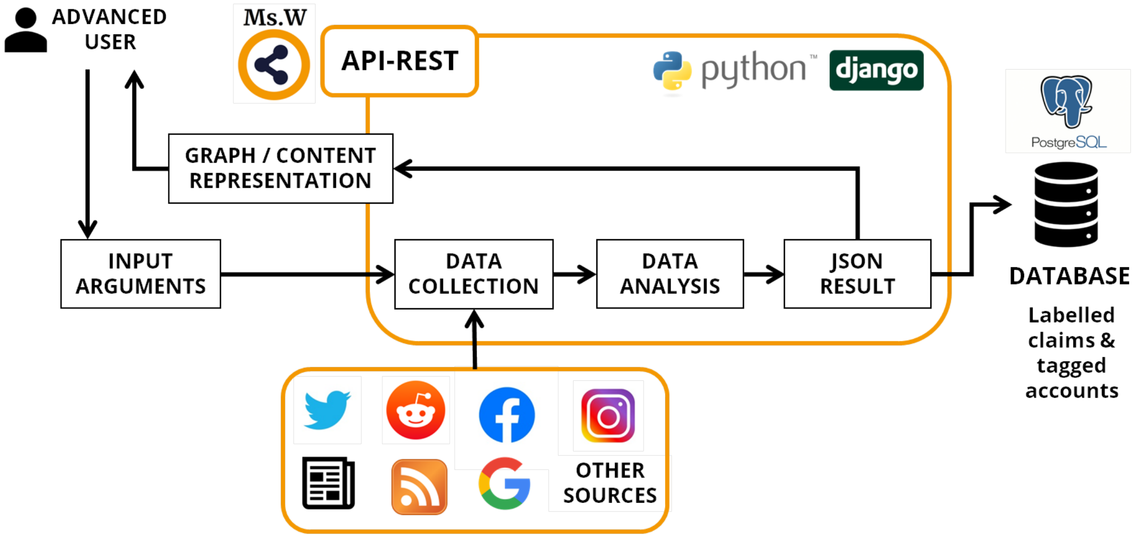


Figure 1. Ms.W API REST for advanced users.

4. Ms.W REST API

In the methodology section we argue that source credibility and claim accuracy are related and need to be assessed in combination. Ms.W REST API includes external and internal tools for dealing with the evaluation of both claim accuracy and source credibility. All functionalities return outputs that can be downloaded in JSON format.

Internal tools, such as the clickbait algorithm, have been developed by this research team based on previous research [10] (which has been extended in our work that is currently under revision [11]).

External tools leverage open-source intelligence (OSINT), which refers to the analysis of publicly available information that may come from media such as newspapers, television and websites and that can help establish the identity, reputation and network of supporters and detractors of an argument or a user account. As shown in figure two, Ms.W includes tools for assessing the presence of bots.

We know that bots play an important role in disinformation campaigns. For instance, between January and April 2020, bots promoting anti-Asian hate speech were highly vocal and hateful (compared to non-bot users) and comprised 10.4 percent of hateful users on Twitter [2]. Thus, we have integrated the Botometer into Ms.W Toolbox to help users assess whether Twitter accounts are real humans or bots. Ms.W also relies on blacklists such as *Stop Funding Misinformation*, Iffy+ and the dataset compiled by [12] to identify malicious accounts.

4.1. Ms.W users' profiles and access privileges

Ms.W envisions three types of users: a super-user or system administrator (admin), an advanced user and a basic user. The admin has all privileges and has access to the all system. The admin can create users' profiles and let them access the REST API. The advanced user accesses the REST API directly and has writing privileges. Advanced users can upload RSS feeds from their favourite news outlets, add definitions to the glossary, and upload labelled datasets. Basic users only have reading privileges and access the REST API through Ms.W frontend, which is under development.

