Defending Against Neural Fake News

Rowan Zellers[♠], Ari Holtzman[♠], Hannah Rashkin[♠], Yonatan Bisk[♠]
Ali Farhadi[♠♡], Franziska Roesner[♠], Yejin Choi[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
`https://rowanzellers.com/grover`

Abstract

Recent progress in natural language generation has raised dual-use concerns. While applications like summarization and translation are positive, the underlying technology also might enable adversaries to generate *neural fake news*: targeted propaganda that closely mimics the style of real news.

Modern computer security relies on careful *threat modeling*: identifying potential threats and vulnerabilities from an adversary's point of view, and exploring potential mitigations to these threats. Likewise, developing robust defenses against neural fake news requires us first to carefully investigate and characterize the risks of these models. We thus present a model for controllable text generation called Grover. Given a headline like 'Link Found Between Vaccines and Autism,' Grover can generate the rest of the article; humans find these generations to be more trustworthy than human-written disinformation.

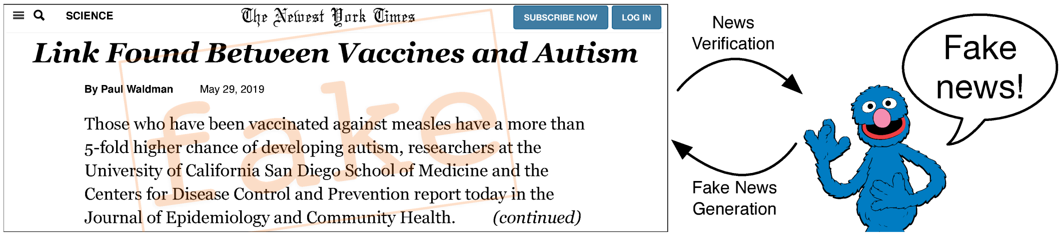
Developing robust verification techniques against generators like Grover is critical. We find that best current discriminators can classify neural fake news from real, human-written, news with 73% accuracy, assuming access to a moderate level of training data. Counterintuitively, the best defense against Grover turns out to be Grover itself, with 92% accuracy, demonstrating the importance of public release of strong generators. We investigate these results further, showing that exposure bias – and sampling strategies that alleviate its effects – both leave artifacts that similar discriminators can pick up on. We conclude by discussing ethical issues regarding the technology, and plan to release Grover publicly, helping pave the way for better detection of neural fake news.

1 Introduction

Online fake news – news designed to intentionally deceive – has recently emerged as a major societal problem. Malicious actors spread fallacious viral stories in order to gain advertising revenue, influence opinions, and even tip elections (Faris et al., 2017; Wardle and Derakhshan, 2017). As such, countering the spread of disinformation online presents an urgent technical and political issue.

To the best of our knowledge, most disinformation online today is manually written (Vargo et al., 2018). However, as progress continues in natural language generation, malicious actors will increasingly be

Preprint. Under review.



Link Found Between Vaccines and Autism

By Paul Waldman May 29, 2019

Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health. *(continued)*

Figure 1: In this paper, we explore Grover, a model which can detect *and generate* neural fake news. Humans find the articles difficult to distinguish from "real news" without high levels of scrutiny.

able to controllably generate realistic-looking propaganda at scale. Thus, while we are excited about recent progress in text generation (Józefowicz et al., 2016; Radford et al., 2018; 2019), we are also concerned with the inevitability of AI-generated 'neural' fake news.[1]

With this paper, we seek to understand and respond to neural fake news *before* it manifests at scale. We draw on the field of computer security, which relies on *threat modeling*: analyzing the space of potential threats and vulnerabilities in a system to develop robust defenses. To scientifically study the risks of neural disinformation, we present a new model called Grover.[2] Our model allows for controllable yet efficient generation of an entire news article – not just the body, but also the title, news source, publication date, and author list. This lets us study an adversary with controllable generations (e.g. Figure 1, an example anti-vaccine article written in the style of the New York Times).

Humans rate the disinformation generated by Grover as trustworthy, even more so than human-written disinformation. Thus, developing robust verification techniques against generators such as Grover is an important research area. We consider a setting in which a discriminator has access to 5000 Grover generations, but unlimited access to real news. In this setting, the best existing fake news discriminators are, themselves, deep pretrained language models (73% accuracy) (Peters et al., 2018; Radford et al., 2018; 2019; Devlin et al., 2018). However, we find that Grover, when used in a discriminative setting, performs even better at 92% accuracy. This seemingly counterintuitive finding represents an exciting opportunity for defense against neural fake news: the best models for generating neural disinformation are also the best models at detecting it.

We investigate how deep pretrained language models distinguish between real and machine-generated text. We find that key artifacts are introduced during generation as a result of exposure bias: the generator is not perfect, so randomly sampling from its distribution results in generations that fall increasingly out-of-distribution as length increases. However, sampling strategies that alleviate these effects also introduce artifacts that strong discriminators can pick up on.

We conclude with a sketch of the ethical territory that must be mapped out in order to understand our responsibilities as researchers when studying fake news, and the potential negative implications of releasing models (Hecht et al., 2018). Accordingly, we suggest a provisional policy of how such models should be released and why we believe it to be safe – and perhaps even imperative – to do so. We believe our proposed framework and accompanying models provide a concrete initial proposal for an evolving conversation about ML-based disinformation threats and how they can be countered.

2 Fake News in a Neural and Adversarial Setting

We present a framework – motivated by today's dynamics of manually created fake news – for understanding what *adversaries* will attempt with deep models, and how *verifiers* should respond.

Scope of fake news. There are many types of *false* news, ranging from satire to propaganda (Wardle, 2017). In this paper, we focus on text-only documents formatted as news articles: stories and their corresponding metadata that contain purposefully false information. Existing fake news is predominantly human-written, for two broad goals: monetization (ad revenue through clicks) and propaganda (communicating targeted information) (Bradshaw and Howard, 2017; Melford and Fagan, 2019). Achieving either goal requires the adversary to be selective about the news that they make, whether by producing only viral content, or content that advances a given agenda.

Fact checking and verification: related work. There is considerable interest in fighting online disinformation. Major platforms such as Facebook prioritize trustworthy sources and shut down accounts linked to disinformation (Mosseri, 2018; Dwoskin and Romm, 2018). Some users of these platforms avoid fake news with tools such as NewsGuard and Hoaxy (Shao et al., 2016) and websites like Snopes and PolitiFact. These services rely on manual fact-checking efforts: verifying the accuracy of claims, articles, and entire websites. Efforts to automate fake news detection generally point out stylistic biases that exist in the text (Rashkin et al., 2017; Wang, 2017; Pérez-Rosas et al., 2018). These efforts can help moderators on social media platforms shut down suspicious accounts.

[1] We thank past work, such as OpenAI's Staged Release Policy for GPT2 for drawing attention to neural disinformation, alongside other dual-use implications.

[2] Short for **G**enerating a**R**ticles by **O**nly **V**iewing m**E**tadata **R**ecords.

However, fact checking is not a panacea – cognitive biases such as the backfire effect and confirmation bias make humans liable to believe fake news that fits their worldview (Swire et al., 2017).

Framework. We cast fake news generation and detection as an adversarial game, with two players:

- **Adversary**. Their goal is to generate fake stories that match specified attributes: generally, being viral or persuasive. The stories must read realistically to both human users as well as the verifier.
- **Verifier**. Their goal is to classify news stories as real or fake. The verifier has access to unlimited real news stories, but few fake news stories from a specific adversary. This setup matches the existing landscape: when a platform blocks an account or website, their disinformative stories provide training for the verifier; but it is difficult to collect fake news from newly-created accounts.

The dual objectives of these two players suggest an escalating "arms race" between attackers and defenders. As verification systems get better, so too will adversaries. We must therefore be prepared to deal with ever- stronger adversarial attacks, which is the focus of the next section.

3 Grover: Modeling Conditional Generation of Neural Fake News

Given existing online disinformation, we have reason to believe adversaries will try to generate targeted content (e.g. clickbait and propaganda). Recently introduced large-scale generative models produce realistic-looking text (Radford et al., 2019), but they do not lend themselves to producing controllable generations (Hu et al., 2017).[3] Therefore, to probe the feasibility of realistic-looking neural fake news, we introduce Grover, which produces both realistic *and* controlled generations.

The current state-of-the-art in unconditional text generation views it as a language modeling problem (Bengio et al., 2003), in which the probability of a document $\boldsymbol{x}$ is the product of the conditional probability of generating each token x_i given previous tokens:

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i | x_1 \ldots x_{i-1}). \tag{1}$$

The document is typically treated as a single unstructured *text field*, beginning with a `<start>` token and ending with an `<end>` token. The latter, `<end>`, is particularly important because it indicates the end of the field, and when to should stop generating. However, a news article has necessary structure beyond the running text, or body field. Metadata fields include the domain where the article is published (indirectly marking the style), the date of publication, the names of the authors, and the headline of the article itself. Not only does generating a news article require producing all of these components, these fields also allow significant control over the generations (e.g. specifying a headline helps control the generated body). An article can be modeled by the joint distribution:

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body}). \tag{2}$$

However, it is not immediately obvious how to sample from Equation 2. One option is to define a *canonical order* among the article's fields $\mathcal{F}$: $(f_1 {<} f_2 {<} \ldots {<} f_{|\mathcal{F}|})$, and model the article left-to-right in that order using Equation 1: $x_1^{f_1}, x_2^{f_1}, \ldots, x_{|f_{|\mathcal{F}|}|}^{f_{|\mathcal{F}|}}$. However, this ordering would forbid sampling certain fields without prohibitively expensive marginalization. Alternatively, one could generate fields in any order, but this requires the model to learn to handle $|\mathcal{F}|!$ potential orderings during inference time.

