FAKE NEWS THREATENS LIBERAL DEMOCRACY: WILL ARTIFICIAL INTELLIGENCE HELP PREVENT GASLIGHTING PSYCHOTIC ENEMIES?

MING MING CHIU
The Education University of Hong Kong

JEONG-NAM KIM
University of Oklahoma

HYELIM LEE
University of Oklahoma

YU WON OH
Myongii University

CHONG-HYUN PARK
Sungkyunkwan University

Volume 5 · Number 2 · 2021 · Pages 363 - 387

Effective liberal democracy relies on citizens and their government representatives accepting shared facts to solve problems and make decisions. By contrast, "false or misleading content presented as news" (*fake news*)[1] threatens a healthy marketplace of ideas.[2] Persuasive fake news divides people to perceive different realities, view each other as psychotic, and hear advocacy for opposing realities as enemy gaslighting. Failing to agree on facts, shared problems, or acceptable solutions paralyzes democratic processes,[3] and denial of actual problems (e.g., climate change) blocks the cooperation and compromise

---

[1] Nolan Higdon. *The anatomy of fake news*. University of California Press, 2020: 27. For the purposes of this article, consider reports of events, such as rain, as true or fake news. Epistemological explication of subtler distinctions of truth, validity, or contested concepts are beyond the scope of this short article.

[2] Jeong-Nam Kim and Homero Gil de Zúñiga, "Pseudo-information, media, publics, and the failing marketplace of ideas: Theory," *American Behavioral Scientist* 65, no. 2 (2021): 164, https://doi.org/10.1177/0002764220950606.

[3] Civiqs, "Do you think the nation's economy is getting better or worse?" Civiqs, accessed July 7, 2021, https://civiqs.com/results/economy_us_direction; Michael Tesler, "Republicans' pessimistic views on the economy have little to do with the economy," *FiveThirtyEight*, May 5, 2021, https://fivethirtyeight.com/features/republicans-pessimistic-views-on-the-economy-have-little-to-do-with-the-economy/.

needed to solve them.[4] Moreover, solutions to false problems (e.g., voter fraud)[5] not only waste time, effort, and resources, but might cause harm (allowing identification cards of soldiers but not college students[6] can disenfranchise students and change election outcomes). Even worse, viewing fellow citizens as enemies fosters self-destructive behaviors that hurt them but also harm ourselves.[7] Hence, fake news threatens liberal democracy.

News Validity, Volume, and Velocity

Detecting fake news to inform suitable interventions is challenging. Despite 1,480 public tweets discussing a violent US Capitol Insurrection against election fraud (spurred by fake news), government authorities failed to stop it from

---

[4] Kari Marie Norgaard, "Climate denial: Emotion, psychology, culture, and political economy," *Oxford handbook on climate change and society* 18, (Summer, 2011): 399.
[5] Minnite, Lorraine C. *The myth of voter fraud*. Cornell University Press, 2011:4.
[6] "Required identification for voting in person," VoteTexas, accessed 2021, https://www.votetexas.gov/register-to-vote/need-id.html.
[7] Dominic J. Packer and Jay Van Bavel. *The power of us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony*. 2021:153.

killing 5 people and injuring over 140 others.[8] Even with training, most humans cannot identify fake news,[9] especially as alternative media (e.g., *209 Times*) can publish 99% real news (e.g., Associated Press news) mixed with 1% fake news—which itself mostly contains facts.[10] Political dominance of media of mostly true news (e.g., Viktor Orban's Fidesz party in Hungary)[11] further facilitates embedding of fake news and hinders its detection.

---

[8] Jessica Guynn, "'Burn down DC': Violence that erupted at Capitol was incited by pro-Trump mob on social media," *USA Today*, February 12, 2021, https://www.usatoday.com/story/tech/2021/01/06/trump-riot-twitter-parler-proud-boys-boogaloos-antifa-qanon/6570794002/

[9] Lutzke, L., Drummond, C., Slovic, P., and Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, *58*, 101964; Pimmer, Christoph, Christoph Eisemann, and Magdalena Mateescu. "Fake news resilience through online games? Tentative findings from a randomized controlled trial in higher education." Paper presented at the 17th International Conference on Cognition and Exploratory Learning in Digital Age, 2020.

