ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

Response to the White House's Council of Advisors on Science and Technology's (PCAST) Invitation for Input for Its Working Group on Generative AI

July 18, 2023

Prepared with the expertise and guidance of these ASA entities:*
Section on Text Analysis, Section on Statistical Learning and Data Science, *and Committee on Data Science and AI*

The American Statistical Association (ASA) appreciates this opportunity to provide the President's Council of Advisors on Science and Technology (PCAST) our input for its Generative AI Working Group regarding threats posed by use of large language models to spread disinformation. The call poses five related questions:

1. In an era in which convincing images, audio, and text can be generated with ease on a massive scale, how can we ensure reliable access to verifiable, trustworthy information? How can we be certain that a particular piece of media is genuinely from the claimed source?
2. How can we best deal with the use of AI by malicious actors to manipulate the beliefs and understanding of citizens?
3. What technologies, policies, and infrastructure can be developed to detect and counter AI-generated disinformation?
4. How can we ensure that the engagement of the public with elected representatives—a cornerstone of democracy—is not drowned out by AI-generated noise?
5. How can we help everyone, including our scientific, political, industrial, and educational leaders, develop the skills needed to identify AI-generated misinformation, impersonation, and manipulation?

Before addressing the questions separately, this response treats the problem in general while respecting the page limit in the call. Responses specific to the questions are given at the end.

Note that in this response, the definition of "large language models" ("LLMs") extends the current technical definition from "a computerized language model, embodied by an artificial neural network using an enormous amount of parameters … resulting in a tokenized vocabulary with a probability distribution", while separate from techniques using neural networks with generative pre-trained transformers (GPT), includes any Deep Learning model that uses a Generative Pre-Trained Transformer (GPT) framework that results in a tokenized vocabulary with a probability distribution and is defined by statistical methods.

Ideas That May Help

There is no technology on the horizon that will stop malicious actors from telling lies, nor some people from believing those lies. Nonetheless, there are ways to mitigate the damage, and many of them involve statistics and data science. Some of the proposed approaches apply to fake news in general, not just that produced by AI. It is worth noting that "fake news" has been generated manually by human beings for many years. In this response, we are also examining two new scenarios: human beings who now use AI to create and distribute fake news, as well as fake news purely generated by AI without human intervention.

One approach to mitigating the damage caused by fake news is the creation of an integrity score for any digital artifact, such as an article, broadcast, or image. An illustrative example of an integrity score might be where articles in the New York Times or Wall Street Journal have a probability of 0.95 of being true, articles in The Times (in London) have a probability of 0.98 of being true, and articles in Politifact have a probability of 0.93 of being true (these probabilities could be created using a Bayesian method or a frequentist (quantitative) historical accuracy method). When someone wanted to quantify the accuracy of a news article, representing the extent to which the article agreed with coverage in selected benchmarks, the integrity score would be one of multiple signals that would indicate trustworthiness. Other signals might include be a score based on an algorithm similar to an Erdős number, that is, a collaboration network that may the "collaborative distance" from the source of the artifact to a large and broad number of peers that have historically high integrity scores (Newman, 2001).

One possible criticism of the integrity score approach is the creation of an "echo chamber", that is, where the integrity score reinforces existing views and alternative ideas have a lower probability. The statistical techniques discussed above can control for this issue, using detection of originality, for example, that would mitigate an echo chamber effect (Campbell, 2020). Less directly, it seems clear that there has been a recent breakdown of comity and trust in political and other spheres. Statistical research on the causes and exacerbating factors is ongoing (e.g., Bail et al., 2018) as well as the impact of using an integrity score within such an environment, but the phenomenon demands more study and interdisciplinary research.

A second approach is taking advantage of a digital record trail that cannot be counterfeited. It would not be needed for all disseminated information, but if a consequential claim is being made, it should have a verified source. By "consequential", we mean one that has direct and significant impact on society and has a measurable level of specificity in its assertions. By "verified", we mean that the original source can be traced and identified with a high degree of confidence. Blockchain networks are the obvious tool for tracing content through the Internet (Xiao et al., 2020). A related approach is a digital signature whose authenticity is ensured by a hash code (Kuznetsov et al., 2018).