Our solution is Grover, a new approach for efficient learning and generation of multi-field documents. We adopt the language modeling framework of Equation 1 in a way that allows for flexible decomposition of Equation 2. During inference time, we start with a set of fields $\mathcal{F}$ as context, with each field f containing field-specific start and end tokens. We sort the fields using a standard order[4] and combine the resulting tokens together. To generate a target field τ, we append the field-specific start token `<start`$-\tau$`>` to the context tokens; then, we sample from the model until we hit `<end`$-\tau$`>`.

Figure 2 shows an example of using Grover to generate an anti-vaccine article. Here, the adversary specifies a domain, date, and headline. After Grover generates the body, it can be used to generate a fake author, before finally generating a new and more appropriate headline.

[3] A common workaround is to have a human seed the text to provide context. However, this **a)** is a heavy handed technique for biasing which may not capture the desired attributes, and **b)** leaves in place a human-written beginning (as tokens are only generated left-to-right), which may create distributional artifacts.

[4] Our ordering is the following field types in order: domain, date, authors, headline, and then the body.

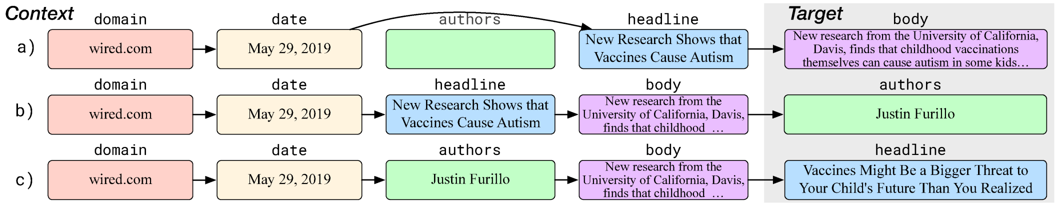

Context
Target
a)
b)
c)
domain
date
authors
headline
body
wired.com
May 29, 2019
New Research Shows that Vaccines Cause Autism
New research from the University of California, Davis, finds that childhood vaccinations themselves can cause autism in some kids…
domain
date
headline
body
authors
wired.com
May 29, 2019
New Research Shows that Vaccines Cause Autism
New research from the University of California, Davis, finds that childhood …
Justin Furillo
domain
date
authors
body
headline
wired.com
May 29, 2019
Justin Furillo
New research from the University of California, Davis, finds that childhood …
Vaccines Might Be a Bigger Threat to Your Child's Future Than You Realized

Figure 2: A diagram of three Grover examples for article generation. In row a), the body is generated from partial context (the authors field is missing). In b), the model generates the authors. In c), the model uses the new generations to regenerate the provided headline to one that is more realistic.

During training, we simulate inference by randomly partitioning an article's fields into two disjoint sets $\mathcal{F}_1$ and $\mathcal{F}_2$. We also randomly drop out individual fields with probability 10%, and drop out all but the body with probability 35%. This allows the model to learn how to perform unconditional generation. The metadata fields in each set are sorted using the standard order, and the model is trained to minimize the cross-entropy of predicting tokens in $\mathcal{F}_1$ followed by $\mathcal{F}_2$.[5]

Architecture. We draw on recent progress in training large Transformers for language modeling (Vaswani et al., 2017), building Grover using the same architecture as for GPT2 (Radford et al., 2019). We consider three model sizes. Our smallest model, Grover-Base, has 12 layers and 117 million parameters, on par with GPT and BERT-Base (Radford et al., 2018; Devlin et al., 2018). Our next model, Grover-Large, has 24 layers and 345 million parameters, on par with BERT-Large. Our largest model, Grover-Mega, has 48 layers and 1.5 billion parameters, the same as GPT2.

Dataset. We present RealNews, a large corpus of news articles from Common Crawl. Training Grover requires a large corpus of news articles with metadata, but none currently exists. Thus, we construct one by scraping dumps from Common Crawl, limiting ourselves to the 5000 news domains indexed by Google News. We used the Newspaper Python library to extract the body and metadata from each article. News from Common Crawl dumps from December 2016 through March 2019 were used as training data; articles published in April 2019 from the April 2019 dump were used for evaluation. After deduplication, RealNews is 120 gigabytes without compression.

Learning. We trained each Grover model on randomly-sampled sequences from RealNews with length 1024. Other optimization hyperparameters are in Appendix A. We trained Grover-Mega for 800k iterations, using a batch size of 512 and 256 TPU v3 cores. Training time was two weeks.

3.1 Language Modeling results: measuring the importance of data, context, and size

We validate Grover, versus standard unconditional language models, on the April 2019 test set. We consider two evaluation modes: *unconditional*, where no context is provided and the model must generate the article body; and *conditional*, in which the full metadata is provided as context. In both cases, the perplexity is only calculated only over the article body.

Our results, shown in Figure 3, show several conclusions. First, Grover noticeably improves (between .6 to .9 perplexity points) when conditioned on metadata. Second, perplexity decreases with size, with Grover-Mega obtaining 8.7 perplexity in the conditional setting. Third, the data distribution is still important: though the GPT2 models with 117M parameters and 345M parameters respectively match our Grover-Base and Grover-Large architectures, our model is over 5 perplexity points lower in both cases, possibly because the OpenAI WebText corpus also contains non-news articles.

3.2 Carefully restricting the variance of generations with Nucleus Sampling

Sampling from Grover is straightforward as it behaves like a left-to-right language model during decoding. However, the choice of decoding algorithm is important. While likelihood-maximization strategies such as beam search work well for *closed-ended* generation tasks where the output contains the same information as the context (like machine translation), these approaches have been shown

[5] This trick means that Grover is only required to handle $2^{|\mathcal{F}|}$ orderings during training, versus $|\mathcal{F}|!$.

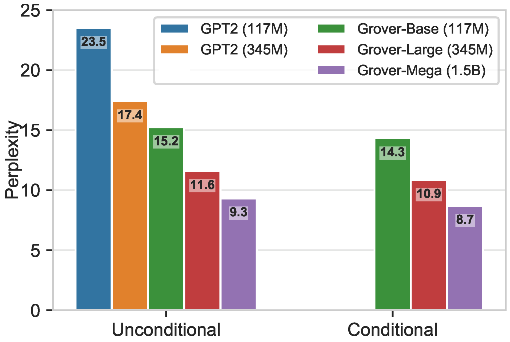


Figure 3: Language Modeling results on the body field of April 2019 articles. We evaluate in the *Unconditional* setting (without provided metadata) as well as in the *Conditional* setting (with all metadata). Grover sees over a 0.6 point drop in perplexity when given metadata.

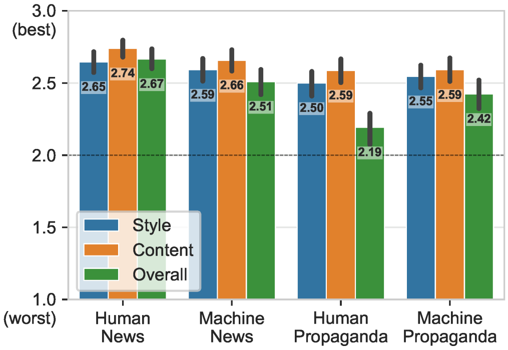


Figure 4: Human evaluation. For each article, three annotators evaluated style, content, and the overall trustworthiness; 100 articles of each category were used. The results show that propaganda generated by Grover is rated more plausible than the original human-written propaganda.

to produce degenerate text during *open-ended* generation (Hashimoto et al., 2019; Holtzman et al., 2019). However, as we will show in Section 6, restricting the variance of generations is also crucial.

In this paper, we primarily use Nucleus Sampling (top-p): for a given threshold p, at each timestep we sample from the most probable words whose cumulative probability comprises the top-p% of the entire vocabulary (Holtzman et al., 2019). We also compare with top-k sampling, wherein the most probable k tokens are used at each timestep (Fan et al., 2018).

4 Humans are Easily Fooled by Grover-written Propaganda

We evaluate the quality of disinformation generated by our largest model, Grover-Mega, using p=.96. We consider four classes of articles: human-written articles from reputable news websites (`Human News`), Grover-written articles conditioned on the same metadata (`Machine News`), human-written articles from known *propaganda* websites (`Human Propaganda`), and Grover-written articles conditioned on the propaganda metadata (`Machine Propaganda`).[6] The domains used are in Appendix B; examples are in Appendix E. We asked a pool of qualified workers on Amazon Mechanical Turk to rate each article on three dimensions: stylistic consistency, content sensibility, and overall trustworthiness.

Results (Figure 4) show a striking trend: though the quality of Grover-written news is not as high as human-written news, it is adept at rewriting propaganda. The overall trustworthiness score of propaganda increases from 2.19 to 2.42 (out of 3) when rewritten by Grover.[7]

5 Neural Fake News Detection

The high quality of neural fake news written by Grover, as judged by humans, makes automatic neural fake news detection an important research area. Using models (below) for the role of the *Verifier* can mitigate the harm of neural fake news by classifying articles as `Human` or `Machine` written. These decisions can assist content moderators and end users in identifying likely (neural) disinformation.

a. Grover. We consider a version of our model adapted for discrimination. Similar to GPT (Radford et al., 2018), we place a special `[CLS]` token at the end of each article, and extract the final hidden state at that point. The hidden state is fed to a linear layer to predict the label `Human` or `Machine`.

To simulate real conditions, and ensure minimal overlap between the generator and discriminator parameters, we initialize Grover for discrimination using the checkpoint at iteration 700k, whereas the generator uses the checkpoint at iteration 800k.

b. GPT2, a 117M or 345M parameter pretrained Transformer language model. Similar to Grover, we follow the GPT approach and extract the hidden state from a newly-added `[CLS]` token.

[6] We use the technique described in Figure 2 to rewrite the propaganda: given the metadata, generate the article first, and then rewrite the headline.

[7] This difference is statistically significant at $p = 0.01$.

