"All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation."

Sarah Kreps, Miles McCain, and Miles Brundage
August 2020

Abstract

Online misinformation has become a constant; only the way actors create and distribute that information is changing. Advances in artificial intelligence (AI) mean that actors can now synthetically generate text in ways that mimic the style and substance of human-created news stories. We carried out three original experiments to study whether these AI-generated texts are credible and can influence opinions on foreign policy—a likely target of real-world misinformation. The first evaluated human detection of AI-generated text relative to the original story from which it was generated. The second examined the credibility distribution across different model sizes to gauge whether improvements in processing produce commensurate increases in credibility. The third investigated the interaction between partisanship and AI-generated news. We find that individuals are largely incapable of distinguishing between AI and human-generated text; partisanship affects the perceived credibility of the story; and exposure to the text does little to change individuals' policy views. The findings have important implications for the way malicious actors might employ AI in online misinformation campaigns and electoral interference.

Replication Data available at:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XVYU3

Introduction

The 2016 United States presidential election brought into sharp relief the ways that foreign actors can influence media content, shape public attitudes about political candidates, and potentially impact electoral outcomes. Online hyper-targeting—directing particular news stories or advertising to specific demographics with the aim of polarizing and tilting political preferences—has triggered countermeasures: social media sites such as Twitter have restricted political advertising.[1] A 2019 US Senate committee on foreign interference in US elections recommended that the government "reinforce with the public the danger of attempted foreign interference in elections."[2] Implicit in this statement is that the public is a bulwark against interference but only through awareness and proper circumspection on how to jettison misinformation.

At the same time, the legislative group noted that Russian efforts to interfere are becoming "more sophisticated" in ways that would thwart the public's ability to discern fact from fiction. Indeed, new state-of-the-art artificial intelligence (AI) technologies can now generate text that mimics the style and substance of real news stories while overcoming the bandwidth limits that face human-generated text. If these models enable malicious actors to generate and publish credible-sounding news stories at scale, then the prospect for misinformation, defined as "false or misleading information,"[3] is high: the volume of inauthentic media could balloon, and the ease of text synthesis might further enable the coordinated hyper-targeting of articles to individual groups. However, if the

[1] Justin Sherman, "The fight over social media's potent political ads just got more interesting," *The Bulletin*, 1 November 2019.
[2] Maggie Miller, "Senate Intel report urges action to prevent Russian meddling in 2020 election," *The Hill*, 8 October 2019.
[3] David Lazer, Matthew Baum, Yochai Benkler, Adam Berinsky, Kelly Greenhill, Filippo Menczer, Miriam Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven Sloman, Cass Sunstein, Emily Thorson, Duncan Watts, and Jonathan Zittrain, "The Science of Fake News," *Science*, 359 no. 6380, 9 March 2018; 1094-1096; 1096.

public does not find these synthetic news stories credible, the advent of these technologies currently poses little threat to democratic institutions.

Despite the potential for increasingly sophisticated technologies to be vehicles for misinformation, Lazer et al. note that "there is little research focused on fake news and no comprehensive data-collection system to provide a dynamic understanding of how pervasive systems of fake news provision are evolving."[4] Moreover, there is no systematic evidence as to whether or how the use of these emerging technologies might influence the public, either creating confusion or shaping political beliefs, and whether the public is in a position to "decide what is credible," an approach Facebook CEO Mark Zuckerberg endorsed as the appropriate way to filter out misinformation online.[5]

In this research, we carried out three main experiments to study the degree to which emerging AI tools can generate credible-sounding texts and shape attitudes about foreign policy. The first evaluated the upper bound of human credibility perceptions of AI-generated text compared to a real news baseline. The second investigated the effect of partisanship on AI-generated news credulity—engaging debates about whether partisans are more likely to believe politically congenial news stories and more likely to continue believing those even in the face of disclaimers, making the prospect of targeted synthetic misinformation high.[6] Finally, the third

[4] Lazer et al., "The Science of Fake News," 1096.
[5] "Facebook's Mark Zuckerberg: Private companies should not censor politicians," *CNN*, 18 October 2019.
[6] Kevin Arceneaux, Martin Johnson, and Chad Murphy, Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure," *Journal of Politics*, Vol 74, No. 1 (Jan 2012), 174-186; 176; Katherine Clayton, Spencer Blair, Jonathan Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholtz-Bright, Austin Welch, Andrew Wolff, Amanda Zhou, and Brendan Nyhan, "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media," *Political Behavior,* forthcoming.

experiment evaluated three differently-sized AI text generation models' ability to synthesize news stories at scale without human intervention.