[10] Yowei Shaw, "The chaos machine: An endless hole," *Hear Every Voice*, April 29, 2021, https://www.npr.org/programs/invisibilia/992214107/the-chaos-machine-an-endless-hole.

[11] Zoltan Simon, "Hungary's strongman leader nears full control of national media," *Bloomberg*, July 24, 2020, https://www.bloomberg.com/news/articles/2020-07-24/top-hungarian-independent-news-site-staff-quits-citing-pressure.

Social media motivates and enables people to find, curate, and quickly share messages, regardless of validity, within their social networks.[12] So, purveyors of fake news launch them via publics' information trafficking to enhance credibility, approval, and velocity (i.e., how many users see the fake news within a time interval; = users/time; e.g., 2,000 users per hour).[13] As fake news spreads faster than true news, social media's facility accelerates its spread[14] (e.g., over 500 million daily tweets). Quickly disseminating fake news to a million people within a day rather than a decade escalates its threat by overwhelming human analysts and demanding faster responses. Likewise, viral fake news across many diverse, online communities (rather than only one homogeneous community) appeals to many different audiences, accelerates its spread, and escalates its danger.

Even without fake news, unimportant viral news can distract from unflattering true news (e.g., after Russia's

---

[12] Soroush Vosoughi, Deb Roy, and Sinan Aral, (2018). "The spread of true and false news online," *Science* 359, no. 6380 (Spring, 2018): 1146.
[13] Kim and Gil de Zúñiga: 165, 167.
[14] "The number of tweets per day in 2020," David Sayce, May 2020, https://www.dsayce.com/social-media/tweets-day/.

Fancy Bear hacked John Podesta emails, Wikileaks released them to deluge Trump's "grab 'em by the p***y" scandal).[15] Indeed, mass media influence audiences mostly by directing their attention, not by imposing specific views[16] (e.g., Trump supporters rejected mainstream media's call to wear masks to slow the pandemic). Unlike mainstream media, social media reflects many people's thoughts, so it reflects public understanding and hence, the information health of a liberal democracy.

Artificial Intelligence: Machine Learning

As humans cannot readily determine the validity of high volume, high-velocity messages, we advocate using computers to quickly collect, organize, and analyze them.[17]

---

15 "Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election," Senate Intelligence Committee, November 10, 2020, https://www.intelligence.senate.gov/sites/default/files/documents/report_volume5.pdf.

16 *Agenda setting theory*, Maxwell E. McCombs, and Donald L. Shaw, "The agenda-setting function of mass media," *Public Opinion Quarterly* 36, no. 2 (Summer, 1972): 176.

17 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Cambridge: MIT press, 2018: 2.

Notably, experts in computer systems that perform tasks typically requiring human intelligence (*artificial intelligence*), have the knowledge, big data, and money (e.g., government grants and company budgets) to identify fake news, its authors, its dissemination speed, and its extent. Specifically, they develop computer systems that can learn and adapt without explicit instructions (*machine learning*, ML) by using many messages to create filters that detect fake news.[18] Based on many input messages labeled as true, false, or neither (e.g., vanilla tastes better than chocolate), ML creates a series of filters from message features to estimate the likelihood that a new message is fake news (e.g., 23% fake, 60% true, 17% neither).[19] Complementing ML, statistical analyses of linguistic, demographic, and other

[18] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems* 117 (Spring 2021): 47.
[19] Michael I. Jordan, and Tom M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science* 349, no. 6245 (Summer 2015): 255.

features of a message can identify and predict its dissemination scope and speed.[20]

Reducing ML Biases with Diverse Views

However, ML can suffer from biases and opacity.[21] Biased ML systems systematically disadvantage underprivileged subpopulations.[22] As no system can fully satisfy all 21 mathematical definitions of fairness, mitigating bias typically involves trade-offs.[23] For example, fairness can demand that an author's race does not (a) predict fake news (*independence*) or (b) alter ML prediction of fake news (*separation*). However, if an author's race is correlated with

[20] Gabriel Rossman, Ming Ming Chiu, and Joeri M. Mol. "Modeling Diffusion of Multiple Innovations via Multilevel Diffusion Curves: Payola in Pop Music Radio." *Sociological Methodology* 38, no. 1 (2008): 222.
[21] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1.
[22] Heinrich Jiang, and Ofir Nachum, (2020). "Identifying and correcting label bias in machine learning," https://arxiv.org/abs/1901.04966.
[23] Arvind Narayanan, "21 Fairness Definitions and Their Politics, ACM Conference on Fairness, Accountability and Transparency," YouTube, March 2, 2018, https://www.youtube.com/ watch?v=jIXIuYdnyyk.

a feature linked to true or fake news (e.g., many professors are white), a ML model cannot have both race independence and race separation, so it must be somewhat unfair.

