Advances in Artificial Intelligence and Machine Learning; Research 1 (3) 191-202 Received 15-11-21; Accepted 15-12-21; Published 23-12-21

Machine Learning Language Models: Achilles Heel for Social Media Platforms and a Possible Solution

Richard F. Sear, Rhys Leahy
The Dynamic Online Networks Lab, George Washington University, Washington D.C. 20052 USA.
Nicholas J. Restrepo
The Dynamic Online Networks Lab, George Washington University, Washington D.C. 20052 USA.
ClustrX LLC, Washington D.C. 20007 USA.
Yonatan Lupu
The Dynamic Online Networks Lab, George Washington University, Washington D.C. 20052 USA.
Department of Political Science, George Washington University, Washington D.C. 20052 USA.
Neil F. Johnson NEILJOHNSON@GWU.EDU
The Dynamic Online Networks Lab, George Washington University, Washington D.C. 20052 USA.
Department of Physics, George Washington University, Washington D.C. 20052 USA.

Corresponding Author: Neil F. Johnson.

Abstract

Any uptick in new misinformation that casts doubt on COVID-19 mitigation strategies, such as vaccine boosters and masks, could reverse society's recovery from the pandemic both nationally and globally. This study demonstrates how machine learning language models can automatically generate new COVID-19 and vaccine misinformation that appears fresh and realistic (i.e. human-generated) even to subject matter experts. The study uses the latest version of the GPT model that is public and freely available, GPT-2, and inputs publicly available text collected from social media communities that are known for their high levels of health misinformation. The same team of subject matter experts that classified the original social media data used as input, are then asked to categorize the GPT-2 output without knowing about its automated origin. None of them successfully identified all the synthetic text strings as being a product of the machine model. This presents a clear warning for social media platforms: an unlimited volume of fresh and seemingly human-produced misinformation can be created perpetually on social media using current, off-the-shelf machine learning algorithms that run continually. We then offer a solution: a statistical approach that detects differences in the dynamics of this output as compared to typical human behavior.

Keywords: Language models, Social media, Misinformation.

Citation: Richard F. Sear, et al. Machine Learning Language Models: Achilles Heel for Social Media Platforms and a Possible Solution. Advances in Artificial Intelligence and Machine Learning. 2021;1(3):12.

1. INTRODUCTION

Online misinformation, hate, and other harmful content on social media have reached the status of a societal crisis with tangible, often fatal, real-world consequences in which people can end up harming themselves or each other (e.g. shunning vaccines) [1-3]. Even Facebook, the richest and largest social media platform, struggles to prevent misinformation from moving into its own mainstream. Recently, it has made significant efforts to curtail this problem, but nonetheless faces an uphill battle given the huge volume of material and users to moderate every day.

The problem of online misinformation in public health has reached a particularly acute level, with narratives circulating about COVID-19 and its unofficial "cures," incorrect claims against the vaccines, and misinformation about the disease's origins [4-10]. Starbird et al.'s continually updated literature review spells out the societal dangers of such online information, misinformation, and disinformation [11-13]. Thinking beyond COVID-19, the challenge of minimizing malignant online influences such as hate, bullying, and extremism is also a crucial one to overcome. Stopping such misinformation is important for securing widespread adherence to establishment policies such as household best practices for fighting climate change or acceptance of 5G towers and infrastructure near family residences.

This misinformation crisis was amplified during the pandemic. Lockdowns understandably drove many to the online world, and in particular to the communities (i.e. pages and groups) that they belong to on Facebook, in order to seek advice and share their concerns with others. This is consistent with prior studies that have shown explicitly that parents, for example, rely on community features of platforms such as Facebook for sharing guidance on issues such as family health [14-16]. The potential impacts of adding misinformation into this system become even more dangerous when one considers the finding of Centola et al. [17], who showed experimentally and theoretically that an online community can suddenly tip to an alternate stance in a reproducible way if there is a committed minority of around 25%.