We found first that readers perceived AI-generated text curated by a human editor to be as credible as an original human-written news article about the same event. Second, we concluded that while partisanship heavily conditioned beliefs about media credibility—indicating the potential viability of hyper-targeted synthetic misinformation—the AI-generated text nonetheless did little to change attitudes about contentious foreign policy issues. Finally, we observed that the AI text generation models were able to synthesize credible-sounding news articles at scale without any human intervention, but that average marginal improvements in perceived credibility of the synthetic text diminished as the power of the model increased.

The public's overall credulity suggests a propensity for manipulation. In this case, we found that people can be manipulated by AI-generated text such that they cannot discern real from synthetic content. The narrower consequence of a manipulable public is that misinformation campaigns have a ripe target. Malicious actors can easily produce AI-generated content and generate confusion about the truth, undermining trust in democratic institutions such as the media.[7] More generally, however, the ease of manipulation suggests avenues for misinformation not in service of political persuasion but instead in sowing confusion and distrust. Following Kant's belief in "a common sense as the necessary condition of the universal communicability of our knowledge," the erosion of common reference points has the potential to undermine the basis of coherent public policy.[8]

[7] Jen Weedon, William Nuland, and Alex Stamos, *Information Operations and Facebook,* 2017, 8.
[8] Andrew Norris, "Arendt, Kant, and the Politics of Common Sense," *Polity* Vol 29, No. 2 (Winter 1996), 165-191; 166.

Advances in AI-Generated Text

Until now, misinformation campaigns have been limited by human resources and bandwidth. Employees of the Internet Research Agency, a Russian company that engaged in online influence operations on behalf of Russian political interests, worked 12-hour shifts writing articles or social media posts about topics that the government assigned.[9] The demands are onerous because individuals must create new content by hand; posting recycled content would make detection by social media platforms and law enforcement agencies far more likely.

New technologies stand to ease that resource burden. Through advances in machine learning and artificial intelligence, language models are able to generate credible-sounding continuations to short prompts. The applications are already transforming journalism, alleviating the resource-constrained media environment by automating the task of writing local news stories, warnings about earthquakes, and earnings reports while shifting the thought-intensive human efforts instead to editing and curation.[10] Language models can also alleviate the task of generating misinformation: malicious actors looking to produce large quantities of misinformation that is not plagiarized nor easily filtered can input a headline or lede from an actual or fake news story, and generate an entirely original article in seconds.

Though there is no evidence that AI text generation models have been used to systematically synthesize politically motivated misinformation, Russia is already known to use automated social media bots to amplify pro-Russian content or anti-West content that seeks to exacerbate domestic division within Europe or the United States.[11] Adopting AI powered systems is a logical next step:

[9] Neil MacFarquhar, "Inside the Russian troll factory: Zombies and a breakneck pace," *New York Times*, 18 February 2018.
[10] Nicholas Diakopoulos, *Automating the News: How Algorithms are Rewriting the Media* (Harvard University Press, 2019).
[11] Todd Helmus et al., *Russian Social Media Influence* (Rand Corporation, 2018).

the Russian president has claimed that "artificial intelligence is the future," noting it as a tool to overcome power and resource asymmetries vis-à-vis countries in Europe and the United States that are able to spend more on defense.[12] AI-generated misinformation therefore represents a plausible future mechanism for the spread of false and misleading information.

Despite their potential utility as a tool of misinformation, research on the plausible misuse case for these language models—generating credible-sounding news text for political misinformation—has been minimal.[13] Previous studies instead have tended to focus on internal performance tasks, such as a language model's capacity for reading comprehension, translation, summarization, and answering questions, and shown that increases in the number of parameters improves task performance in a log-linear manner.[14] Our study is the first to analyze the potential effect of these models on the media landscape. We focused our empirical study on GPT-2,[15] one of the most powerful language models available. It generates coherent text given minimal inputs—whether a sentence or even just one word—and does so in the same genre of the input and without domain-specific training datasets. Below we describe the design, objectives, and results of three experiments that investigated whether readers can detect original versus synthetic text, the relationship between model power and perceived credibility of news outputs, and the impact of partisanship on perceptions of news credibility.[16]

[12] Alina Polyakova, "Weapons of the weak: Russia and AI-driven asymmetric warfare," *Brookings Institution*, 15 November 2018.
[13] One exception is the research of Zellers et al., which found that Grover, one particular language model, produces outputs that individuals view as trustworthy as human-generated text. Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi, "Defending against Neural Fake News," https://arxiv.org/pdf/1905.12616.pdf
[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language Models are Unsupervised Multitask Learners," https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[15] GPT-2 was created by the research group OpenAI.
[16] See the Appendix for a complete discussion of the three experimental designs and instruments.