Table 1: Results of discriminators versus generators, in both the paired and unpaired settings and across architecture sizes. We also vary the generation hyperparameters for each generator-discriminator pair, reporting the discrimination test accuracy for the hyperparameters with the *lowest* validation accuracy. Compared with other models such as BERT, Grover is the best at detecting its own generations as neural fake news.

		Unpaired Accuracy			Paired Accuracy		
		Generator size			Generator size		
Discriminator size		1.5B	345M	117M	1.5B	345M	117M
	Chance		50.0			50.0	
1.5B	Grover-Mega	**92.0**	**98.5**	**99.8**	**97.4**	**100.0**	**100.0**
345M	Grover-Large	**80.8**	**91.2**	**98.4**	**89.0**	**96.9**	**100.0**
	BERT-Large	73.1	75.9	97.5	84.1	91.5	99.9
	GPT2	70.1	78.0	90.3	78.8	87.0	96.8
117M	Grover-Base	**70.1**	**80.0**	**89.2**	77.5	88.2	95.7
	BERT-Base	67.2	76.6	84.1	**80.0**	**89.5**	**96.2**
	GPT2	66.2	71.9	83.5	72.5	79.6	89.6
11M	FastText	63.8	65.6	69.7	65.9	69.0	74.4

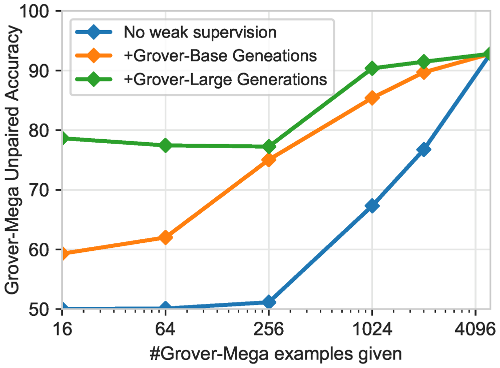


Figure 5: Exploring weak supervision for discriminating Grover-Mega generations. With no weak supervision, the discriminator sees x machine-written articles (from Grover Mega). For +Grover-Base and +Grover-Mega, the discriminator sees $5000-x$ machine-written articles given by the weaker generator in question. Seeing weaker generations improves performance when few in-domain samples are given.

c. BERT, a 117M parameter (BERT-Base) or 345M parameter (BERT-Large) bidirectional Transformer encoder commonly used for discriminative tasks. We perform domain adaptation to adapt BERT to the news domain, as well as to account for long articles; details in Appendix C.

d. FastText, an off-the-shelf library for bag-of-ngram text classification (Joulin et al., 2017). Though not pretrained, similar models do well at detecting human-written fake news.

All models are trained to minimize the cross-entropy loss of predicting the right label. Hyperparameters used during discrimination are in Appendix D.

5.1 A semi-supervised setting for neural fake news detection

While there are many human-written articles online, most are from the distant past, whereas articles to be detected will likely be set in the present. Likewise, there might be relatively few neural fake news articles from a given adversary.[8] We thus frame neural fake news detection as a semi-supervised problem. A neural verifier (or *discriminator*) has access to many human-written news articles from March 2019 and before – the entire RealNews training set. However, it has limited access to generations, and more recent news articles. Using 10k news articles from April 2019, we generate article body text; another 10k articles are used as a set of human-written news articles. We split the articles in a balanced way, with 10k for training (5k per label), 2k for validation, and 8k for testing.

We consider two evaluation modes. In the **unpaired** setting, a discriminator is provided single news articles, and must classify each independently as `Human` or `Machine`. In the **paired** setting, a model is given two news articles with the same metadata, one real and one machine-generated. The discriminator must assign the machine-written article a higher `Machine` probability than the human-written article. We evaluate both modes in terms of accuracy.

5.2 Discrimination results: Grover performs best at detecting Grover's fake news

We present experimental results in Table 1 for all generator and discriminator combinations. For each pair, we show the test results using the most adversarial generation hyperparameters (top-p&k) as judged on the val set.[9] The results show several trends. First, the paired setting appears significantly easier than the unpaired setting across the board, suggesting that it is often difficult for the model to calibrate its predictions. Second, model size is highly important in the arms race between generators and discriminators. Using Grover to discriminate Grover's generations results in roughly 90%

[8] Moreover, since disinformation can be shared on a heterogeneous mix of platforms, it might be challenging to pin down a single generated model.

[9] For each discriminator/generator pair, we search over $p \in \{.9, .92, .94, .96, .98, 1.0\}$ and $k \in \{1, 20, 40\}$.

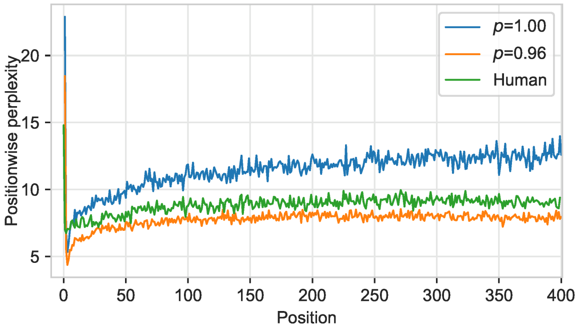


Figure 6: Perplexities of Grover-Mega, averaged over each position in the body (after conditioning on metadata). We consider human-written with Grover-Mega generated text at p=1 (random sampling) and p=.96. The perplexity of randomly sampled text is higher than human-written text, and the gap increases with sequence length. This suggests that sampling without variance reduction often causes generations to fall increasingly out-of-distribution.

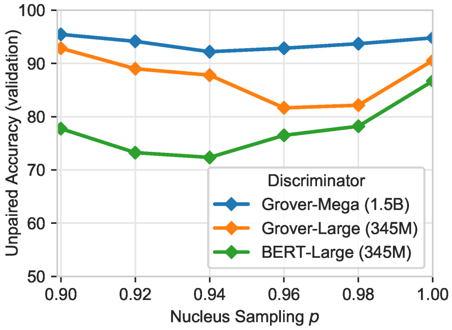


Figure 7: Unpaired validation accuracy, telling apart generated news articles (from Grover Mega) from real articles, at different variance reduction thresholds p (for Nucleus Sampling) and k for top-k. Results varying p show a sweet spot (p = 0.92 – 0.96) wherein discrimination is hardest.

accuracy across the range of sizes. If a larger generator is used, accuracy slips below 81%; conversely, if the discriminator is larger, accuracy is above 98%. Lastly, other discriminators perform worse than Grover overall, even when controlling for architecture size and (for both BERT models) the domain. This suggests that effective discrimination requires having a similar *inductive bias* as the generator.[10]

5.3 Weak supervision: what happens if we don't have access to Grover-Mega?

These results suggest that Grover is an effective discriminator when we have a medium number of fake news examples from the exact adversary that we will encounter at test time. What happens if we relax this assumption? Here, we consider the problem of detecting an adversary who is generating news with Grover-Mega and an unknown top-p threshold.[11] In this setup, during training, we have access to a weaker model (Grover-Base or Grover-Large). We consider the effect of having only x examples from Grover-Mega, and sampling the missing $5000-x$ articles from one of the weaker models, where the top-p threshold is uniformly chosen for each article in the range of $[0.9, 1.0]$.

We show the results of this experiment in Figure 5. The results suggest that observing additional generations greatly helps discrimination performance when few examples of Grover-Mega are available: weak supervision with between 16 and 256 examples from Grover-Large yields around 78% accuracy, while accuracy remains around 50% without weak supervision. As the portion of examples that come from Grover-Mega increases, however, accuracy converges to 92%.

6 How does a model distinguish between human and machine text?

Why does Grover perform best at detecting its own fake news? We hypothesize that the reason may be due in part to *exposure bias*, the phenomenon whereby models maximizing Equation 1 are trained only conditioned on human-written text (Ranzato et al., 2016). To test our hypothesis, in Figure 6 we plot the perplexities given by Grover-Mega over each position for body text at top-p thresholds of 0.96 and 1, as well as over human text. Generating the first token after `<startbody>` results in high perplexity. However, the rest of the positions show a curious pattern: the perplexity of human-written text is lower than randomly sampled text, and this gap increases with sequence length, suggesting that random sampling causes Grover to fall increasingly out of the distribution of human language. However, limiting the variance (p=0.96) lowers the resulting perplexity and limits its growth.

Limiting the variance of a model also creates artifacts On the other hand, clipping the model's variance also leaves an artifact, as prior work has observed for top-k sampling (Strobelt and Gehrmann,

[10] This matches findings on the HellaSwag dataset (Zellers et al., 2019b). Given human text and machine text written by a finetuned GPT model, a GPT discriminator outperforms BERT-Base at picking out human text.

[11] The top-p threshold used was p=0.96, but we are not supposed to know this!

2019). A similar phenomenon holds for Nucleus (top-p) sampling. The probability of observing a human-written article where all tokens are drawn from the top-p% of the distribution is p^n, where n is the document's length. This probability goes to zero as n increases. However, for Nucleus Sampled text – in which the final $1-p$ is cut off – all tokens come from the top-p.

The visibility of the artifacts depends on the choice of discriminator. The top-p at each timestep is calculated under the generator's worldview, meaning that if the discriminator models text in a different way, it might have a harder time pinpointing the empty $1-p$ tail. This could explain BERT's lower performance during discrimination.

A sweet spot of careful variance reduction Not reducing the variance, as well as significantly reducing the variance, both cause problems. Might there be a *sweet spot* for how much to truncate the variance, to make discrimination maximally hard? In Figure 7, we show results varying the top-p threshold for the discrimination task applied to Grover-Mega's generations. The results indeed show a sweet spot, roughly between p=0.92 and p=0.98 depending on the discriminator, wherein discrimination is hardest. Interestingly, we note that the most adversarial top-p threshold for BERT-Large is considerably lower than the corresponding top-p for Grover-Large of the same size. This supports our hypothesis that BERT's view of language differs markedly from Grover; using a lower top-p threshold does not seem to give it much more information about the missing tail.