Biases can occur via sampling, history, labeling, or proxies. An input data sample might not represent the target population (*sampling bias;*[24] e.g., a social media sample with mostly tweets and few Facebook posts overvalues tweets and undervalues Facebook posts). Also, inputting historical data with biased outcomes can reproduce them (*historical bias*;[25] e.g., ignoring the recent increase in news articles authored by women both overvalues men's articles and undervalues women's articles). Incorrect labeling of true news as fake can bias ML results (*labeling bias*;[26] e.g., Brexit supporters incorrectly label the message "European

---

[24] Joy Buolamwini, and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77.
[25] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California University Law Review* 104, no. 3 (Autumn, 2016): 671, http://www.californialawreview. org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.
[26] Cade Metz, "Using A.I. to Find Bias in A.I.," *New York Times,* June 30, 2021, https://www.nytimes.com/2021/06/30/technology/artificial-intelligence-bias.html

membership increases British trade" as false). As features of the outcome (true vs. fake) or input (credibility) often lack precise definition, ML designers might use simplistic/imprecise proxies (*proxy bias*;[27] e.g., BBC news is not always true). Hence, each ML team should include diverse individuals with sensitivity to such biases (e.g., women, racial minorities, etc.). Then, all team members can attend to these biases, brainstorm proposals to minimize them, and test their effectiveness.

Also, unlike mathematical or statistical models whose variables predict outcomes (e.g., age predicts fake news), ML's filters typically yield cryptic numbers (black box). Thus, ML designers often do not know (a) how to interpret the numbers and (b) how their system decides[28] whether a news article is fake. As such opacity exacerbates invisible

[27] Aaron Klein, "Credit Denial in the Age of AI," *Brookings Institution*, April 11, 2019, https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/; Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (Winter, 2019): 447.
[28] Mohri et al.:5.

biases, ML scholars devised methods to increase ML transparency.[29]

Overcoming ML Opacity and Bias with Greater Transparency

We can increase ML transparency via (a) global descriptions of features and (b) local interpretation of specific predictions.[30] By organizing and visualizing the overall ML results (e.g., via *Yellowbrick* software), designers can hypothesize key input features of the message, author, online community, or time, and then test how assigning different weights to each feature affects outcomes.[31] For example, ML output indicates 7 million fake news messages and 3 million true news messages, so we examine the

[29] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo, "Statistical stability indices for LIME: obtaining reliable explanations for machine learning models," *Journal of the Operational Research Society*, (Winter, 2020): 1.
[30] Josua Krause, Adam Perer, and Kenney Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," *Proceedings of the 2016 CHI conference on human factors in computing systems*, (Summer, 2016): 5686.
[31] Benjamin Bengfort, and Rebecca Bilbro, "Yellowbrick: Visualizing the scikit-learn model selection process," *Journal of Open Source Software* 4, no. 35 (Spring 2019): 1075.

proportion of messages in each category with specific features (e.g., certainty words ["always"] appear in 80% fake news but only 10% of true news). Then, we create models (e.g., statistical) with features like certainty, and test whether the model results resemble the ML results. Likewise, we test for bias. To detect gender bias, we test whether adding a *female* (vs. male) *author* variable to the ML system yields substantially different results.

Also, users can examine specific predictions to consider whether slight changes in a message's input features affect the outcome (e.g., via *Lime software*).[32] For example, does sending two similar sentences ("immigrants *always* help the French economy" and "immigrants *can* help the French economy") into the ML system yield different results? Likewise, we test for racial bias with "*Black* immigrants help the French economy" and "*White* immigrants help the French economy."