Outlier detection is much studied in statistics (Ben-Gal, 2005). If a news item is flagged as an outlier, its accuracy may be questionable. Techniques have already been developed that apply to digital text (Kannan et al., 2017) and to images (Marchette and Solka, 2003), although some additional work would surely be needed to adapt those methods of identifying outliers to false information sourced or disseminated by generative AI applications, and disinformation more generally.

Cluster analysis, that is, statistically grouping similar items together, would also be helpful (Jaeger and Banks, 2022). Disinformation is usually tailored to further a specific agenda. Automatic identification of groups of media posts that share a common theme enables the public to recognize coordinated efforts at deception. Some clusters will correspond to accurate news, but others will correspond to fake news. Cui and Potok (2005) developed methods for clustering documents, and Verma, Verma and Tiwari (2021) explored methods for clustering images. Little work has been done on clustering videos (Asano et al., 2020). Again, research would need to be done to tailor such methods to disinformation detection.

Adversarial risk analysis (ARA) is a research area that may be relevant to countering the spread of disinformation. ARA enables one to build a model for the decision-making of a strategic opponent, then place a subjective (Bayesian) distribution over the unknowns, and using this information, choose the action that maximizes expected utility (Rios et al., 2009). Using ARA to identify disinformation would require some knowledge of the goals of the opponent (e.g., to manipulate an election) and a subjective distribution over the opponent's capabilities (priors). Sensitivity to the assumed subjective distribution is easily explored in this setting.

There is ongoing discussion of the use of AI to recognize deepfakes created by other AIs (Salazar, 2020). Deepfakes are digitally manipulated media where the manipulation cannot be identified by human beings without technological assistance. Statisticians and computer scientists have developed methodologies that are useful for assessing the performance of such systems. To improve the classification power, a wide array of approaches exist. Statistically-based machine learning techniques such as boosting, stacking, and ensemble methods are all strategies for improving the accuracy of classifiers that distinguish deepfakes (Hastie, Tibshirani and Friedman, 2009).

Likewise, there is ongoing research of the use of AI to recognize AI-produced content, whether or not the content is considered a deep fake (Liang and Tadesse, 2022). The use of recognizing AI-produced content would allow for a standard (not a regulation) of "tagging" the content as AI-produced, signaling to the consumer additional information about the potential accuracy of the content. In addition, the tag could be considered as statistical input to the generation of an integrity score, even if the tag is not shown to the end consumer.

Approaches That Probably Will Not Work

There have been some discussions of the use of regulation requiring enforcing copyrights and placing watermarks on AI generated content, laws to prohibit deliberate introduction of AI material into public discourse, and the formation of agreements to slow the pace of AI development. We believe these are at best temporary patches. Generative AI research is an international enterprise: foreign actors will not be impeded by such measures, and domestic disruptors will find loopholes and evasions. Statistical methods such as the ones discussed above are more durable, transcend language barriers and cultures, and may evolve to keep pace of AI development.

Finally, we accept many false convictions as part of our everyday life, such as an over-emphasis on the likelihood of the statistical improbability of rare events such as airline crashes or winning the lottery (reference Thayer, Kahneman, etc.). We may want to accept or measure the level of risk generated by innocuous, inconsequential disinformation and solve instead for disinformation that disrupts individual or collective lives.

Responses to Specific Questions

1. How can we ensure reliable access to verifiable, trustworthy information? How can we be certain that a particular piece of media is genuinely from the claimed source?

To ensure access to trustworthy information, one can assign integrity scores to news outlets, or to news anchors, or to politicians. It would reward careful digital content and warn of purveyors of disinformation.

To ensure that a particular piece of media is from the claimed source, one needs a return address that cannot be counterfeited and will work in a network of transactions. Blockchain and other systems could work.

2. How can we best deal with the use of AI by malicious actors to manipulate the beliefs and understanding of citizens?

We can provide new tools, often statistical, that make it easy for citizens to assess the accuracy of a statement. Politifact offers one such method today, rooted in fact-checking using humans to perform some of the automated techniques discussed above. If properly and transparently constructed, the public may very well buy into such a tool. Other techniques, such as semantic search used with generative AI, might provide a very powerful way to implement such a tool.

For example, building a system that could query for semantically similar articles/paragraphs and then generating a referenced summary of supporting/refuting resources.