The key challenge of how to address online misinformation and harms is one of scale. There are many solutions offered within the social science literature concerning fact-checking and even inoculation to prevent online receivers of misinformation from being significantly impacted. Many of these can work if done properly, but they are hard to scale up to the millions or billions that use social media. In addition, the moderators hired by social media platforms have a practically impossible task to manually analyze online texts for misinformation given the huge volume of new material every day. One technique to help automate this process is to have software that looks for similar occurrences of known misinformation. This assumes that misinformation remains relatively static, when in reality, narratives, terminology, and misinformation evolve rapidly over time.

This issue of misinformation evolving would not be a major problem if the timescale over which this evolution happens were still slow compared to the operating time of human moderators, since they could then adapt their automated systems in real time in order to include these more gradual changes in the database of known misinformation text. But what if the misinformation evolves so quickly that human moderators cannot keep up? Going further, what if it is produced perpetually by some generative algorithm that constantly produces new material that appears human-made?

The main contribution of this paper is to show that off-the-shelf and publicly available language models (such as GPT-2) can indeed play that role as perpetual creators of health misinformation, simply by learning from existing text found in online communities that currently feature such health misinformation [18]. Specifically, our study shows that they can produce misinformation concerning COVID-19 and vaccines, that appears even to subject matter experts as new and human-made. Our team of subject matter experts that had identified the initial misinformation-rich social media data used as input to the machine model, were asked to categorize its output without knowing its origins. None of them successfully identified all the synthetic text strings as being machine-made. This presents a clear message for social media platforms and their moderation efforts: an unlimited volume of fresh misinformation can be produced perpetually on social media using current, off-the-shelf machine learning algorithms running continually.

2. GPT-2 STUDY

The acronym GPT stands for "Generative Pre-trained Transformer" which is a class of models that are autoregressive language models that produce human-like text using deep learning. Such language models are already garnering huge interest in the commercial and academic sectors [19-26]. There have already been studies showing that more advanced versions (GPT-3) can produce new pieces of literature and music. Here, we show that the same is also true in the context of health misinformation, even for a slightly older version called GPT-2. GPT-2 has the added relevance for online use of being freely available and well documented, and hence more likely to be employed by many of the existing creators of misinformation. Therefore, we consider GPT-2 exclusively in this study.

As stated in a 2021 *Nature* review article [18], such a GPT model works by "observing the statistical relationships between the words and phrases it reads but doesn't understand their meaning." Among human-like realistic texts, it can therefore produce ones that are clearly meaningless and would, in the context of sophisticated literature, alert an observer as being inauthentic and machine-generated. Indeed, most prior studies of GPT models concerned this challenge of producing text with sentences, grammar, and a flow of topics, for comparison to sophisticated, human-made literature. The significant difference in online social media is that the texts tend to be short, are not well-crafted, and can blend topics in less refined ways than more formal writing does. In particular, the online communities peddling misinformation, such as the anti-vaccination communities from which we obtain the input text to GPT-2, frequently write such blended and informal text when opposing or at least questioning best-science guidance regarding health. An example is given in FIGURE 1.

This suggests that applying GPT-2 to the system of narratives from actual social media content offers a new task in which GPT-2 might perform even better than in prior trials, hence further motivating our study. The threat of perpetually fresh online misinformation from such language models has not been examined to our knowledge. Our study's scope is limited but provides a proof of principle that GPT-2 can indeed be used successfully in this way. Though the GPT series of machine learning language models has shown remarkable ability to produce fresh and realistic text output that can pass as being generated by humans, we know of no study to date where such an input vocabulary involves text from existing online discussions surrounding public health, e.g. vaccine hesitancy and misinformation.