Overall, our analysis suggests that Grover might be the best at catching Grover because it is the best at knowing where the tail is, and thus whether it was truncated.

7 Conclusion: a Release Strategy for Grover

This paper investigates the threats posed by adversaries seeking to spread disinformation. Our sketch of what these threats might look like – a controllable language model named Grover – suggests that these threats are real and dangerous. Grover can rewrite propaganda articles, with humans rating the rewritten versions as more trustworthy. At the same time, there are defenses to these models – notably, in the form of Grover itself. We conclude with a discussion of next steps and ethical considerations.

The Era of Neural Disinformation. Though training Grover was challenging, it is easily achievable by real-world adversaries today. Obtaining the data required through Common Crawl cost $10k in AWS credits and can be massively parallelized over many CPUs. Training Grover-Mega is relatively inexpensive: at a cost of $0.30 per TPU v3 core-hour and two weeks of training, the total cost is $25k. Spending more money and engineering time could yield even more powerful generators.

Release of generators is critical. At first, it would seem like keeping models like Grover private would make us safer. However, Grover serves as an effective detector of neural fake news, even when the generator is much larger (Section 5). If generators are kept private, then there will be little recourse against adversarial attacks.

Future of progress in generation. Models like BERT are strong discriminators for many NLP tasks, but they are not as good at detecting Grover's generations as Grover itself, even after domain adaptation. One hypothesis is that the artifacts shown in Section 6 are most visible to a left-to-right discriminator. This also suggests that recent progress on generating text in any order (Gu et al., 2019; Stern et al., 2019; Ghazvininejad et al., 2019) may lead to models that evade a Grover discriminator. Likewise, models that are trained conditioned on their own predictions might avoid exposure bias, however, these objectives often lead to low performance on language tasks (Caccia et al., 2018). One additional possibility is the use of Adversarial Filtering (Zellers et al., 2018; 2019b) to oversample and then select a subset of generations. However, we found this didn't work well for very long sequences (up to 1024 BPE tokens), possibly as these are far from the 'Goldilocks Zone' wherein discrimination is hard for machines.

Future of progress in discrimination. Our discriminators are effective, but they primarily leverage distributional features rather than evidence. In contrast, humans assess whether an article is truthful by relying on a model of the world, assessing whether the evidence in the article matches that model. Future work should investigate integrating knowledge into the discriminator (e.g. for claim verification in FEVER; Thorne et al., 2018). An open question is to scale progress in this task towards entire news articles, and without paired evidence (similar to open-domain QA; Chen et al., 2017).

What should platforms do? Video-sharing platforms like YouTube use deep neural networks to scan videos while they are uploaded, to filter out content like pornography (Hosseini et al., 2017). We suggest platforms do the same for news articles. An ensemble of deep generative models, such as Grover, can analyze the content of text – together with more shallow models that predict human-written disinformation. However, humans must still be in the loop due to dangers of flagging real news as machine-generated, and possible unwanted social biases of these models.

Public Release. We plan to make Grover-Base and Grover-Large publicly available. Interested researchers may also apply to download Grover-Mega and RealNews.[12]

Acknowledgments

We thank Dan Weld for providing feedback on this work. Thanks also to Zak Stone and the Google Cloud TPU team for help with the computing infrastructure. This work was supported by the National Science Foundation through a Graduate Research Fellowship (DGE-1256082) and NSF grants (IIS-1524371, 1637479, 165205, 1703166), the DARPA CwC program through ARO (W911NF-15-1-0543), the Sloan Research Foundation through a Sloan Fellowship, the Allen Institute for Artificial Intelligence, the NVIDIA Artificial Intelligence Lab, Samsung through a Samsung AI research grant, and gifts by Google and Facebook. Computations on beaker.org were supported in part by credits from Google Cloud.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Samantha Bradshaw and Philip Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. Technical report, Oxford Internet Institute, 2017.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. *arXiv preprint arXiv:1811.02549*, 2018.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Rachel Dicker. Avoid These Fake News Sites at All Costs. `https://www.usnews.com/news/national-news/articles/2016-11-14/avoid-these-fake-news-sites-at-all-costs`, 2016. [Online; accessed 22-May-2019].

Elizabeth Dwoskin and Tony Romm. Facebook says it has uncovered a coordinated disinformation operation ahead of the 2018 midterm elections. *The Washington Post*, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication 2017-6.*, 2017.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Constant-time machine translation with conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

[12] More up-to-date information about the release policy available at `https://rowanzellers.com/grover`.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order. *arXiv preprint arXiv:1902.01370*, 2019.

Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *arXiv preprint arXiv:1904.02817*, 2019.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernnst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog*, 2018.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Hossein Hosseini, Baicen Xiao, Andrew Clark, and Radha Poovendran. Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api. In *Proceedings of the 2017 on Multimedia Privacy and Security*, pages 21–32. ACM, 2017.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org, 2017.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431, 2017.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *CoRR*, abs/1602.02410, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Clare Melford and Craig Fagan. Cutting the funding of disinformation: The ad-tech solution. Technical report, The Global Disinformation Index, 2019.

Adam Mosseri. News feed fyi: Helping ensure news on facebook is from trusted sources. *Facebook Newsroom*, 19, 2018.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL `https://blog.openai.com/language-unsupervised/`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*. ICLR, 2016.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4603–4611, 2018.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. Insertion transformer: Flexible sequence generation via insertion operations. *arXiv preprint arXiv:1902.03249*, 2019.

Hendrik Strobelt and Sebastian Gehrmann. Catching a unicorn with gltr: A tool to detect automatically generated text. Technical report, Harvard, 2019.

Briony Swire, Ullrich KH Ecker, and Stephan Lewandowsky. The role of familiarity in correcting inaccurate information. *Journal of experimental psychology: learning, memory, and cognition*, 43 (12):1948, 2017.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20(5): 2028–2049, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc., 2017.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

Claire Wardle. Fake news. it's complicated. *First Draft News*, 16, 2017.

Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report, DGI (2017)*, 9, 2017.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019b.

Supplemental Material

A Optimization Hyperparameters

For our input representation, we use the same BPE vocabulary as (Radford et al., 2019). We use Adafactor (Shazeer and Stern, 2018) as our optimizer. Common optimizers such as Adam (Kingma and Ba, 2014) tend to work well, but the memory cost scales linearly with the number of parameters, which renders training Grover-Mega all but impossible. Adafactor alleviates this problem by factoring the second-order momentum parameters into a tensor product of two vectors. We used a maximum learning rate of 1e-4 with linear warm-up over the first 10,000 iterations, and decay over the remaining iterations. We set Adafactor's $\beta_1 = 0.999$ and clipped updates for each parameter to a root-mean-squared of at most 1. Last, we applied weight decay with coefficient 0.01. We used a batch size of 512 on 256 TPU v3 cores. which corresponds to roughly 20 epochs through our news dataset. The total training time required roughly two weeks.

B Real News and Propaganda Websites

In our generation experiments (Section 4), we consider a set of mainstream as well as propaganda websites. We used the following websites as 'real news': `theguardian.com`, `reuters.com`, `nytimes.com`, `theatlantic.com`, `usatoday.com`, `huffingtonpost.com`, and `nbcnews.com`. For propaganda sites, we chose sites that have notably spread misinformation (Dicker, 2016) and/or are alternative media sites with strong political affiliations[13]. These were `breitbart.com`, `infowars.com`, `wnd.com`, `bigleaguepolitics.com`, and `naturalnews.com`.

C Domain Adaptation of BERT

BERT (Devlin et al., 2018) is a strong model for most classification tasks. However, care must be taken to format the input in the right way, particularly because BERT is pretrained in a setting where it is given two spans (separated by a special `[SEP]` token). We thus use the following input format. The first span consists of the metadata, with each field prefixed by its name in brackets (e.g. '`[title]`'). The second span consists of the body. Because the generations are cased (with capital and lowercase letters), we used the 'cased' version of BERT.

Past work (e.g. Zellers et al. (2019a); Han and Eisenstein (2019)) has found that BERT, like other language models, benefits greatly from domain adaptation. We thus perform domain adaptation on BERT, adapting it to the news domain, by training it on RealNews for 50k iterations at a batch size of 256. Additionally, BERT was trained with a sequence length of at most 512 WordPiece tokens, but generations from Grover are much longer (1024 BPE tokens). Thus, we initialized new position embeddings for positions 513-1024, and performed domain adaptation at a length of 1024 WordPiece tokens.

D Hyperparameters for the Discriminators

For our discrimination experiments, we limited the lengths of generations (and human-written articles) to 1024 BPE tokens. This was needed because our discriminators only handle documents up to 1024 words. However, we also found that the longer length empirically discrimination easier for models (see Section 6).

For our discrimination experiments, we used different hyperparameters depending on the model, after an initial grid search. For BERT, we used the Adam (Kingma and Ba, 2014) optimizer with a learning rate of $2e - 5$ and a batch size of 64. We trained BERT models for 5 epochs, with a linear warm-up of the learning rate over the initial 20% iterations. For GPT2 and Grover, we used the Adam actor optimizer (Shazeer and Stern, 2018) optimizer with a learning rate of $2e - 5$ for all models, and a batch size of 64. We applied an auxiliary language modeling loss for these models with a coefficient of 0.5. These models were trained for 10 epochs, with a linear warm-up over the initial 20% iterations.

[13] See `allsides.com/media-bias/media-bias-ratings`.