5. Testing Ms.W methodology and toolbox through a use case

In this section we present a use case to show how an advanced user can take advantage of Ms.W toolbox and methodology. Let us assume an user wants to assess the veracity of a claim made on a tweet about a scientific finding about COVID19. The tweet shown in the chart is artificially constructed, but reflects the format and content of real tweets used in a real investiga-

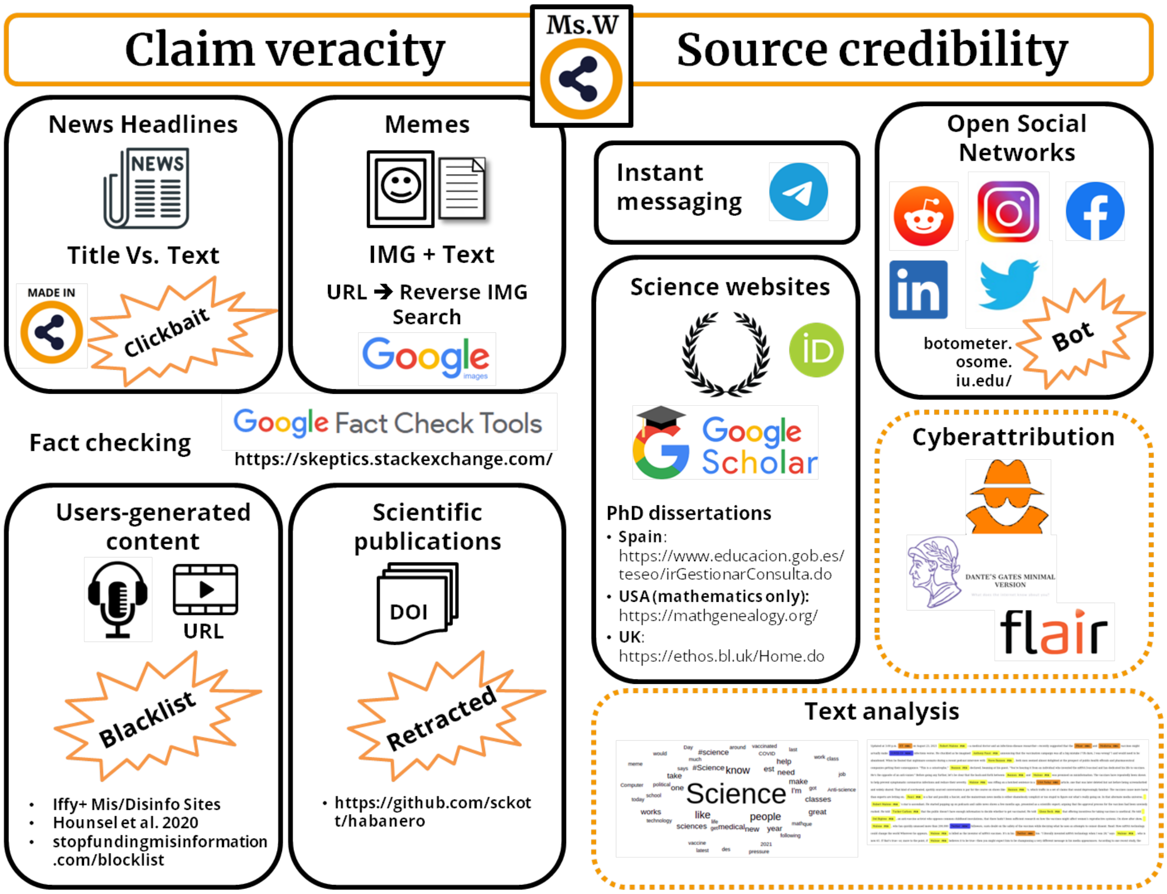


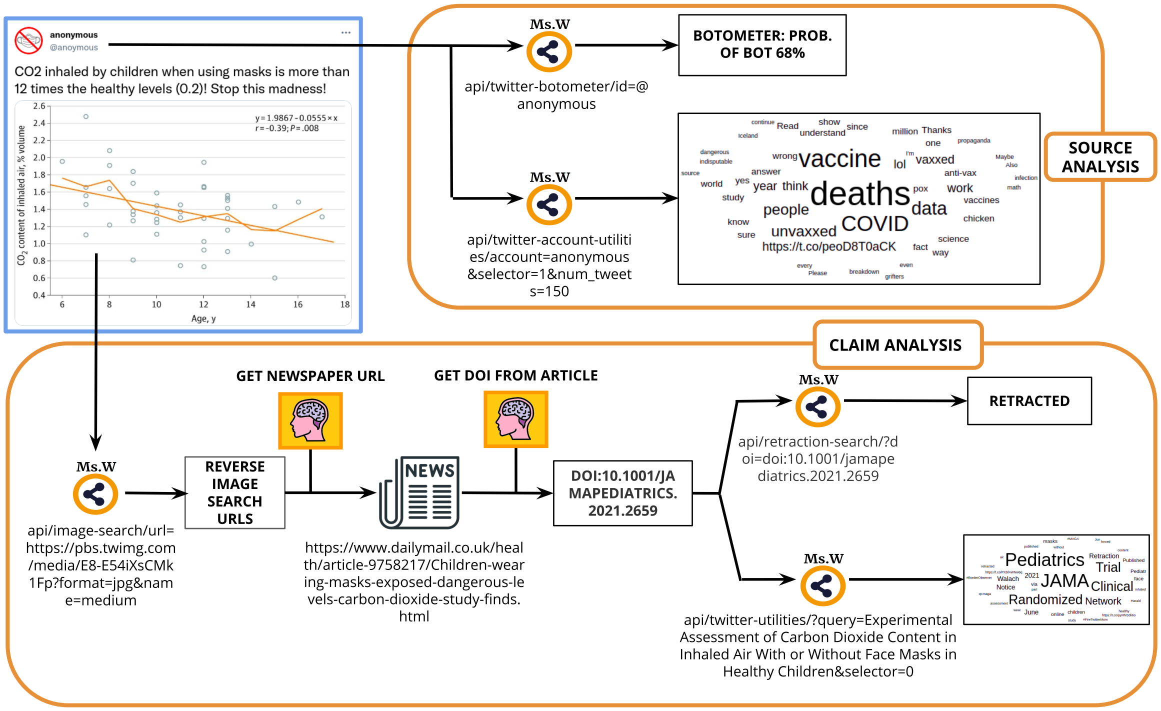
Figure 2. Ms.W Toolbox integrated functionalities

tion. The ideal tweet displayed in figure three is a privacy-preserving artefact adopted to avoid exposing specific accounts and the messages they endorse. Aggregated results, obtained from the analysis of real Twitter accounts, are included as wordclouds.

Following Ms.W methodology, we start from assessing the credibility of the source by evaluating the risk that the account is a bot. The advanced user calls the botometer endpoint and obtains as a result a 68% chance of the Twitter account being a bot (taking into account that 28% of accounts with a bot score above 1.1 are labeled as humans by the Botometer).

To better understand the worldview of the source, the user then obtains from Ms.W REST API a wordcloud of the last 150 tweets the account published (in fig. 3 Twitter handles are removed to protect users' identities). Words appearing in the cloud show that the account has been vocal about COVID19 vaccine. Focusing now on the claim made in the post, the advanced user decides to run, on the endpoint, a reverse image search. From all returned URLs that may include the image, the user selects a URL from a newspaper article where she finds the scientific article, which is the true origin of the image.