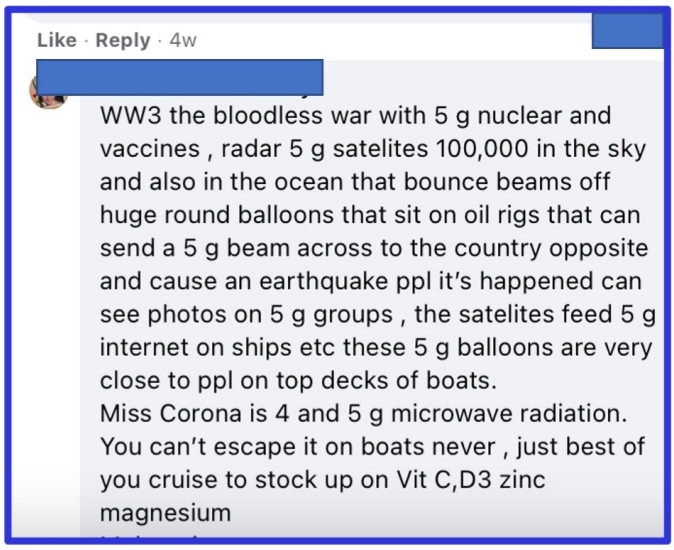

Figure 1: Example of the unpolished nature of actual text posted by a human on a Facebook page. We have purposely blocked out any identifying information.

3. CONTEXT OF PRIOR RESEARCH ON GPT MODELS

Hutson provides an excellent overview of machine language models such as the GPT series, in ref. [18]. The potential of GPT has been explored in a number of settings and application areas. In Ref. [19],Hendrycks et al. present the idea for a test to measure GPT model accuracy. It includes 57 problem areas including law, computer science and mathematics. The test examines the knowledge and problem-solving ability of the model. Their study shows that although many models perform little better than random, the largest of the latest GPT models improves on this by approximately 20 percent. They note that there are still improvements required before expert level is reached, and for GPT models to know when they are wrong. However, we note that for the present case of social media, where language is loose and informal and expertise is hard to identify, such shortcomings are not likely to be a significant problem.

Gehman et al. [20] examine the degree to which models in a similar class to GPT can be prompted to produce toxic language, before moving on to look at how this might eventually be controlled. Alarmingly, they find that even fairly innocent prompts that appear benign, can produce toxic texts when they are input to such pre-trained GPT models. This is consistent with what we report in this paper for the case of health misinformation.

Reference [21], by McGuffie et al. is also interesting in that the authors examined how GPT models could be used for weaponization by extremists to augment their ideologies and to help recruit newcomers. They used prompts inspired by actual right-wing extremist narratives to produce output. They then assessed the extent to which the GPT output could contribute to online radicalization toward violent extremism. They found that it indeed showed strength in emulating such content. They concluded that without further safeguards, weaponization is likely. They recommended action by a broad range of stakeholders to get ahead of a likely future flood of machine-

generated mis/disinformation and propaganda. Our present study suggests that this action needs to be extended to include public health stakeholders.

Brown et al. [22], consider an apparent weakness in GPT: while GPT type models have an architecture that is agnostic of task, there is still a need for significant fine-tuning according to task using datasets containing many thousands of examples. By comparison, human beings typically do reasonably well based on very few prior examples or instructions. However, they then show this limitation can be overcome by demonstrating that scaling up GPT-like models generates significant improvement in settings involving only a few examples. In alignment with the results in the present study in this paper, Brown et al. show explicitly that GPT models can create portions of synthetic news articles which subject matter experts have difficulty distinguishing from human-produced articles.

In terms of future security, Carlini et al. [23], show that if such a GPT misinformation model falls into the hands of an adversary, the adversary can perform a training data extraction attack to recover individual training examples by querying the language model. They find that larger models are more vulnerable than smaller models. This warns that GPT-2 models producing health misinformation could get hijacked and further developed by even more dangerous actors as time goes on.

For additional work in this area of GPT models, we refer to the paper of Tamkin et al. [25] who review the latest GPT model discussions which emerged in a meeting on October 14th, 2020 including researchers from OpenAI and the Stanford Institute for Human-Centered Artificial Intelligence. This includes an in-depth analysis of the technical capabilities of GPT-like models, and the potential societal effects of widespread use.

4. METHODOLOGY OF THIS STUDY

Our study in the area of health misinformation uses the latest version of GPT that is public and freely available: GPT-2. GPT-2 is a pre-trained machine learning model which works by synthesizing the next item in an arbitrary token sequence using a deep neural network. Inputting publicly available text collected from social media communities known for their misinformation, we show that GPT-2 easily produces new text that is also rich in misinformation and appears to have a human origin. Our study draws input text from Facebook, but our findings could apply to any other social media platform with typically short text-based posts.