E Examples

In Figures 8 and 9, we include examples of articles with the average scores given by human raters, who were asked to evaluate the style, content, and overall trustworthiness. In Figure 8, we show a real article (`Human News`) posted by the Guardian along with an article from Grover (`Machine News`) made using the same metadata. Figure 9 shows a real propaganda article from the Natural News (`Human Propaganda`) and an article made with Grover (`Machine Propaganda`) with the original headline and the style of Huffington Post (Grover was used to re-write the title to be more stylistically similar to the Huffington Post, as well).

We also present several other generated examples, generated from Grover-Mega with a top-p threshold of p=0.95. All of the examples are cut off to 1024 generated BPE tokens, since this is our setup for discrimination.

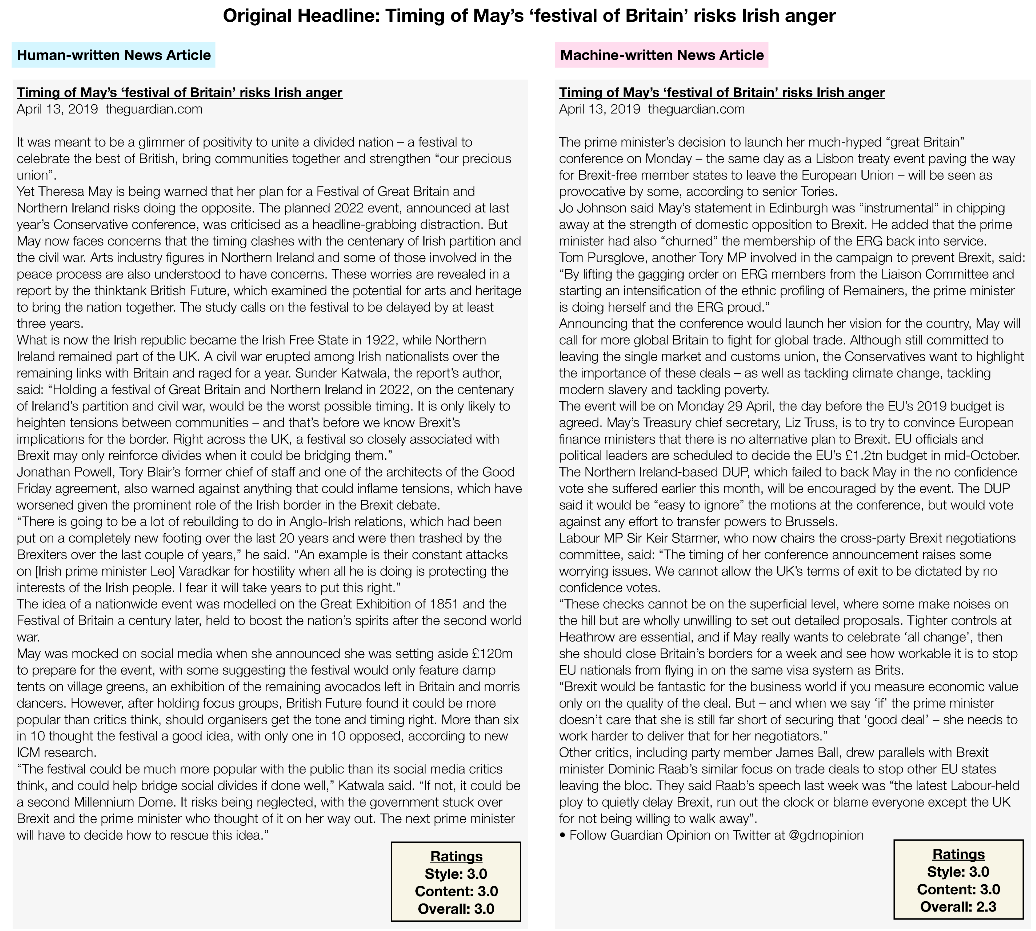
a. Grover can generate controlled propaganda. In Figure 10, we show the continuation from Figure 1, about a link found between autism and vaccines.

b. Grover can spoof the identity of writers. In Figure 11 we show a realistic-looking editorial seemingly from New York Times columnist Paul Krugman.

c. Grover can generate fake political news. In Figure 12 we show an article generated about Trump being impeached, written in the style of the Washington Post.

d. Grover can generate fake movie reviews (opinion spam; Ott et al. (2011)). In Figure 13 we show a movie review, generated in the style of LA Times Movie Critic Kenneth Turan, for Sharknado 6, ‘The Last Sharknado: It’s About Time’

e. Grover can generate fake business news. In Figure 14, we show an article generated about an ‘Uber for Dogs’ startup.



Original Headline: Timing of May’s ‘festival of Britain’ risks Irish anger

Human-written News Article

Timing of May’s ‘festival of Britain’ risks Irish anger
April 13, 2019 theguardian.com

It was meant to be a glimmer of positivity to unite a divided nation – a festival to celebrate the best of British, bring communities together and strengthen “our precious union”.

Yet Theresa May is being warned that her plan for a Festival of Great Britain and Northern Ireland risks doing the opposite. The planned 2022 event, announced at last year’s Conservative conference, was criticised as a headline-grabbing distraction. But May now faces concerns that the timing clashes with the centenary of Irish partition and the civil war. Arts industry figures in Northern Ireland and some of those involved in the peace process are also understood to have concerns. These worries are revealed in a report by the thinktank British Future, which examined the potential for arts and heritage to bring the nation together. The study calls on the festival to be delayed by at least three years.

What is now the Irish republic became the Irish Free State in 1922, while Northern Ireland remained part of the UK. A civil war erupted among Irish nationalists over the remaining links with Britain and raged for a year. Sunder Katwala, the report’s author, said: “Holding a festival of Great Britain and Northern Ireland in 2022, on the centenary of Ireland’s partition and civil war, would be the worst possible timing. It is only likely to heighten tensions between communities – and that’s before we know Brexit’s implications for the border. Right across the UK, a festival so closely associated with Brexit may only reinforce divides when it could be bridging them.”

Jonathan Powell, Tory Blair’s former chief of staff and one of the architects of the Good Friday agreement, also warned against anything that could inflame tensions, which have worsened given the prominent role of the Irish border in the Brexit debate.

“There is going to be a lot of rebuilding to do in Anglo-Irish relations, which had been put on a completely new footing over the last 20 years and were then trashed by the Brexiters over the last couple of years,” he said. “An example is their constant attacks on [Irish prime minister Leo] Varadkar for hostility when all he is doing is protecting the interests of the Irish people. I fear it will take years to put this right.”

The idea of a nationwide event was modelled on the Great Exhibition of 1851 and the Festival of Britain a century later, held to boost the nation’s spirits after the second world war.

May was mocked on social media when she announced she was setting aside £120m to prepare for the event, with some suggesting the festival would only feature damp tents on village greens, an exhibition of the remaining avocados left in Britain and morris dancers. However, after holding focus groups, British Future found it could be more popular than critics think, should organisers get the tone and timing right. More than six in 10 thought the festival a good idea, with only one in 10 opposed, according to new ICM research.

“The festival could be much more popular with the public than its social media critics think, and could help bridge social divides if done well,” Katwala said. “If not, it could be a second Millennium Dome. It risks being neglected, with the government stuck over Brexit and the prime minister who thought of it on her way out. The next prime minister will have to decide how to rescue this idea.”

Ratings
Style: 3.0
Content: 3.0
Overall: 3.0

Machine-written News Article

Timing of May’s ‘festival of Britain’ risks Irish anger
April 13, 2019 theguardian.com

The prime minister’s decision to launch her much-hyped “great Britain” conference on Monday – the same day as a Lisbon treaty event paving the way for Brexit-free member states to leave the European Union – will be seen as provocative by some, according to senior Tories.

Jo Johnson said May’s statement in Edinburgh was “instrumental” in chipping away at the strength of domestic opposition to Brexit. He added that the prime minister had also “churned” the membership of the ERG back into service.

Tom Pursglove, another Tory MP involved in the campaign to prevent Brexit, said: “By lifting the gagging order on ERG members from the Liaison Committee and starting an intensification of the ethnic profiling of Remainers, the prime minister is doing herself and the ERG proud.”

Announcing that the conference would launch her vision for the country, May will call for more global Britain to fight for global trade. Although still committed to leaving the single market and customs union, the Conservatives want to highlight the importance of these deals – as well as tackling climate change, tackling modern slavery and tackling poverty.

The event will be on Monday 29 April, the day before the EU’s 2019 budget is agreed. May’s Treasury chief secretary, Liz Truss, is to try to convince European finance ministers that there is no alternative plan to Brexit. EU officials and political leaders are scheduled to decide the EU’s £1.2tn budget in mid-October.

The Northern Ireland-based DUP, which failed to back May in the no confidence vote she suffered earlier this month, will be encouraged by the event. The DUP said it would be “easy to ignore” the motions at the conference, but would vote against any effort to transfer powers to Brussels.

Labour MP Sir Keir Starmer, who now chairs the cross-party Brexit negotiations committee, said: “The timing of her conference announcement raises some worrying issues. We cannot allow the UK’s terms of exit to be dictated by no confidence votes.

“These checks cannot be on the superficial level, where some make noises on the hill but are wholly unwilling to set out detailed proposals. Tighter controls at Heathrow are essential, and if May really wants to celebrate ‘all change’, then she should close Britain’s borders for a week and see how workable it is to stop EU nationals from flying in on the same visa system as Brits.