Experiment 1: Comparing Upper-Bound Credibility

Our first study focused on three different versions of GPT-2—the 355 million (medium), 724 million (large), and 1.5 billion (extra-large) parameter models. The different versions have similar architectures and rely on the same learning rules and dataset. All models were trained on 40 gigabytes of outbound Reddit links with "karma scores" greater than two, selecting for content that is educational and interesting. The models vary in terms of the number of parameters: more parameters translate into enhanced capacity to capture complex relationships within and across texts. In short, as parameter size increases, so does the "intelligence" of the text generator and the sophistication of the output.

The first experiment tested the American public's perceived credibility of AI-generated text by comparing news output from the three models against human a human-written baseline article from the *New York Times* about a North Korean ship seizure. We chose the *Times* because it is the newspaper of record in the United States, and we selected the topic of North Korea because of its prominence as an ongoing American foreign policy concern. We then used the AI text generation tool to generate 20 different news stories per model. Seeking a best-case scenario for credibility and reflecting what a well-staffed interference campaign might do to maximize its impact, we manually selected the most credible stories based on three criteria: the presence of 1) grammatical or spelling errors (appropriate use of articles, complete sentences); 2) factual errors (correct titles and affiliations for named individuals); and 3) conceptual coherence (stays on topic versus deviates to unrelated topic).

In order to minimize the effect of topic variation on perceived credibility and isolate our assessment of the models to their ability to synthesize political misinformation, we used the

beginning of the single *Times* baseline story as the lede for the models to use in generating the body text. Therefore, all the stories concerned the same North Korean ship seizure; any variations in perceived credibility across stories reflect the models' text synthesis capabilities and not variations in topic. For the lowest-powered model, we tested only outputs generated by using the first two sentences of the *Times* story as inputs. We challenged the two larger models by generating outputs not just with the first two sentences of the *Times* baseline as a prompt, but also with just the first sentence of the baseline as a prompt (in which more inference on the part of the model is required). Each respondent read a randomized story, rated the credibility of the story on a four-point scale, and answered a number of demographic questions.[17]

Results: Experiment 1

Table 1 shows the results comparing the three AI models and the NYT baseline. In the 355M parameter model, most respondents viewed all of the texts as credible.[18] Even in the treatment that respondents thought was least credible, three-fifths of respondents deemed the article to be credible; in the treatment with the highest perceived credibility, nearly 72% of respondents deemed the synthesized article credible. Nonetheless, the *Times* baseline was statistically more credible than any of the outputs from the lowest powered model.[19] In contrast, the outputs from the 774M and 1.5B models were virtually indistinguishable in terms of perceived credibility relative to the baseline

[17] Participants were recruited via Amazon Turk. See Appendix for all details on recruitment, timing, and question ordering for each experiment.
[18] Refers to responses of "somewhat credible" or "very credible." See Appendix for more information about the survey instrument.
[19] Zellers et al. similarly found comparable levels of trustworthiness (100 articles of news and propaganda) for a different natural language model called Grover.

Times story, and one of the 774M treatments (that based on the one-sentence input) was statistically more credible than the baseline.

	355M			774M			1.5B		
Treatment	Mean (%)	95% CI	t	Mean	95% CI	t	Mean	95% CI	t
n		501			507			504	
New York Times (control)	83	77-89	--	71	62-80	--	76	68-84	--
2 sentence input (1)	58	49-67	4.5 ***	73	67-82	-0.3	75	67-84	0.1
2 sentence input (2)	72	63-80	2.2 **	65	56-75	0.9	71	62-80	0.8
2 sentence input (3)	69	61-77	2.8 ***						
1 sentence input (4)				84	77-91	-2.2 **	77	68-85	-0.1
1 sentence input (5)				75	67-84	-.7	70	61-79	1.0