A complete step-by-step recipe for our data collection is given in Johnson et al. [3]. Here we provide a short review. Given the reliance that people have on online communities for fostering trust and hence the potential harms of misinformation shared there, we collect our data from Facebook Pages. Our team of subject matter experts (SMEs) collecting and classifying the data communities (Facebook Pages) consisted of analysts who have several years' worth of experience with online community content on Facebook and other platforms. They are trained through prior work on the content of these online communities.

We start with a core of communities which we obtain with an automated search using keywords associated with vaccines, and then we check these manually. The SMEs work independently and afterwards check for consensus. Without requiring any discussion, consensus was already reached

at this first stage in approximately 80% of the cases. Subsequent cases were then discussed until agreement was reached. Since the number of such communities is not large (thousands rather than billions, which would be the case if we had individual users as our unit of analysis) this process can be done quite easily. We then obtain the links from these pages to other pages, to obtain an expanded set of communities and hence establish the first step in a systematic, iterative process akin to snowball sampling. We infer directed links from a source (i.e. page A) into a target (i.e. page B) when A endorses and subscribes to B. Facebook's Graph API refers to this as "fanned.", and users see it as a "like." This does not mean necessarily A likes B literally but rather is interested in knowing more about it, hence an anti-vaccination community might "like" (link to) a pro-vaccination page and vice versa. Fanning in this way creates a conduit of information from B's feed into A's wall, but most importantly it also increases the chance that followers of A get exposed to B's content. There are two reasons for this: (1) A is more likely to interact with and/or re-post B's content, and these are actions that will be broadcast to A's followers. (2) Facebook's own algorithms are more likely to expose and suggest B's content to A's followers. The additional communities obtained through these links, are then added to the initial core set. We then iterated twice following the above-mentioned steps.

The SMEs then categorize these Facebook communities as anti-vaccination or not, based on their content and "About" sections. We then extract the anti-vaccination communities' content as text, which is used as a corpus of documents to fine-tune an off-the-shelf GPT-2 model [27].Not all text in the communities identified by the SMEs is misinformation, though there are significant amounts there. We do not annotate content for misinformation (this could introduce human bias). Instead, we use a randomly selected subset of the entire corpus of text from the anti-vaccination communities to fine-tune our GPT-2 models.

The default GPT-2 model is trained for a next-word prediction task on a large corpus of English text found online. As such, the model is susceptible to biases present in this text (this has also been noted in prior research). The fine-tuning process allows the off-the-shelf model to learn relationships between words from the corpus of anti-vaccination material, which were perhaps not as prevalent or significant in the generic text corpus that it was initially trained on.

Given the variety of possible parameterizations, we employed two versions of GPT-2. The first ("gen1") was created directly by inputting the anti-vaccination community content, while the second model ("gen2") only used text from the "message" attribute (the body of the post). In practice, this means "gen2" was trained on a text corpus that was more sanitized and lacked items such as URLs which are typically found in Facebook posts.

To seed GPT-2's output, we prompted it with 15 different phrases, such as "it is time" or "we know". These phrases were either generic sentence-starters or phrases related to common topics in anti-vaccination communities, as discovered by our previous work with LDA models [28]. For each prompt, we generated text with 3 different settings for the "temperature" which is a GPT-2 user-defined parameter that controls the variety of trends that GPT- 2 looks for while producing new text. To create our text strings, we set the temperature to 1, 1.1, and 1.2. Each GPT-2 version generated 45 text strings, for a total of 90 strings of generated text.

We bundled all 90 strings together into a dataset. From this dataset, we randomly selected strings such that no string's prompt appeared more than twice. We selected 15 strings with temperature 1, 3 with temperature 1.1, and 2 with temperature 1.2. 8 texts came from "gen1" and 12 came

Table 1: Examples of GPT-2 output text from our study. It mimics the human-generated text shown in FIGURE 1 in terms of its lack of structure and unpolished writing style.