ML scholars can then work together with domain experts (e.g., communication, linguistics, political science) and statisticians to test these hypotheses within transparent,

[32] Visani et al.:1.

comprehensive, ecological, explanatory models.[33] Then, they can contextualize, interpret, and triangulate their ML and statistics results.[34] Such cooperation can yield ML systems that humans can interpret, use, and trust.

Implications

Institutional cooperation, data protection, and publicized findings can help such collaborations succeed and build public support for detecting fake news. As private companies (e.g., Google, Facebook) and government agencies (e.g., Smithsonian in US) store most social media data, their institutional agreements with researchers can both address public concerns (e.g., protect data confidentiality and privacy) and ensure public dissemination of findings for other scholars, companies, and government officials to use.

[33] Littmann et al., "Validity of machine learning in biology and medicine increased through collaborations across fields of expertise," *Nature Machine Intelligence* 2, no. 1 (Spring 2020): 18.
[34] Pratap Dangeti, *Statistics for machine learning* (UK: Packt Publishing Ltd., 2017):125.

Educators can help foster fake news expertise via media literacy and an ML pipeline. Current ML fake news detection systems are too specific or too inaccurate for general users; even the best ML fake news detection systems are biased and flawed,[35] so people need media literacy education (e.g., News Literacy Project curriculum) to reduce their susceptibility to fake news.[36] Teaching ML and its prerequisites (e.g., computer programming) to all students (starting with https://scratch.mit.edu/ for children) both prepares some students for ML careers and helps all students understand ML strengths and weaknesses to better grasp and accept suitable ML uses, such as detecting fake news.

Conclusion

As high-volume, high-velocity, persuasive fake news can threaten a healthy information marketplace of ideas and paralyze democratic processes, ML can help scan billions of

[35] Jiang and Nachum.
[36] *News Literacy Project* offers a free 13-lesson online curriculum (https://get.checkology.org/). Also, *Ground News* compares reporting (https://ground.news/), and *Adfontes Media* assesses the reliability of news sources (http://www.adfontesmedia.com/).

messages a day to identify fake news, its scope and its velocity. However, ML experts across industry, government, and academia must cooperate to create and share comprehensible ML algorithms to detect fake news with minimal bias. Also, educators can help counter fake news by teaching media literacy, ML, and its prerequisites.

References

"Do you think the nation's economy is getting better or worse?" Civiqs, accessed July 7, 2021. https://civiqs.com/results/economy_us_direction.

"Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election." Senate Intelligence Committee, November 10, 2020. https://www.intelligence.senate.gov/sites/default/files/documents/report_volume5.pdf.

"Required identification for voting in person." VoteTexas, accessed 2021. https://www.votetexas.gov/register-to-vote/need-id.html.

Arcas, Blaise Agüera y, Margaret Mitchell, and Alexander Todorov. "Physiognomy's New Clothes." *Medium*, May 6, 2017. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California University Law Review* 104, no. 3 (Autumn, 2016): 671. http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

Bengfort, Benjamin, and Rebecca Bilbro. "Yellowbrick: Visualizing the scikit-learn model selection process." *Journal of Open Source Software* 4, no. 35 (Spring 2019): 1075. https://doi.org/10.21105/joss.01075.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81, (2018): 77-91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Dangeti, Pratap. *Statistics for machine learning* (UK: Packt Publishing Ltd., 2017).

Guynn, Jessica. "'Burn down DC': Violence that erupted at Capitol was incited by pro-Trump mob on social media." *USA Today*, February 12, 2021. https://www.usatoday.com/story/tech/2021/01/06/trump-riot-twitter-parler-proud-boys-boogaloos-antifa-qanon/6570794002/.

Hakak, Saqib, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. "An ensemble machine learning approach through effective feature extraction to classify fake news." *Future Generation Computer Systems* 117 (Spring 2021): 47-58. https://doi.org/10.1016/j.future.2020.11.022.

Jiang, Heinrich and Ofir Nachum. "Identifying and correcting label bias in machine learning." *Proceedings of Machine Learning Research* 108, (2020): 702-712.

Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245

(Summer 2015): 255-260. https://doi.org/10.1126/science.aaa8415.