As the news article provides the DOI of the scientific article, the user calls the retraction endpoint to confirm whether or not the scientific article was accepted or not (thus retracted) by the scientific community. Afterwards, the user requests from the endpoint a wordcloud of the tweets that contain the title of the scientific article in order to obtain aggregated information about it. Besides content specific expressions such as ”clinical trial”, the word ”retracted” is also visible in the results with the name of the scientist authoring the piece. The most important part of this search is that the user is now able to identify those accounts talking about this scientific article. An analysis of this community would be the next



anonymous
@anoymous

CO2 inhaled by children when using masks is more than 12 times the healthy levels (0.2)! Stop this madness!

$y = 1.9867 - 0.0555 \times x$
$r = -0.39; P = .008$
CO_2 content of inhaled air, % volume
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
6
8
10
12
14
16
18
Age, y
Ms.W
api/twitter-botometer/id=@anonymous
BOTOMETER: PROB. OF BOT 68%
Ms.W
api/twitter-account-utilities/account=anonymous&selector=1&num_tweets=150
continue Read show understand since million Thanks one propaganda Iceland dangerous indisputable wrong vaccine lol I'm vaxxed Maybe Also source answer anti-vax infection world yes year think deaths pox work math study vaccines people data know COVID chicken sure unvaxxed https://t.co/peoD8T0aCK fact science way every Please breakdown even grifters
SOURCE ANALYSIS
CLAIM ANALYSIS
GET NEWSPAPER URL
GET DOI FROM ARTICLE
Ms.W
api/image-search/url=https://pbs.twimg.com/media/E8-E54iXsCMk1Fp?format=jpg&name=medium
REVERSE IMAGE SEARCH URLS
NEWS
https://www.dailymail.co.uk/health/article-9758217/Children-wearing-masks-exposed-dangerous-levels-carbon-dioxide-study-finds.html
DOI:10.1001/JAMAPEDIATRICS.2021.2659
Ms.W
api/retraction-search/?doi=doi:10.1001/jamapediatrics.2021.2659
RETRACTED
Ms.W
api/twitter-utilities/?query=Experimental Assessment of Carbon Dioxide Content in Inhaled Air With or Without Face Masks in Healthy Children&selector=0
masks Pediatrics Retraction Published Trial Walach 2021 Notice JAMA Clinical Randomized Network June online children