Table 1. Difference in means between three GPT-2 models and either 2 or 1 sentence inputs (relative to baseline *New York Times* article on the question of whether the news source is "credible"). *=p<0.05, **=p<0.01, ***=p<0.001

Experiment 2: The Effect of Policy and Partisanship

If analyses of online misinformation are any indication, malicious actors tend to spread divisive content on "hot-button issues with racial undertones," thereby stoking social discord in the United States.[20] While our first study did not vary the topic of the treatment texts in order to isolate the analysis to the capabilities of AI text generation, our second study seeks to examine how individuals respond to targeted synthetic text from both congenial and non-congenial political

[20] Maggie Miller, "Senate Intel report," *The Hill*, 8 October 2019.

viewpoints, and whether they are equally disavowed or agitated by synthetic text that comes with a disclaimer about the media's veracity.

For this experiment, we selected one of the most politically salient and contentious issues in the United States,[21] immigration, and varied the political viewpoint and whether the story had a disclaimer. The articles focused on immigration "caravans" that had migrated from Central America to the United States border in 2018. To investigate how individuals respond to media that intentionally comports or conflicts with their ideological priors—as targeted misinformation might—we varied the ideological angle of the story. As a baseline, we selected a descriptive story about immigration caravans from the *Associated Press*. We then selected one story from *Fox News*, known as a credible source for Republicans and distrusted by 81% of liberals, and another from *The Huffington Post*, known to have a progressive bent and viewed as credible by liberals and overwhelmingly distrusted by conservatives.[22] The sources of the stories were not revealed to respondents.

We also varied the stated authenticity of the article. In recent years, various initiatives have attempted to alert readers to the potential of inauthentic media; we set out to test whether these disclaimers have any meaningful effect on the perceived credibility of the synthetic stories. For each original news story, we tested two treatments: in one treatment, we displayed an AI-generated story with no other information, and in the other we showed a disclaimer above the same story.[23] While including real human generated stories with disclaimers as a treatment would have been more internally valid, allowing us to draw clean inferences about the independent effect of disclaimer, we

[21] Frank Newport, "Immigration Surges to Top of Most Important Problem List," *Gallup,* 18 July 2018.
[22] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa, "Political Polarization and Media Habits," *Pew Research*, 21 October 2014.
[23] For more information about this disclaimer, see the Appendix.

opted instead for external validity: not including any disclaimers on the human-written stories in our treatments because it would have meant fabricating a disclaimer for genuine news (problematic from an IRB standpoint for misrepresenting something factual). We recognize that not including a disclaimer treatment for the original texts means we cannot directly compare the AI disclaimer treatments with the original texts; we can, however, still evaluate differences between the AI with disclaimer and AI without disclaimer treatments.

In pre-test questions, we asked participants about their political and ideological orientations so that the story on immigration would not prime their responses. Using a 3x3 experimental design with more than 1,500 respondents, we varied the political ideology of the story (left, right, and center) and whether the body text was the original human content, AI-generated, or AI-generated flagged with a disclaimer. The AI-generated stories used the first two sentences of the corresponding human-written articles as a prompt. Each AI-generated treatment text was selected from a group of 20 initial generations using the same criteria as the first experiment.

We measured attitudes about immigration, seeking to understand whether the synthetically-generated stories independently affected responses to questions whether immigration should be kept at its present level, increased, or decreased, as well as whether respondents favor or oppose the construction of walls along the US border.

Results: Experiment 2

Our analysis, summarized in figure 1, shows that partisans were more likely to find their politically-congenial story credible—indicating that hyper-targeting synthetic articles is a viable strategy to maximize credulity. Republicans found the *Fox News* story more credible (mean of 3.35 on 1-4 Likert scale, 95% CI=3.17-3.53) than the *Huffington Post* (mean 2.98, 95% CI=2.78-3.17) and

Democrats found the *Huffington Post* story (3.19, 95% CI=3.04-3.33) more credible than *Fox* (2.68, 95% CI=2.52-2.84) despite the stories being presented without any explicit attribution to those outlets.[24] Whereas previous research finds that individuals are drawn towards politically-congenial outlets (Republicans to *Fox* branded stories, Democrats to *NPR*),[25] we show that even without brand associations, partisans favor such outlets.