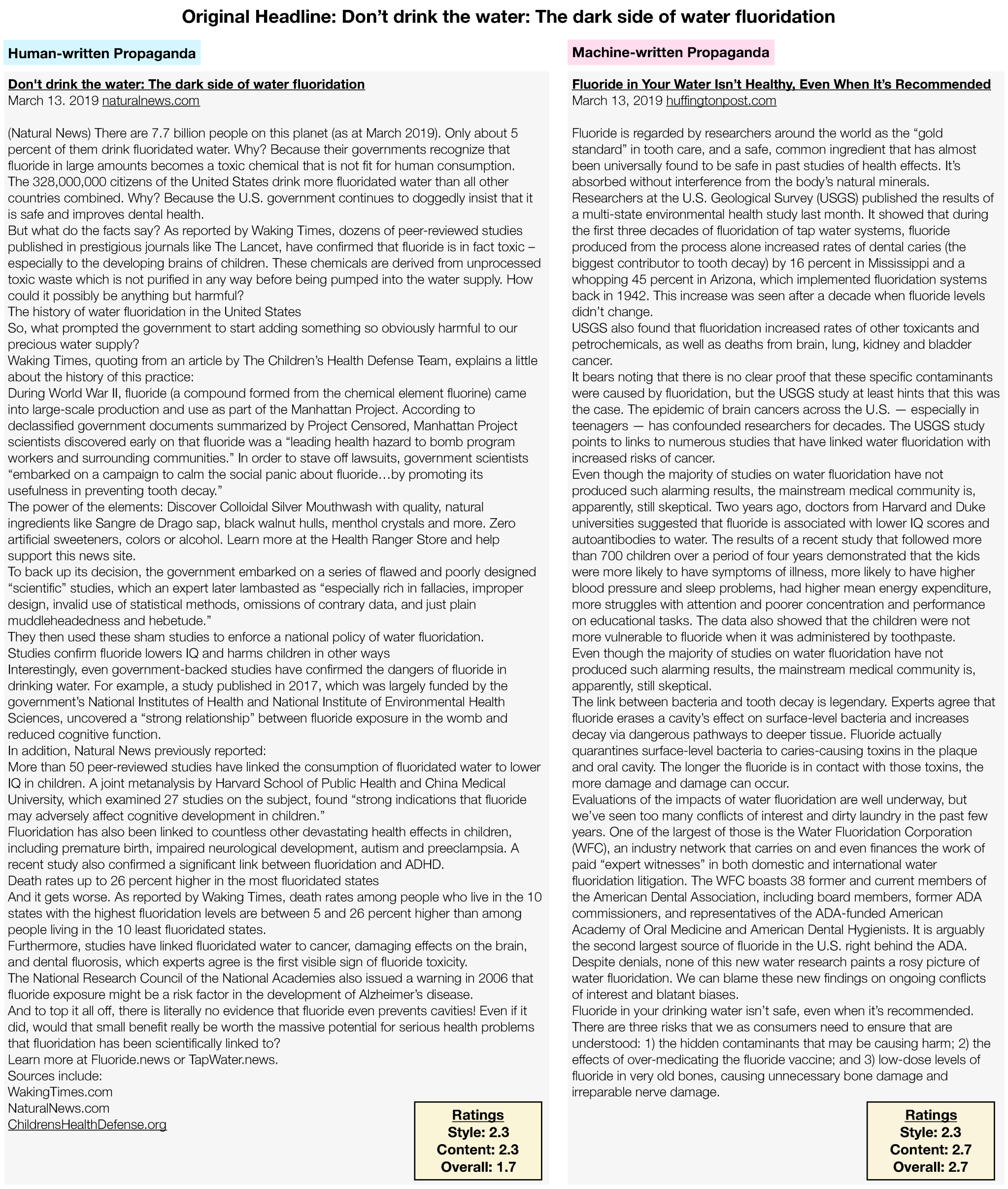
“Brexit would be fantastic for the business world if you measure economic value only on the quality of the deal. But – and when we say ‘if’ the prime minister doesn’t care that she is still far short of securing that ‘good deal’ – she needs to work harder to deliver that for her negotiators.”

Other critics, including party member James Ball, drew parallels with Brexit minister Dominic Raab’s similar focus on trade deals to stop other EU states leaving the bloc. They said Raab’s speech last week was “the latest Labour-held ploy to quietly delay Brexit, run out the clock or blame everyone except the UK for not being willing to walk away”.

- Follow Guardian Opinion on Twitter at @gdnopinion

Ratings
Style: 3.0
Content: 3.0
Overall: 2.3

Figure 8: Example of human-written news and machine-written news articles about the same headline from The Guardian with the average ratings from human rating study.



Original Headline: Don’t drink the water: The dark side of water fluoridation

Human-written Propaganda

Don't drink the water: The dark side of water fluoridation
March 13. 2019 naturalnews.com

(Natural News) There are 7.7 billion people on this planet (as at March 2019). Only about 5 percent of them drink fluoridated water. Why? Because their governments recognize that fluoride in large amounts becomes a toxic chemical that is not fit for human consumption.

The 328,000,000 citizens of the United States drink more fluoridated water than all other countries combined. Why? Because the U.S. government continues to doggedly insist that it is safe and improves dental health.

But what do the facts say? As reported by Waking Times, dozens of peer-reviewed studies published in prestigious journals like The Lancet, have confirmed that fluoride is in fact toxic – especially to the developing brains of children. These chemicals are derived from unprocessed toxic waste which is not purified in any way before being pumped into the water supply. How could it possibly be anything but harmful?

The history of water fluoridation in the United States

So, what prompted the government to start adding something so obviously harmful to our precious water supply?

Waking Times, quoting from an article by The Children’s Health Defense Team, explains a little about the history of this practice:

During World War II, fluoride (a compound formed from the chemical element fluorine) came into large-scale production and use as part of the Manhattan Project. According to declassified government documents summarized by Project Censored, Manhattan Project scientists discovered early on that fluoride was a “leading health hazard to bomb program workers and surrounding communities.” In order to stave off lawsuits, government scientists “embarked on a campaign to calm the social panic about fluoride…by promoting its usefulness in preventing tooth decay.”

The power of the elements: Discover Colloidal Silver Mouthwash with quality, natural ingredients like Sangre de Drago sap, black walnut hulls, menthol crystals and more. Zero artificial sweeteners, colors or alcohol. Learn more at the Health Ranger Store and help support this news site.

To back up its decision, the government embarked on a series of flawed and poorly designed “scientific” studies, which an expert later lambasted as “especially rich in fallacies, improper design, invalid use of statistical methods, omissions of contrary data, and just plain muddleheadedness and hebetude.”

They then used these sham studies to enforce a national policy of water fluoridation.

Studies confirm fluoride lowers IQ and harms children in other ways

Interestingly, even government-backed studies have confirmed the dangers of fluoride in drinking water. For example, a study published in 2017, which was largely funded by the government’s National Institutes of Health and National Institute of Environmental Health Sciences, uncovered a “strong relationship” between fluoride exposure in the womb and reduced cognitive function.

In addition, Natural News previously reported:

More than 50 peer-reviewed studies have linked the consumption of fluoridated water to lower IQ in children. A joint metanalysis by Harvard School of Public Health and China Medical University, which examined 27 studies on the subject, found “strong indications that fluoride may adversely affect cognitive development in children.”

Fluoridation has also been linked to countless other devastating health effects in children, including premature birth, impaired neurological development, autism and preeclampsia. A recent study also confirmed a significant link between fluoridation and ADHD.

Death rates up to 26 percent higher in the most fluoridated states

And it gets worse. As reported by Waking Times, death rates among people who live in the 10 states with the highest fluoridation levels are between 5 and 26 percent higher than among people living in the 10 least fluoridated states.

Furthermore, studies have linked fluoridated water to cancer, damaging effects on the brain, and dental fluorosis, which experts agree is the first visible sign of fluoride toxicity.

The National Research Council of the National Academies also issued a warning in 2006 that fluoride exposure might be a risk factor in the development of Alzheimer’s disease.

And to top it all off, there is literally no evidence that fluoride even prevents cavities! Even if it did, would that small benefit really be worth the massive potential for serious health problems that fluoridation has been scientifically linked to?

Learn more at Fluoride.news or TapWater.news.

Sources include:
WakingTimes.com
NaturalNews.com
ChildrensHealthDefense.org

Ratings
Style: 2.3
Content: 2.3
Overall: 1.7

Machine-written Propaganda

Fluoride in Your Water Isn’t Healthy, Even When It’s Recommended
March 13, 2019 huffingtonpost.com

Fluoride is regarded by researchers around the world as the “gold standard” in tooth care, and a safe, common ingredient that has almost been universally found to be safe in past studies of health effects. It’s absorbed without interference from the body’s natural minerals.

Researchers at the U.S. Geological Survey (USGS) published the results of a multi-state environmental health study last month. It showed that during the first three decades of fluoridation of tap water systems, fluoride produced from the process alone increased rates of dental caries (the biggest contributor to tooth decay) by 16 percent in Mississippi and a whopping 45 percent in Arizona, which implemented fluoridation systems back in 1942. This increase was seen after a decade when fluoride levels didn’t change.

USGS also found that fluoridation increased rates of other toxicants and petrochemicals, as well as deaths from brain, lung, kidney and bladder cancer.

It bears noting that there is no clear proof that these specific contaminants were caused by fluoridation, but the USGS study at least hints that this was the case. The epidemic of brain cancers across the U.S. — especially in teenagers — has confounded researchers for decades. The USGS study points to links to numerous studies that have linked water fluoridation with increased risks of cancer.

Even though the majority of studies on water fluoridation have not produced such alarming results, the mainstream medical community is, apparently, still skeptical. Two years ago, doctors from Harvard and Duke universities suggested that fluoride is associated with lower IQ scores and autoantibodies to water. The results of a recent study that followed more than 700 children over a period of four years demonstrated that the kids were more likely to have symptoms of illness, more likely to have higher blood pressure and sleep problems, had higher mean energy expenditure, more struggles with attention and poorer concentration and performance on educational tasks. The data also showed that the children were not more vulnerable to fluoride when it was administered by toothpaste.

Even though the majority of studies on water fluoridation have not produced such alarming results, the mainstream medical community is, apparently, still skeptical.