3. What technologies, policies, and infrastructure can be developed to detect and counter AI-generated disinformation?

Machine learning methodology can be used to detect AI-generated misinformation, and perhaps human generated misinformation. But it is an arms race and the classification will not be perfect. In addition, we can build tools to enhance rather than replace a human’s ability to investigate validity of claims directly - AI Augmented Human Judgement - using building blocks such as semantic search and generative AI to provide immediate access to authoritative sources to support/refute claims, making it easier to judge real from fake news.

Policies that impact international actors or clever and resourced domestic ones will be difficult to regulate. Attempting to slow the pace of research gives an advantage to potential opponents who will ignore any roadblocks. Policy development in a technologically advanced and rapidly developing environment will be an ongoing challenge.

4. How can we ensure that the engagement of the public with elected representatives—a cornerstone of democracy—is not drowned out by AI-generated noise?

The volume of AI generated noise seems less of a problem, considering the need for curation of trusted information. Such curation entails a combination of assessing accuracy and assessing significance---a minor error of fact and malicious disinformation are both wrong, but the latter is more consequential. Elected representatives should lead the way in establishing a system for quantifying the trustworthiness of media reports, but they will need the support of statisticians, sociologists, computer scientists, and others.

5. How can we help everyone develop the skills needed to identify AI-generated misinformation, impersonation, and manipulation?

Ironically, flooding social media, news or the public domains with disinformation may drive people to believe with greater caution what they are told. We have learned not to answer emails from “catfish”, nor to share passwords online. However, to increase skills in discriminating truth from clever AI fakes, we must create easy-to-use mechanisms that allow fact checking.

Questions can be sent to ASA Director of Science Policy, Steve Pierson: pierson@amstat.org.

References

Asano, Y., Patrick, M., Rupprecht, C., & Vedaldi, A. (2020). Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, *33*, 4660-4671.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115, no. 37 (2018): 9216-9221.

Ben-Gal, I. (2005). Outlier detection. *Data mining and knowledge discovery handbook*, 131-146.

Berkhout, J. (2016, May). Google's PageRank algorithm for ranking nodes in general networks. In *2016 13th International Workshop on Discrete Event Systems (WODES)* (pp. 153-158). IEEE.

Colin Campbell, Kirk Plangger, Sean Sands & Jan Kietzmann (2022) Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising, Journal of Advertising, 51:1, 22-38.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65.

Cui, X., & Potok, T. E. (2005). Document clustering analysis based on hybrid PSO+ K-means algorithm. *Journal of Computer Sciences (special issue)*, *27*, 33.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Jaeger, A., & Banks, D. (2022). Cluster analysis: A modern statistical review. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1597.

Kannan, R., Woo, H., Aggarwal, C. C., & Park, H. (2017, June). Outlier detection for text data. In *Proceedings of the 2017 siam international conference on data mining* (pp. 489-497). Society for Industrial and Applied Mathematics.

Kuznetsov, A., Pushkar'ov, A., Kiyan, N., & Kuznetsova, T. (2018, May). Code-based electronic digital signature. In *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 331-336). IEEE.

Liang, W., Tadesse, G.A., Ho, D. et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 4, 669–677 (2022).

Marchette, D. J., & Solka, J. L. (2003). Using data images for outlier detection. *Computational Statistics & Data Analysis*, *43*(4), 541-552.

Newman ME. The structure of scientific collaboration networks. Proc Natl Acad Sci U S A. 2001 Jan 16;98(2):404-9. doi: 10.1073/pnas.98.2.404. Epub 2001 Jan 9. PMID: 11149952; PMCID: PMC14598.

Rios Insua, D., Rios, J., & Banks, D. (2009). Adversarial risk analysis. *Journal of the American Statistical Association*, *104*(486), 841-854.

Salazar, A. P. (2020). AI tools on fake news detection: An overview and comparative study. *Researchgate*.

Verma, H., Verma, D., & Tiwari, P. K. (2021). A population-based hybrid FCM-PSO algorithm for clustering analysis and segmentation of brain image. *Expert systems with applications*, *167*, 114121.

Xiao, Y., Zhang, N., Lou, W., & Hou, Y. T. (2020). A survey of distributed consensus protocols for blockchain networks. *IEEE Communications Surveys & Tutorials*, *22*(2), 1432-1465.