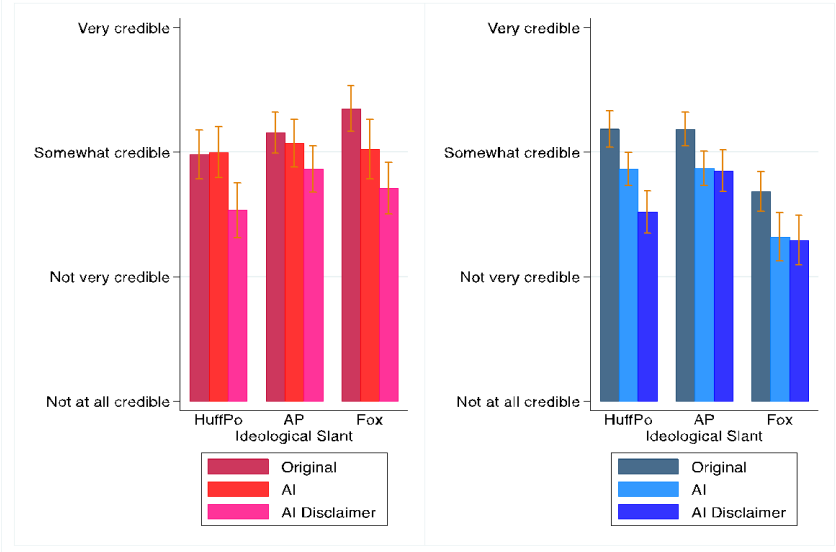

Figure 1. Perceived credibility of news stories by party identification with 95% CI. Left panel shows Republican respondents; right panel shows Democratic respondents. See Appendix for additional demographic analysis.

Next, we evaluated the effect of partisanship on whether individuals are susceptible to directionally motivated reasoning, in this case whether partisans are likely to persevere in their beliefs about credibility or particular policy views when exposed either to disclaimers or opposing ideological perspectives.[26] Figure 1 illustrates declines in perceived credibility for the AI-disclaimer compared to the AI-generated story on the basis of partisanship. For Democrats, while the decline

[24] Differences in mean perceived credibility for Republicans *Fox* vs *Huffington Post*, p=.007; Democrats' p=0.000.
[25] Shanto Iyengar and Kyu Hahn, "Red media, blue media: evidence of ideological sensitivity in media use," *Journal of Communication*, 59 (2009): 19-39.
[26] DJ Flynn, Brendan Nyhan, and Jason Reifler, "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics," *Advances in Political Psychology*, Vol 38, Suppl 1 (2017), 127-150.

in perceived credibility was statistically significant for the congenial story, from a mean of 2.86 on a 4 point scale (without disclaimer; 95% CI=2.73-2.99) to 2.51 (same story with disclaimer; 95% CI=2.35-2.68),[27] the disclaimer had no effect on the perceived credibility of the AI-generated, politically non-congenial story—possibly the result of a floor effect given the already low levels of perceived credibility for the AI-generated story that used the *Fox* article as its input.[28]

Republicans, in contrast, were receptive to the AI disclaimer for the politically non-congenial story, with their perceived credibility decreasing from 3.00 (95% CI=2.79-3.2) to 2.53 (95% CI=2.31-2.75).[29] They were statistically unmoved by the AI disclaimer for the politically congenial *Fox* story, as the left panel of figure 1 illustrates (from 3.02, 95% CI=2.78-3.26; to 2.71, 95% CI=2.5-2.92).[30] These heterogeneous effects are consistent with other studies that have found a resistance among Republicans to corrections of politically-congenial fake news stories.[31]

Finally, we assessed how the partisan orientation of the story interacted with reader's own partisan views and the effect on attitudes about immigration.[32] As figure 2 suggests, exposure to AI-generated news on immigration, regardless of whether the angle was politically congenial, did not change attitudes toward immigration for either Republicans nor Democrats. In terms of attitudes toward immigration, the disclaimer also had no effect relative to the AI-generated story it corrected. These results suggest either that beliefs are sufficiently entrenched that individuals experience inoculation to new frames,[33] or relatedly, that the information is further reinforcing individuals'