Figure 3. Ms.W use case: back and forward from a tweet to a retracted scientific article

stage here (not represented in fig. three for privacy reasons and lack of space). The user could label accounts according to the worldviews expressed in tweets associated with the scientific article. A group of COVID19 denialists could be isolated from the overall community discussing the findings and their activities could be further investigated. The connection between the scientific article and the presence of clickbait in associated newspaper articles could also be assessed. All these pieces together would contribute to advance users' fact-checking and investigative work.

6. Discussion and conclusions

The weaponisation of scientific information as part of broader cloaked science and disinformation campaigns is an especially dangerous hybrid threat. Debunking health misinformation is a priority as a worrisome proportion of fabricated posts online is crafted as part of large disinformation campaigns and information operations conceived to deceive the public opinion and shape their views in favour of some party or against a specific opponent.

In 2016, the European Commission adopted a Joint Framework to foster the resilience to countering hybrid threats in cooperation with NATO, while in 2018, it issued a Communication titled "Tackling online disinformation: a European approach", followed in 2021 by a Code of Practice on Disinformation, which has been signed by digital platforms (Facebook, Google, Mozilla, Twitter, Microsoft, and TikTok) and trade associations. Even though since 2014 there has been an exponential growth in the number of active fact-checkers, there are still no specific applications helping users to tackle the problem of scientific misinformation and cloaked science. We argue that the problem of debunking disinformation can only be partially automated. To reduce risks of manipulation of automated fact-checking services, a human-in-the-loop approach needs to be followed in designing information debunking tools. Thus, in this article we have presented the design of the TRESCA Misinformation Widget, or Ms.W, which comprises a methodology and a toolbox for helping users verify the veracity of scientific claims and the credibility of the sources making those claims.

Ms.W comprises two elements: (a) a methodology with instructions users should follow for forming opinions before expressing judgements on the truthfulness of information; (b) a toolbox with automatic solutions for verifying the accuracy of claims and the credibility of sources. Ms.W toolbox consists of a REST API and of a web interface.

Even though it is often challenging to clearly attribute an information operation to a specific attacker, the investigation of attackers' intentions and identities is a necessary condition to disentangle mis- from dis-information. Being able to make this distinction has relevant operational, judicial and public policy implications. Fact-checkers focus their efforts on debunking misinformation by focusing on content accuracy more than on cyberattribution. Cybersecurity researchers and secret services tend to concentrate their efforts on identifying suspicious accounts, which can be bots and trolls, or State-sponsored accounts. *Unclear attribution* refers to cases where there is insufficient evidence to definitively identify campaign operators or participants.

Overall, Ms.W is meant to help journalists, policy-makers and lay people not only investigate the accuracy of claims they find online (posts, memes, videos, newspaper articles, etc.), but also reveal sources of orchestrated disinformation campaigns leveraging scientific information.

Ms.W can benefit a variety of stakeholders and can be adapted to respond to their specific needs. Ms.W as a toolbox and as a methodology can be used in teaching courses to increase digital and media literacy. It can also be used to promote investigative journalism, following the example of bellingcat.com or by private corporations who want to assess the cross-platform penetration of marketing campaigns run by their competitors. It can of course be used by law enforcement agencies in the investigation of disinformation campaigns.

By focusing on the interdependence between the veracity of claims and the credibility of sources, Ms.W approach demonstrates to be suitable for the investigation of hybrid threats, which requires the constant integration of different streams of information flowing online and offline. Furthermore, Ms.W is an expression of human-centered security as the automation of tasks never replaces human intelligence, but complements it.

In this respect, Ms.W helps users (a) automate tasks that demands the processing of huge amounts of information (as in the case of bot detection or community analysis), and (b) visualise and summarise text data. By promoting users' critical thinking, Ms.W is also a promising tool for improving the quality of data labelling. By constantly moving from small to big data in their investigative efforts, users can triangulate findings, revise their assumptions and reach robust conclusions. Another advantage of Ms.W as a REST API for advanced users is its adaptability and the possibility of developing and integrating more investigative tools to the Ms.W toolbox. Ms.W methodology help also users become more aware of ideological and psychological biases, without leaving the quest for truth only in the hands of machines.