The link between bacteria and tooth decay is legendary. Experts agree that fluoride erases a cavity’s effect on surface-level bacteria and increases decay via dangerous pathways to deeper tissue. Fluoride actually quarantines surface-level bacteria to caries-causing toxins in the plaque and oral cavity. The longer the fluoride is in contact with those toxins, the more damage and damage can occur.

Evaluations of the impacts of water fluoridation are well underway, but we’ve seen too many conflicts of interest and dirty laundry in the past few years. One of the largest of those is the Water Fluoridation Corporation (WFC), an industry network that carries on and even finances the work of paid “expert witnesses” in both domestic and international water fluoridation litigation. The WFC boasts 38 former and current members of the American Dental Association, including board members, former ADA commissioners, and representatives of the ADA-funded American Academy of Oral Medicine and American Dental Hygienists. It is arguably the second largest source of fluoride in the U.S. right behind the ADA.

Despite denials, none of this new water research paints a rosy picture of water fluoridation. We can blame these new findings on ongoing conflicts of interest and blatant biases.

Fluoride in your drinking water isn’t safe, even when it’s recommended.

There are three risks that we as consumers need to ensure that are understood: 1) the hidden contaminants that may be causing harm; 2) the effects of over-medicating the fluoride vaccine; and 3) low-dose levels of fluoride in very old bones, causing unnecessary bone damage and irreparable nerve damage.

Ratings
Style: 2.3
Content: 2.7
Overall: 2.7

Figure 9: Example of human-written and machine-written articles arguing against fluoride with the average ratings from human rating study.

Provided headline: ‘Link found between autism and vaccines’

Generated body: Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health.

Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism. Earlier studies had linked past vaccines with the condition, but it was not clear whether those links were causal or indirect. Among research groups, the UCSD and CDC teams used new methods that enabled them to determine the associations between the virus and the disease.

While researchers in the previous studies had observed associations between those vaccinated against measles and autism in children whose mothers had received the virus, the current UCSD/CDC team focused on childhood vaccines given during the first two years of life because it is commonly believed those vaccines begin to clear the body’s immune system before the onset of autism. The UCSD/CDC team compared children who had been vaccinated against measles, mumps and rubella (MMR) and those who had not.

“Because this is the first study to focus on MMR-vaccine syndrome, it was also the first time that this syndrome had been scientifically characterized,” said study lead author William Chan, MD, MPH, a researcher in the Department of Pediatrics in the UC San Diego School of Medicine. “We were able to identify certain subgroups of those vaccinated against MMR and compare them to those not vaccinated.”

Mumps, measles and rubella vaccines are administered in either three or five doses, depending on the age. A recent study by the Public Health Agency of Canada found only 10 percent of children vaccinated against measles had a developing autism disorder when compared to 73 percent of children vaccinated with one dose, 50 percent of those vaccinated with two doses and 45 percent of those vaccinated with three doses. Those results do not surprise the UC San Diego researchers, who found those vaccinated with three or more doses were at an additional 5- to 8-fold increased risk.

The current UCSD/CDC analysis followed the same years of data as earlier studies, but used new methods to accurately pinpoint the timing of those immunity-shifting epidemics. For example, previous studies showed that autism was on the rise in 1995-96, but never topped 50 cases until 2001-02, when measles overtook mumps as the leading cause of infant death due to disease. In 2003, several measles outbreaks were triggered by parents who had ignored warnings about the vaccine.

“Our research was critical in providing details of when outbreaks occurred in regions where there was significant vaccination efforts,” said study co-author Louis Cooley, MD, PhD, professor of Pediatrics and professor of Communicable Diseases and Epidemiology at UCSD School of Medicine. “We were able to show that if you get your MMR vaccine in early childhood, you should reduce your risk of autism by 33 percent, with every passing year bringing a slightly lower risk.”

Researchers found that low vaccination rates — 42 percent in 2003-04 and 75 percent in 2007-08 — and more frequent, and severe, cases of measles predicted higher rates of autism in later years. In 2004-05, the annual rate of autism and other disorders — including intellectual disability — increased by 20 percent, far more than the 5- to 7-fold increase in their model predicted, suggesting there may be a more subtle link between vaccine exposure and autism than previously understood. Preliminary findings from a separate analysis of the 2006-07 data suggest that autism risk spikes occurred during times when children had fewer routine vaccinations and received older doses of vaccines.

Autism spectrum disorders include autism spectrum disorders (ASD), a group of neurodevelopmental disorders that can occur with or without symptoms of the immune system disorder, including Asperger’s syndrome, pervasive developmental disorder not otherwise specified (PDD-NOS) and pervasive developmental disorder-not otherwise specified (PDD-ENG).

The study was supported by the National Institutes of Health’s National Institute of Child Health and Human Development (CDER), the NIH Program on Smoking and Health and the San Diego Veterans Affairs (SDVA) Healthcare System.

About the University of California, San Diego School of Medicine

Founded in 1897, the School of Medicine has been recognized nationally for outstanding achievement in educating clinical and research physicians. Nobel laureates and more than 170 physicians who practice in the United States and abroad are the graduates of UC San Diego. The school has been recognized as the first medical school in the world to offer its entire physician-practitioner curriculum online, currently serving more than 1,000 medical students worldwide. The school also trains the next generation of scientists, physicians and public health leaders and ranks first in the nation for its graduates’ ability to train other healthcare professionals. UC San Diego is associated with UC San Diego Health System, which has a faculty of more than 9,000 and an annual

Figure 10: The rest of the example from the teaser figure, about a hypothetical link found between autism and vaccines.

Provided domain: `nytimes.com`
Provided author: Paul Krugman
Provided date: May 23, 2019

Generated headline: Why Trump could get impeached

Generated body: President Donald Trump could get impeached. Or he could get fired. Or he could be indicted for special-counsel matters involving Stormy Daniels, or the Russians, or his own campaign. But, at the end of the day, what would matter is what House Speaker Nancy Pelosi decides. And here's why:

First, Pelosi seems open to the idea of impeachment, to the point where last week she endorsed the use of the word in a speech. And this is not to claim that "impeachment" is a possible solution to the current impasse on immigration. Nobody seriously thinks that. Rather, it's to say that the dividing line between Trump's impeachment and, on the other hand, the notion that he's being treated adequately in his courtship of the Republican base, is soft. If Trump runs a replay of his 2016 campaign, the focus on immigration will fade as an issue. Then his campaign's fiery rhetoric on the economy and taxes will dominate, and get bigger. But as we pointed out yesterday, the "soul" of the Trump campaign was white nationalist sloganeering and authoritarianism. It seems unlikely that this will change.

If it doesn't change, impeachment seems like the only way to constrain him.

Yet the power of the impeachment path is out of Trump's hands. The recent passage of the Defense Authorization Act, to cite just one example, which codifies within the law the requirement that presidents get an honest break from investigations, seems to have guaranteed that Bob Mueller will be working hard into the fall. And "executive privilege" — the right of a president to withhold certain materials from the government — is clearly limited by the extent to which federal judges have looked at and rejected it.

So impeachment is less a decision than a historical fact: If you think the election of a president who promised to enrich himself through corruption has happened, a court-ordered investigation, or both, is probably necessary to correct the damage. But it can't happen while he's president, or until he's gone. Even then, there are a few ways that a president can claim executive privilege, or whatever other arguments he can throw out to minimize the pressure on him.

So Congress must seize the moment. The first and most urgent must be an investigation of the Trump campaign's possible ties to the Russians, and the subsequent question of whether the president knew what was going on. (For full coverage of the Russia story, see here.) The second, more abstract but no less important issue is whether the president's own actions constitute obstruction of justice. And third, Trump needs to be impeached because his rhetoric has helped put members of his own party who could back his impeachment into near paralysis.

If Republicans don't take a bold stand, and appoint a special prosecutor to investigate all this — just as Democrats did during Watergate — then their caucus will become dependent on impeachment proceedings to hold its seats. While that might be a good outcome in the short run, it could also signal a likely government shutdown — just as it signaled the risk of a government shutdown during the Clinton impeachment. So it's a bad decision for the country, or at least for the country's self-interest.

That said, it could be an even worse one for Trump. True, other presidents have been impeached in part for behavior far more extreme than Trump's — for cases like Bill Clinton's loveless marriage, and, more frequently, for their handling of women in private. But never before has the impeachment authority come at the point where a man at the pinnacle of power and influence has cozied up to a hostile foreign power, and openly attacked not just the judiciary, but the rule of law itself.

For those who would recommend removal now for that alone, it's worth noting that on a serious level, that kind of behavior isn't akin to the brinkmanship that some U.S. presidents have brought to bear during a confrontation with Cuba. It's as close as Trump has come to laying out some of the first lines of battle for our uncertain and uneasy world.

And to me, that is a form of obstruction of justice.

Figure 11: An example of spoofing someone's identity using Grover-Mega. We generate a convincing New York Times Op-Ed that is seemingly from columnist Paul Krugman.

Provided domain: `washingtonpost.com`
Provided date: May 29, 2019
Provided headline: US House Votes to Impeach Donald Trump After New Evidence Released from Mueller Report

Generated body: WASHINGTON — The House voted to impeach President Donald Trump Wednesday after releasing hundreds of pages of text messages that point to clear evidence of obstruction of justice and communication with the head of the Trump Organization about a potential business deal in Russia.

The 220-197 vote came after weeks of debate over whether new evidence released by special counsel Robert Mueller's office signaled sufficient grounds for Trump's removal from office. The president personally denounced the move, announcing his intent to veto the resolution and accusing Democrats of plotting to remove him from office through a "con job."

"This is all a big con job," he said during an appearance in Texas. "They got caught."

A number of Democratic lawmakers went further, calling for Trump's impeachment on the grounds that he obstructed justice. Democrats hold the House majority, and if the Senate fails to convict Trump, the president can be removed from office by a two-thirds vote.