[27] Differences in mean perceived credibility for Democrats, AI-disclaimer-*HP* versus AI-*HP*, p=0.002.
[28] Differences in mean perceived credibility for Democrats AI-disclaimer-*Fox* vs AI-*Fox*, p=0.836.
[29] Difference in mean perceived credibility for Republicans AI-disclaimer-*HP* vs AI-*HP*, p=0.002.
[30] Difference in mean perceived credibility for Republicans between AI-disclaimer-*Fox* vs AI-*Fox*, p=0.053.
[31] Clayton et al. 2018; R Kelly Garret, Jacob Long, and Min Seon Jeong, "New evidence on group polarization from partisan media to misperception: Affective polarization as mediator," *Journal of Communication*, 69, 5 (Oct 2019), 490-517.
[32] We show support for building a wall in the Appendix because the results are similar.
[33] Dennis Chong and James Druckman, "Counterframing Effects," *The Journal of Politics*, Vol 75, No. 1 (Jan 2013), 1-16; 3.

priors. Growing polarization and intensification of attitudes, or perhaps further study of the issue, may explain the muted effects of exposure to politically non-congenial positions compared to previous studies.[34]

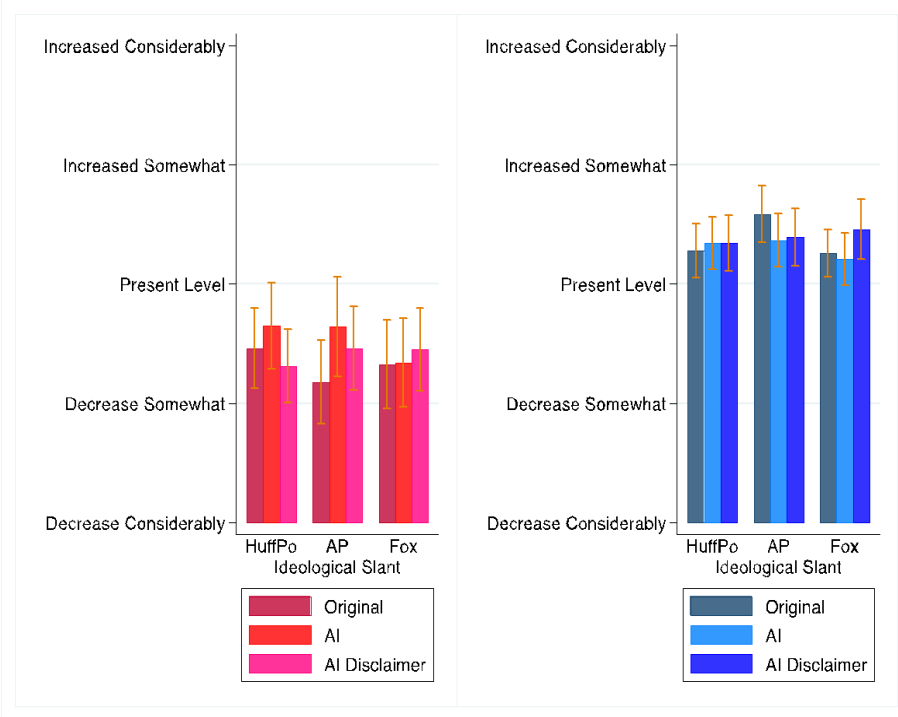

Figure 2. Attitudes Toward Immigration by Party Identification. Left panel shows Republican respondents, and right panel shows Democratic respondents. Y-axis scale reflects whether immigration should be increased, maintained, or decreased.

Experiment 3: Testing Greater Automation of Synthetic Text Generation

In the first two studies, we generated dozens of outputs and tested the perceived credibility of those with the fewest grammatical or spelling errors, factual errors, and the most conceptual coherence in order to assess the upper bound of perceived credibility. While it is certainly possible for a hypothetical misinformation campaign to cherry pick several credible-sounding stories of even the least powerful model, shifting its human resources from article generation to merely curation, we

[34] Arceneaux et al. "Polarized Political Communication," 183-184.

also sought to assess whether the models could produce misinformation without any human in the loop. In our third study we therefore focused not just on the upper tails of each model's credibility distribution, but on the entire distribution. Using the same *New York Times* story as the first experiment, we generated 300 outputs each for the 355M, 774M, and 1.5B models, all based on a one-sentence input. We developed and deployed a custom computer program to automatically perform minimal text cleaning to remove advertising and ensure consistent formatting across articles.[35] A total of 600 respondents (200 per model)[36] read a randomized AI-generated output, then reported whether the story was credible. Thus, no two respondents read the same story.