7. Funding

This work was funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 872855 (TRESCA) and partially supported by the project ”CYNAMON – Cybersecurity, Network Analysis and Monitoring for the Next Generation Internet”, funded by the Madrid Region under “Programas de Actividades de I+D entre grupos de investigación de la Comunidad de Madrid en tecnologías 2018” (P2018/TCS-4566; BOCM. No. 304; 21/12/2018).

REFERENCES

1. Islam MS, Sarkar T, Khan SH, Mostofa Kamal AH, Hasan SMM, Kabir A, et al. COVID-19-Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. The American journal of tropical medicine and hygiene. 2020;103(4):1621–1629. doi:10.4269/ajtmh.20-0812.
2. Ziems C, He B, Soni S, Kumar S. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv preprint arXiv:200512423. 2020;.
3. Rasmussen AL. On the origins of SARS-CoV-2. Nature Medicine. 2021;27(1):9–9. doi:10.1038/s41591-020-01205-5.
4. Donovan J, Nilsen J. Cloaked Science: The Yan Reports. Media Manipulation Casebook. 2021;.
5. Schwartz C, Overdorf R. Disinformation from the Inside: Combining Machine Learning and Journalism to Investigate Sockpuppet Campaigns. In: Companion

Proceedings of the Web Conference 2020;. p. 623–628. Available from: https://cybersafety-workshop.github.io/2020/papers/disinformation.pdf.

6. Jonason PK, Webster GD. The dirty dozen: A concise measure of the dark triad. Psychological assessment. 2010;22(2):420.
7. Tsikerdekis M, Zeadally S. Detecting Online Content Deception. IT Professional. 2020;22(2):35–44.
8. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 7534–7550. Available from: https://aclanthology.org/2020.emnlp-main.609.
9. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: Detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media. vol. 11; 2017.
10. Palacio I, Arroyo D. Fake News Detection: Do Complex Problems Need Complex Solutions? In: 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020). CISIS; 2019.
11. Oliva C, Palacio Marín I, Lago-Fernández LF, David A. Fake news detection: When complex problems demand complex solutions;.
12. Hounsel A, Holland J, Kaiser B, Borgolte K, Feamster N, Mayer J. Identifying disinformation websites using infrastructure features. In: 10th {USENIX} Workshop on Free and Open Communications on the Internet ({FOCI} 20); 2020.
13. Literat I, Kligler-Vilenchik N. Youth collective political expression on social media: The role of affordances and memetic dimensions for voicing political views. New Media & Society. 2019;21(9):1988–2009.
14. Wilczek B. Misinformation and herd behavior in media markets: A cross-national investigation of how tabloids' attention to misinformation drives broadsheets' attention to misinformation in political and business journalism. Plos one. 2020;15(11):e0241389.
15. Grimes DR. Medical disinformation and the unviable nature of COVID-19 conspiracy theories. Plos one. 2021;16(3):e0245900.
16. Kolluri NL, Murthy D. CoVerifi: A COVID-19 news verification system. Online Social Networks and Media. 2021;22:100123. doi:https://doi.org/10.1016/j.osnem.2021.100123.
17. Guarino S, Trino N, Chessa A, Riotta G. Beyond fact-checking: Network analysis tools for monitoring disinformation in social media. In: International Conference on Complex Networks and Their Applications. Springer; 2019. p. 436–447.
18. Hassan N, Arslan F, Li C, Tremayne M. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017. p. 1803–1812.
19. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or fiction: Verifying scientific claims. arXiv preprint arXiv:200414974. 2020;.
20. Fraga-Lamas P, Fernandez-Carames TM. Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. IT Professional. 2020;22(02):53–59. doi:10.1109/MITP.2020.2977589.

David Arroyo is Tenured Scientist in the *Institute of Physical and Information Technologies* (ITEFI) of CSIC, Spain. Email: david.arroyo@csic.es.

Alberto Gómez-Espés is Research Assistant at ITEFI-CSIC. Email: alberto.g.e@csic.es.

Santiago Palmero-Muñoz is Research Assistant at ITEFI-CSIC. Email: santiago.palmero@csic.es.

Sara Degli-Esposti is the corresponding author and Research Scientist at IFS-CSIC. Email: sara.degli.esposti@csic.es.