Kim, Jeong-Nam and Homero Gil de Zúñiga. "Pseudo-information, media, publics, and the failing marketplace of ideas: Theory." *American Behavioral Scientist* 65, no. 2 (2021): 163-179. https://doi.org/10.1177/0002764220950606.

Klein, Aaron. "Credit Denial in the Age of AI." *Brookings Institution*, April 11, 2019. https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/.

Krause, Josua, Adam Perer, and Kenney Ng. "Interacting with predictions: Visual inspection of black-box machine learning models." *Proceedings of the 2016 CHI conference on human factors in computing systems,* (Summer, 2016): 5686-5697. https://doi.org/10.1145/2858036.2858529.

Littmann, Maria, Katharina Selig, Liel Cohen-Lavi, Yotam Frank, Peter Hönigschmid, Evans Kataka, Anja Mösch, Kun Qian, Avihai Ron, Sebastian Schmid, Adam Sorbie, Liran Szlak, Ayana Dagan-Wiener, Nir Ben-Tal, Masha Y. Niv, Daniel Razansky, Björn W. Schuller, Donna Ankerst, Tomer Hertz, and Burkhard Rost. "Validity of machine learning in biology and medicine increased through collaborations across fields of expertise." *Nature Machine Intelligence* 2, no. 1 (Spring 2020): 18-24. https://doi.org/10.1038/s42256-019-0139-8.

Manning, Christopher D.. "Computational linguistics and deep learning." *Computational Linguistics* 41, no. 4 (2015): 701-707.

McCombs, Maxwell E., and Donald L. Shaw. "The agenda-setting function of mass media." *Public Opinion Quarterly* 36, no. 2 (Summer, 1972): 176-187. doi: 10.1086/267990.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1.

Metz, Cade. "Using A.I. to Find Bias in A.I." *New York Times.* June 30, 2021. https://www.nytimes.com/2021/06/30/technology/artificial-intelligence-bias.html.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Cambridge: MIT press, 2018.

Narayanan, Arvind. "21 Fairness Definitions and Their Politics, ACM Conference on Fairness, Accountability and Transparency." Youtube, March 2, 2018. https://www.youtube.com/ watch?v=jIXIuYdnyyk.

Norgaard, Kari Marie. "Climate denial: Emotion, psychology, culture, and political economy." *Oxford handbook on climate change and society* 18, (Summer, 2011): 399-413.

https://doi.org/10.1093/oxfordhb/9780199566600.003.0027.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (Winter, 2019): 447-453. https://science.sciencemag.org/ content/366/6464/447.

Packer, Dominic J., and Jay Van Bavel. *The power of us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony*. New York: Little, Brown Spark, 2021.

Sayce, David. "The number of tweets per day in 2020." May 2020. https://www.dsayce.com/social-media/tweets-day/.

Shaw, Yowei. "The chaos machine: An endless hole." *Hear Every Voice*, April 29, 2021. https://www.npr.org/programs/invisibilia/992214107/the-chaos-machine-an-endless-hole.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD Explorations Newsletter* 19, no. 1 (Summer, 2017): 22-36.

Simon, Zoltan. "Hungary's strongman leader nears full control of national media." *Bloomberg*. July 24, 2020. https://www.bloomberg.com/news/articles/2020-07-24/top-hungarian-independent-news-site-staff-quits-citing-pressure.

Tesler, Michael. "Republicans' pessimistic views on the economy have little to do with the economy." *FiveThirtyEight*, May 5, 2021. https://fivethirtyeight.com/features/republicans-pessimistic-views-on-the-economy-have-little-to-do-with-the-economy/.

Vaughn, Julie, Avital Baral, Mayukha Vadari, and William Boag. "Dataset Bias in Diagnostic AI systems: Guidelines for Dataset Collection and Usage." *ACM Conference on Health,* *Inference and Learning*, 2020. http://juliev42.github.io/files/CHIL_paper_bias.pdf.

Visani, Giorgio, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. "Statistical stability indices for LIME: obtaining reliable explanations for machine learning models." *Journal of the Operational Research Society*, (Winter, 2020): 1-11.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The spread of true and false news online." *Science* 359, no. 6380 (Spring, 2018): 1146-1151. https://doi.org/10.1126/science.aap9559.