In this experiment, we also unpacked the idea of "credibility." Credibility, even if it logically reduces to believability,[37] is not a unidimensional concept. Understanding what people consider when they respond that a story is more or less credible can shed light on the features of the AI-generated story that are persuasive to readers and therefore likely to correspond to more successful misinformation campaigns. Leveraging studies that decompose the potential features of credibility, we asked whether the story was believable, accurate, and clear.[38] Respondents scored each of these on a four-point scale, which we then aggregated into an indexed credibility score.[39] To assess the reliability of the measure, we calculated Cronbach's alpha for the credibility assessments of each

[35] This program is detailed in the Appendix.
[36] As with the previous experiments, the details on sample, recruitment, and timing are available in the Appendix.
[37] Philip Meyer, "Defining and Measuring Credibility of Newspapers: Developing an Index," *Journalism Quarterly*, Vol 65, No. 3 (1988), 573.
[38] We exclude depth or completeness because we select a fairly narrow foreign policy issue of North Korea within one type of media source, a hypothetical newspaper, whereas the depth or thoroughness measures are geared toward comparison of different media types, for example newspapers, magazines, and candidate literature, which are qualitatively different in terms of their theoretical and actual depth measures (Johnson and Kaye 2000, 328-329).
[39] Andrew Flanagin and Miriam Metzger, "Perceptions of Internet Information Credibility," *Journalism and Mass Communication Quarterly*, Vol 77, No. 3 (Autumn 2000), 515-540; 522.

model and found levels greater than 0.85 for each, suggesting reasonable degrees of internal consistency.[40]

Results: Experiment 3

Figure 3 below plots the credibility distribution of the three models—355M, 774M, and 1.5B—comparing the distributions with each other rather than to the original *Times* story as in experiment one. The distributions vary little in terms of their means. For the smallest model, the mean for the credibility index was 6.65 (95% CI=6.37-6.93), compared to 6.72 (95% CI=6.45-6.99) for the 724M model and 6.93 (95% CI=6.67-7.19) for the 1.5B model. However, we then conducted a Kolmogorov-Smirnov test to compare the cumulative distributions between different model sizes. Comparing the frequency distributions, we found evidence that the smallest model was from populations with different distributions from either of the two larger models.[41] The experiment reinforces evidence from experiment one that the best outputs of the 774M and 1.5B models are perceived to be more credible than that of the 355M model, the credibility means credibility of the full distributions are not statistically distinguishable (although might be with a larger sample).

[40] Cronbach's alpha calculations included in the Appendix.
[41] Kolmogorov-Smirnov test information included in the Appendix.

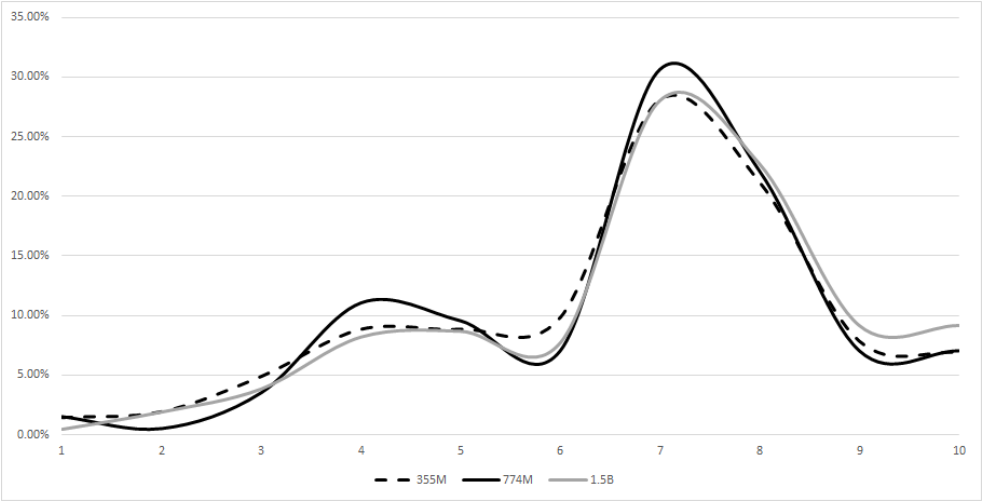

Figure 3. Credibility distribution (index of 1-10) for the 355M, 774M, and 1.5B GPT-2 model sizes. Credibility index based on a 4-point scale of whether the story was 1) believable, 2) accurate, and 3) clear, then aggregated to a 3-12 scale and rescaled to range between 1 and 10. The y-axis is the percentage of respondents (n= ~200 per model) whose credibility index registered at that 1-10 level.