Model	Text
gen2	It isn't for the good of the state, but for a Pharma-controlled Government making it all about you instead? You cannot make the public sector go along with the lockdown." That's pretty consistent with the charade of the establishment order. Don't forget to click here to read the press release about the charges of infection. It's like they are saying the symptoms of a disease are suddenly different, in many cases causing more harm than good. "Should people refuse because of what the virus has don
gen2	What for you? Would YOU receive the flu vaccine to protect you from an influenza infection? How does CNN report viral interference (flu) and mortality? #Antibiotics are a critical part of preventing #coronavirus infection. The new edition of Raising a Vaccine Free Child is now available in digital format from Kindle, as well as in paperback from vaccinefreechild.com. Please consider donating today! "The chances of death or serious infection from all causes are very low, and people should be pr
gen1	That can be detected in several ways: • Eyes are closed and the fibers on the inside are not transparent. The fibers cannot pass the virus into the blood. • The vaccination dose is fractioned by the number of doses administered. For example, a single dose of 100 micrograms may be sufficient to reduce infection in 10 to 20 days. • The virus can be detected only in the elderly and those with underlying health conditions. However, vaccination against SARS-CoV-2 can effectively raise these condition
gen1	We know for a fact that covid-19 does not spread through feces, and it's not airborne. The only thing that will keep the infection going is what is injected into our cells with a powerful immune response. If we take steps to prevent infection and reduce infections, health and personal safety will improve. So when we come down on the side of freedom, we'll be in many different hands. us□us□us□us□us□us□ Want to stay informed? You don't have to be a doctor in order to join PIC.Sign up today! #i
gen2	Never been the only ' Silent Age' to have stopped protecting individuals from being infected with Ebola. The tale of the "cessibilite" deaths is telling in Africa. But it's from vaccines. □The beginning of this year, we need to stand united. We must win! #MichiganForVaccineChoice www.michiganvaccinechoice.org:=: No reliable data available. Where is everyone? Please get sharing like this, via FB, everywhere because: Q: "Parents are the real heroes in the fight. Doctors and nurses are the real l
gen1	The news reinforces the message that the government is hiding behind a façade, spreading fear andPopulation control instead of improving health and immunity. The fact that 99.9% of Americans are not panicking about COVID-19 and will not panic about anything new is very scary. If the goalposts are true, then other policies such as restriction and defiance of personal orders, vaccination, job protection and parental rights, can only accelerate the spread of the virus. Patients Without Fear Should
gen2	We know for a fact that the effectiveness of masks does not mean they protect us from viruses and bacteria, let us not forget that most public sector healthcare workers and nurses wear masks, either in class or in the wider community. We want to remind you that face coverings can be very harmful and very dangerous. To find out why, please read this article and the story at the following link. Thanks to Del for this info. "Should people refuse because of what we say? Should we all get on with

from "gen2". We truncated them to exactly 500 characters, which is approximately the typical post length in authentic content. Although these texts are clearly casual, low-quality writing as compared to longform works like journalism or books, they are typical of the non-academic level of language, grammar, and writing that appears in community narratives – and they all have apparent relevance, despite being entirely machine-generated. TABLE 1 shows examples of these texts.

We then carried out a trial among the SME co-authors of this paper. Their duties as paid researchers involve examining sets of texts and content. They are accustomed to assessing whether such content is likely to be written by a human. They are used to seeing automated bot-produced content that has the characteristic signatures of repeating phrases several times or an awkward machine-like structure, so we therefore gave them a subset of texts produced by GPT-2. They had no prior reason to believe these texts were any different from the ones obtained from Facebook anti-vaccination communities in their usual work. We placed these selected strings of synthetic content in a format that they were used to using in order to classify the texts.

Part of the SMEs' expertise is that they are skilled at looking at human texts and hence are well positioned – better than casual online users – to find posts which are not human. During their daily job of classifying texts, the above batch was presented to them in a sequential way. None of the SMEs successfully identified all the synthetic text strings as being produced by GPT-2. Hence, GPT-2 successfully fooled our SMEs, meaning that it would likely fool many everyday users of social media – and in particular those who only casually use it as a way to seek guidance in a period of isolation, such as during the lockdowns induced by the COVID-19 pandemic.