But even with Trump's political allies organizing from the White House, Republicans in the House stood united behind their leader, noting the historical difficulty of making the case for impeachment.

"This is not a decision for the president," said House Speaker Paul Ryan, R-Wis. "It's not one for this body, and I think it's a dereliction of duty for the minority party to be making these calls."

Democrats on the House Judiciary Committee on Tuesday said they were not bound by Ryan's defense. They released an 11-page joint report with Republicans on the House Oversight and Government Reform Committee that took a different tack, emphasizing that Mueller had not decided whether the case warranted impeachment. Rep. Jerrold Nadler, D-N.Y., the committee's chairman, told reporters before the vote that he hoped the two reports would "send a clear message to the president that we in this body are not going to stand for obstruction of justice."

But when asked whether the report would build the case for impeachment, he said: "What's far more likely to build a case for impeachment are the indictments and evidence produced through the special counsel's work."

On Tuesday night, the White House issued a statement proclaiming that the book by a former FBI deputy director, which has been cited by Democrats in support of impeachment, had been "made up and defamatory" to the president.

That statement drew swift and heavy rebukes from a number of Democrats, who said they had found the statement "beyond astonishing."

The Democratic report, citing redacted text messages sent by former FBI counterintelligence agent Peter Strzok, revealed what the party called "a pattern of bias" and "an affinity for the Russian regime."

The report did not include a full, uncensored text message Strzok sent to a friend after Trump was elected that said: "I'll stop it." That led Democrats to bemoan the language used, questioning what is or is not in the special counsel's investigative file.

The report also quoted a review of emails and witness testimony to Mueller that said Trump "has obstructed justice in the Russia investigation" by firing FBI Director James Comey, by urging Attorney General Jeff Sessions to recuse himself from the Russia investigation and by urging his son, Donald Trump Jr., to meet with a Kremlin-connected Russian lawyer at Trump Tower in 2016.

"This has never happened in the history of our country," Nadler said of the president. "No president has ever tried to interfere with an investigation into his own conduct."

The vote was also led by Democrats who have been outspoken against Trump, including Rep. Al Green, D-Texas, who took out newspaper ads calling for the president's impeachment. Rep. Maxine Waters, D-Calif., another outspoken Trump critic, said Trump's ties to Russia were "becoming more and more ominous," and said "as we go on, we are going to see if we can't get the articles of impeachment out."

Rep. Tony Cárdenas, D-Calif., a member of the Judiciary Committee, said many Democrats would be eager to see whether the newly discovered evidence could be used to support impeachment.

"I think it will be the nail in the coffin of the 'non-con,'" he said.

Generated headline: Congress voted to impeach President Trump Wednesday; the President vowed to veto the resolution.

Figure 12: An example of using Grover-Mega to generate fake political news, written in the style of the Washington Post. Here, it generates a mostly-realistic-looking article about Trump being impeached.

Provided domain: `latimes.com`
Provided date: August 1, 2018
Provided authors: Kenneth Turan
Provided headline: Review: "The Last Sharknado: It’s About Time" is a gem of an action movie

---

Generated body: Some heroes take a vacation every now and then, even temporarily, a couple of beachfront condos in a movie world they hope, a la Rick James in “The Beautiful Daze,” won’t have to live through. Unfortunately, that has never happened to one of the most exceptional directors in the business, Anthony & Joe Russo, so the moment is altogether too good to spoil, at least for us Southern Californians.

In “The Last Sharknado: It’s About Time,” which had its Los Angeles premiere at the Tower Theatre on Tuesday, we can expect the San Diego-born Anthony Russo to somehow make it through seven years without being inside a sharknado.

Oh, he has a couple of other movies during that time, mostly that stupid’d-up New York “Avengers” picture, but they have nothing to do with sharks or serious action-film plotting.

Don’t worry, he lives to fight another day.

Written by Jason Friedberg and Aaron Seltzer (“Vampires Suck”) and directed by John Hlavin (“Code Name Verity”), “The Last Sharknado” is exactly what you would expect: very silly and thrilling.

Once again the Ryan Newman- and Ian Ziering-costumed stars of the YouTube smash “Sharknado” have a lot of fun poking fun at themselves and the silly film business (“I don’t know if being caged for a few days after we get out is too much to ask”) along with helping a once-beleaguered government figure out what to do.

We start on a lake in North Carolina as Newman, now an exotic dancer, does a good-natured impersonation of Seltzer’s Agent Slutz. He’s not a bad actor, though, and by the end of the first reel “it’s clear he’s never a lunatic.”

So Slutz has just been rehired by the Trump Administration, even though the agency has been effectively taken over by the tag team of the Governator and Alex Jones (the viewer is never told why they are being hired).

As Slutz, Newman is working with his deputy, Ziering, the impetuous marine officer no longer needled by his loyalty to his best friend, Tara Reid, who showed up in the third “Sharknado” and is absent this time.

The three are doing their best to fix things, with Newman looking in particular for more information on how to deal with the five remaining sharks in New York who seem to think they should go with the Trump administration.

Svelt’s plotting goes way beyond the call of duty, of course, for despite the presence of Anthony Mackie and Juliette Lewis, for example, there are many disappointments here, especially in the casting of a grouchy Jeff Goldblum as the villain of the piece.

Still, like the original “Sharknado,” all the elements are exactly right, including some fancy effects (stunt coordinator Zak Penn, “The Avengers”) and some deeply, bizarrely funny dialogue (try spending a few days in a sharknado and after a while questioning your sanity).

Knowing that everything is going to be fine, Ziering tells Newman: “Whether we go sharknado-free is on you.”

Come again?

Might we welcome it? If by a miracle we’re not sated by our popcorn, that would be the best way to look at it.

kenneth.turan@latimes.com

Follow me on Twitter @KennethTuran

Generated headline: ’The Last Sharknado’ returns with the dramatic thriller you expected — and some unexpected laughs

---

Figure 13: An example of using Grover-Mega to generate a fake movie review. Here, we spoof LA Times Film Critic Kenneth Turan and generate a positive movie review for ‘The Last Sharknado: It’s About Time,’ the sixth installment in the Sharknado series of movies.

Provided domain: techcrunch.com
Provided date: May 29, 2019
Provided headline: Rowan Zellers' new startup raises 17M in Series A round for "Uber for Dogs"

Generated body: In today's competitive eco-system of startups trying to convince consumers to eschew popular apps in favor of their own, you would think that the dog-walking market would be the easiest to crack. Well, it's not. But at least one startup believes it has a new and clever solution: apply its core offering of mobile-based recommendation engine algorithms to locate dog owners near them with the end goal of making dog walking more efficient.

The company, owned by Australian startup Yfrog, is called DogSpotter. It is co-founded by entrepreneur Rowan Zellers, along with King's College London-trained dogsman Craig Osbaldoon.

Launched in Melbourne in November, DogSpotter has now launched in Sydney, Canberra, Adelaide, Perth, and is set to be introduced to Brisbane and Hobart in June, with a platform the company expects to be available in all capital cities in Australia this year. In addition, it is now prepared to accept payment via credit cards in Australia and New Zealand.

In exchange for its name recognition and the unique proposition that it is aiming to turn into a software-as-a-service, DogSpotter has raised $17 million in a Series A round of funding led by Plug and Play Tech Center, with participation from Open Ocean Capital and previous investors, including the Cahill Foundation.

So what does DogSpotter offer, beyond a fairly basic service that could theoretically be used by anyone, which is at the base of a human-side algorithm that determines the best potential location for a dog owner's pet — a problem that dogs are notoriously impatient for in cities where dog walking is often unprofitable or a lost art for many?

Well, DogSpotter leverages its core recommendation engine to process the thousands of now readily available reviews of dog care providers via its app that users can find by searching for their own city or a curated selection of recommendations. That way, it's much quicker than traditional alternatives, where you have to wade through a fair number of reviews to find the best-rated providers in your city, and does it with far less hassle.

From there, the DogSpotter app makes use of the data it provides to recommend relevant walking services, among them a variety of preferred types of walks for your dog, based on everything from their preferences and sorts of paths to the weather, geographic differences, and degree of safety.

DogSpotter's founders also contend that their software can — and in some cases already has been — used to make sure that the actual human responsible for dog walking is nearby, by using the location-based data to facilitate long-distance walkings or picking out a dog for a veterinarian who can be contacted.

Speaking of which, they're not just operating on the basis that it's an effective way to walk dogs without suffering losses. Although I won't deny that I've seen some big cities suffer from too many small businesses advertising purely online and with little regard for actual human location. As it happens, DogSpotter tells me it has its own network of paying businesses in places like U.S. cities, but it claims to be seeing a company like DogBuddy.com.au as a market that it is disrupting.

My hope is that the company ends up developing a number of other application-specific verticals such as veterinary procedures — again, I have not seen this yet but that would be a great and easy way to monetize.

The business model is to offer DogSpotter services for a set fee per user, per location per month. It includes complimentary services for canine community or adoption groups, dog run assessment and mapping, dog boarding facilities, dog walking packages, and a myriad of other services.

"DogSpotter is different from other dog walking platforms because we have the leading mobile app which is the most effective tool for location-based visual targeting for relevant content and experiences," said Rowan Zellers, founder and CEO of DogSpotter, in a statement. "This means that DogSpotter is able to provide more relevant products and services and, for early users, we can allow companies to target only the users they want to engage with."

The problem, of course, is that not all dog walkers have the best apps. DogSpotter says it will compete on performance of its offering, not with other dog walking apps but rather with making it easier for dog owners to find the people they are looking for and, importantly, for those individual dog walkers to save a small bit of money on their operating costs, by saving time and effort with a website. And so on.

Generated headline: DogSpotter raises $17M to improve your search for walking services

Figure 14: An example of using Grover-Mega to generate fake business news. This generates an article about a fake startup for 'Uber for Dogs', ostensibly created by the first author of this paper.