The evidence suggests diminished marginal increases in performance, defined as perceived credibility of text, as the model size increases. Our analysis suggests that the development of new AI-based natural language programs, such as Facebook's transformer-based Blender with 9.4 billion parameters,[42] Google's T5 (with 11 billion parameters),[43] and OpenAI's newer GPT-3 (with 175 billion parameters)[44] are therefore unlikely to be many orders of magnitude more capable than the 1.5 billion parameter model tested here.[45]

Conclusion

[42] Kyle Wiggers, "Facebook open-sources Blender, a chatbot people say 'feels more human,'" *Venture*, 29 April 2020.
[43] Adam Roberts and Colin Raffel, "Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer," *Google AI Blog,* 24 February 2020.
[44] Tom Brown et al., "Language Models are Few-Shot Learners,"
[45] Importantly, number of parameters is not the *only* determinant of model quality and power; underlying architecture improvements and training data also strongly affect skill at a given task. However, model parameter size nonetheless roughly correlates to power.

While governments have long practiced misinformation and propaganda, the danger of new AI-based tools is scale and velocity: the ability to produce large volumes of credible-sounding misinformation quickly, then to leverage networks to distribute it expeditiously online.[46] Our analysis shows that these tools are capable of producing news content that readers deem as equally or more credible than human-written news stories. The potential consequences will be more noise in a news environment already characterized by growing volumes of misinformation, lower levels of trust in the media,[47] and citizen disengagement with the political landscape if they cannot discern fact from fiction.[48] While there is no direct evidence that foreign or domestic adversaries have begun employing AI text generation tools in service of misinformation, post-2016 legislative investigation of the 2016 election suggests that Russia has sought ways to make the generation of misinformation more efficient while still appearing credible—the hallmark of these tools. As this analysis shows, lower-powered earlier versions of the technology did not meet these criteria but the more powerful recent versions do.

Our findings also corroborate earlier studies suggesting that disclaimers are not consistently effective.[49] However, evidence from our analysis suggests that digital media literacy interventions may be effective in educating individuals how to discern between human and synthetically-generated text. Respondents in our experiments frequently identified markers of synthetic text, whether contradictions in the story, grammatical or factual errors, while still overwhelmingly reporting that the synthetic stories sound authentic and believable. Given the frequency with which respondents in our experiments identified dubious aspects of the story, even if they nonetheless thought the story

[46] Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, Vol 359, No. 6380; 9 March 2018; 1146-1151.
[47] Megan Brenan, "Americans' Trust in Mass Media Edges Down to 41%," *Gallup*, 26 September 2019.
[48] Sabrina Tavernise and Aidan Gardiner, "'No One Believes Anything': Voters Worn Out by a Fog of Political News," *New York Times,* 18 November 2019.
[49] Clayton et al., 2019.

was credible, the prospects for effective education interventions that help individuals understand the markers of synthetic text are promising.[50]

A different avenue for detecting synthetic text may be technology itself. One study suggested that "the best way to detect neural fake news is to use a model that is also a generator. The generator is most familiar with its own habits, quirks, and traits" and therefore "can easily spot its own generated fake news articles, as well as those generated by other AIs."[51] Platforms analyzing metadata associated with posts—including origin IP addresses, the timing and frequency of new activity, and the social graph of different accounts—may also have potential.[52]

While we have taken an important first step in analyzing the credibility of synthetic text tools, we urge future research in a number of directions. Are individuals' policy attitudes resistant to politically non-congenial viewpoints because the foreign policy issue of North Korea is insufficiently salient, or because attitudes on immigration are fairly entrenched? We would suggest research on other issues that might push the question of generalizability further. Additionally, how does the credibility of synthetic text compare to human-created fake news? Many of the viral stories from the 2016 election cycle were from fake news outlets but authored by humans.[53] Are AI-generated stories perceived as more or less credible than these human-generated stories posted on for-profit websites? How does the propensity to share the story vary? How does partisanship condition the interpretation of each? We leave these questions for future research.

[50] Guess et al. 2019; Pennycook and Rand 2018.
[51] "Grover-A State-of-the-Art Defense against Neural Fake News," https://grover.allenai.org/detect
[52] Irene Solaiman et al., "Release strategies and the social impacts of language models," 2019, https://arxiv.org/pdf/1908.09203.pdf.
[53] Brendan Nyhan, "Why Fears of Fake News are Overhyped," *Medium*, 4 February 2019.