5. A SOLUTION TO GPT-PRODUCED MISINFORMATION

The above result of GPT-2 output appearing as fresh, human-produced (misinformation) content to SMEs, is of course concerning. GPT-3 is even more powerful and there are significant efforts underway to remove the quirks in such language models that produce occasional highly nonsensical content which might set off 'common sense' alarm bells in everyday social media users. One can therefore reasonably ask what lies in store in a decade's time: potentially a daunting threat. Are we seeing here the Achilles Heel of social media platforms, or is there a way out?

Just as complex machinery within the GPT series generates the material, there is hope that a different kind of machinery can help thwart it – the machinery of complex systems science. Specifically, it is known from complex systems studies that human activity patterns and human production patterns follow particular statistical signatures. Instead of being Poisson-like, which means that the next 'event' has a fixed probability per unit time, the reality is that there are strong positive and negative feedbacks. If you haven't been to a gym for a long time, the chance that you won't go to a gym in the next time period goes up – and vice versa. The same holds true with replying to emails, writing conventional letters, and other tasks. It even stretches to the microscopic scale of neurons firing in the brain. Given this, there is no reason that an 'event' such as the switching of topics in a text, for example, should not also follow such irregular non-Poisson patterns. This offers a way out that we will now discuss in more detail, i.e. statistical measures in association with human behavior. For full details of the implementations, we refer to earlier papers of ours mentioned below [29-32], where we implemented successfully these two schemes in biological systems and other examples of human behavior. While we have not yet implemented such measures for the texts that we produce

using GPT-2, or compared them to the real ones, there is no reason this cannot be done using these same techniques.

The first statistical measure that we propose captures the dynamics through the text in terms of topic interactions. It was proposed by Ruszczycki et al. [29] as a tool for analyzing multiple co-evolving time-series. It works as follows. As shown in FIGURE 2, we list possible topics vertically while the horizontal axis is the number n of the word as it appears sequentially in the text. As the words are counted, if Topic 2 (e.g. 'lockdown') is mentioned at the tenth word then a vertical bar appears along the horizontal axis at n=10 and along 'Topic 2'. We draw arrows into it from events in any topic that occur within a previous window Δn, and arrows from it to events in any topic that occur in the next time window Δn. Counting these arrows means that each event in each topic has a (k_"in",k_"out"). We can then plot these as shown and count the number of each type (sinks, sources, etc.). We repeat the same process for the GPT-2 text and the real text separately and see if there are any statistical differences between them.

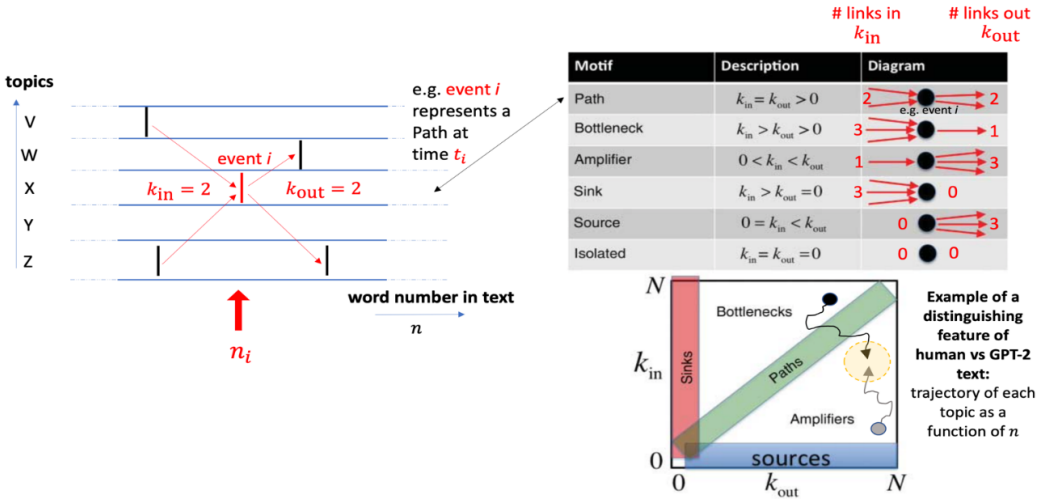

Figure 2: Example of a measure from complex systems science to be used for detecting human vs. GPT-2 text.

The second statistical measure that we propose was invented by Barabasi et al. [33]. It involves analyzing the deviations from random behavior (Poisson process) of the events (e.g. mentions) for each topic, i.e. each row in FIGURE 2 as a function of n. The idea is to compare the way in which GPT-2 text and real human text deviates from Poisson behavior in terms of the event series for each topic. We stress that neither the GPT-2 text nor the real human text are Poisson, or are assumed to be Poisson. Our focus is instead in the way that they deviate from it, and whether they do so in the same way. This statistical measure has two parts: (1) the deviation of the distribution of intervals between events for a given topic, measured by a quantity B and (2) the autocorrelation of consecutive n values, which is measured by a quantity M [33]. For a stationary Poisson process (i.e. fixed random probability of event per unit n) we should see B=0 and M=0, or close to it.

6. CONCLUSIONS AND FUTURE WORK

We have reported on a study of machine learning language models where the input vocabulary involves text from existing online discussions surrounding public health; more specifically, vaccine hesitancy and misinformation. We used the latest version of GPT that is public and freely available: GPT-2. Inputting publicly available text collected from social media communities known for their misinformation, we show that GPT-2 easily produces new text that is also rich in misinformation and appears to have a human origin. Though not meant as an exhaustive test, our team of subject matter experts that identified the initial misinformation-rich social media data, then attempted to categorize the output of our fine-tuned GPT-2 model. None of them successfully identified all output text strings as being synthesized by GPT-2. A few of these output strings were particularly convincing: our team members confidently misclassified them as being produced by humans. This shows that misinformation can be produced perpetually on social media using current, off-the-shelf machine learning algorithms running continually.

We then presented a possible solution: to train other automated tools to recognize statistical patterns in this output that would be expected of actual human behavior. Specifically, we discussed how the fluctuation between topics within a text could help flag content that may superficially appear human-like and fresh, but which has subtle statistical patterns that are too extreme to be human. This offered solution does not in any way constrain such language model development, but instead enhances it with input from the science of complex systems.

There are many possible directions for future work. First, the same exercise should be repeated for new versions of GPT as they become publicly available. Second, the process can be repeated for other areas of societal misinformation, using the same methodology, such as climate change. Third, the text can be collected from other social media platforms, such as Gab and VKontakte, and the results compared to those presented here. Fourth, the analysis performed here by subject matter experts should be extended to larger numbers. It could also be mimicked in the setting of the general public using large numbers of volunteers online. Fifth, the use of text could be extended to include memes and images. Finally, the performance of the automated tool that we propose for detecting GPT produced content, should be evaluated in real time.

7. ACKNOWLEDGMENT

We are grateful for funding for this research from the U.S. Air Force Office of Scientific Research under award numbers FA9550-20-1-0382 and FA9550-20-1-0383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

8. CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

References

[1] Johnson NF, Leahy R, Johnson Restrepo N, Velasquez N, Zheng M, et. al. Hidden Resilience, and Adaptive Dynamics of the Global Online Hate Ecology. Nature. 2019;573:261-265.

[2] www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

[3] Johnson NF, Velasquez N, Johnson Restrepo N, Leahy R, Gabriel N, et. al. The Online Competition Between Pro- And Anti-vaccination Views. Nature. 2020; 582:230-233

[4] https://www.who.int/teams/risk-communication/infodemic-management/1st-who-infodemiology-conference

[5] Starbird K, Spiro ES, Koltai K. Misinformation, Crisis, and Public Health—Reviewing the Literature V1.0. Social Science Research Council. MediaWell. 2020.

[6] Larson H. A Lack of Information Can Become Misinformation. Nature. 2020;580:306.

[7] Ball P, Maxmen A. The Epic Battle Against Coronavirus Misinformation and Conspiracy Theories. Nature. 2020.

[8] Siddiqui M, Salmon DA, Omer SB. Epidemiology of Vaccine Hesitancy in the United States. Human Vccines& Therapeutics. 2013;9:2643-2648.

[9] Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, et. al. The science of fake news. Science. 2018;359, 1094-1096.

[10] Centola D, Becker J, Brackbill D, Baronchelli A. Experimental evidence for tipping points in social convention. Science 2018;360:1116-1119.

[11] Bessi A, Coletto M, Davidescu GA, Scala A., Caldarelli G, et. al. Science vs Conspiracy: Collective Narratives in the Age of Misinformation. PloS one. 2015; 10:e0118093.

[12] https://nautilus.org/napsnet/napsnet-special-reports/of-virality-and-viruses-the-anti-vaccine-movement-and-social-media/

[13] Starbird K. Disinformation's Spread: Bots, Trolls and All of Us. Nature. 2019;571:449.

[14] Ammari T, Schoenebeck S. "Thanks for Your Interest in Our Facebook Group, but It's Only for Dads": Social Roles of Stay-At-Home Dads. In CSCW '16: Proc. 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.2016;1363–1375.

[15] Moon YY, Mathews A, Oden R, Carlin R. Mothers' Perceptions of the Internet and Social Media as Sources of Parenting and Health Information: Qualitative Study. Journal of Medical Internet Research. 2019;21:e14289.

[16] LawsR, Walsh AD, Hesketh KD, Downing KL, Kuswara K, et. al. Differences Between Mothers and Fathers of Young Children in Their Use of the Internet to Support Healthy Family Lifestyle Behaviors: Cross-Sectional StudyJournal of Medical Internet Research. 2019;21:e11454.

[17] Centola D, Becker J, Brackbill D, Baronchelli D. Experimental Evidence for Tipping Points in Social Convention. Science. 2018;360:1116–1119.

[18] Hutson M. The Language Machines. Nature. 2021;591:22-25.

[19] https://arxiv.org/abs/2009.03300

[20] https://arxiv.org/abs/2009.11462

[21] https://arxiv.org/abs/2009.06807

[22] https://arxiv.org/abs/2005.14165

[23] https://arxiv.org/abs/2012.07805

[24] Guan J, Huang F, Zhao Z, Zhu X, Huang M. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation.Transactions of the Association for Computational Linguistics. 2020;8:93–108.

[25] https://arxiv.org/abs/2102.02503

[26] Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, et al.Learning to Summarize with Human Feedback.In:LarochelleH, Ranzato M, Hadsell R, Balcan M F, Lin H, editors.In ProceedingsAdvances in Neural Information Processing Systems33. 2020;3008-3021.

[27] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, et al."Transformers: State-Of-The-Art Natural Language Processing", in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020:38–45.
£

[28] Sear RF, Velasquez N, Leahy R, Johnson Restrepo N, El Oud S, et. al. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. IEEE Access. 2020;8:91886-91893.

[29] Ruszczycki B, Zhao Z, Johnson N, Johnson NF. Temporal Network Approach to Unraveling Collective Neuron Firings. Journal of Complex Networks. 2014;2:74-84.

[30] Johnson NF, Zhao G, Caycedo F, Manrique P, Qi H, et al. Extreme alien light allows survival of terrestrial bacteria. Nature, Scientific Reports. 2013;3:2198.

[31] Zhao G, McDonald M, Fenn D, Williams S, Johnson N, et al. Transition in the Waiting-Time Distribution of Price-Change Events in a Global Socioeconomic System. Physica A. 2013;392:6458.

[32] Johnson NA, Johnson NF. An Explanation for the Universal 3.5 Power-Law Observed in Currency Markets. Results in Physics. 2016;6:48.

[33] Goh KI, Barabasi AL. Burstiness and Memory in Complex Systems. EPL.2008;81:48002.