Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
September 2024

How Persuasive Is AI-Generated Propaganda?

Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz

MAJOR BREAKTHROUGHS IN AI TECHNOLOGIES, ESPECIALLY LARGE LANGUAGE MODELS (LLMS), have prompted concerns that these tools could enable the mass production of propaganda at low cost. Machine learning models that generate original text based on user prompts are increasingly powerful and accessible, causing many to worry that they could supercharge already frequent and ongoing online covert propaganda and other information campaigns. Indeed, companies and external researchers have already begun uncovering covert propaganda campaigns that are using AI.

Research into the risk of AI-generated propaganda is emerging: Scholars have examined if people find AI-generated news articles credible, if people recognize when AI-generated content is false, and whether elected officials reply to AI-written constituent letters. To date, however, no studies have examined the persuasiveness of AI-generated propaganda against a real-world benchmark.

Our paper, "How Persuasive Is AI-Generated Propaganda?," addresses this gap. We conducted an experiment with U.S. respondents to compare the persuasiveness of foreign propaganda articles written by humans

Key Takeaways

Major breakthroughs in large language models have catalyzed concerns about nation-states using these tools to create convincing propaganda—but little research has tested the persuasiveness of AI-generated propaganda compared to real-world propaganda.

We conducted a preregistered survey experiment of U.S. respondents to measure how persuasive participants find six English-language foreign propaganda articles sourced from covert campaigns compared to articles on the same six topics generated by OpenAI's GPT-3 model.

GPT-3-generated articles were highly persuasive and nearly as compelling as real-world propaganda. With human-machine teaming, including editing the prompts fed to the model and curating GPT-3 output, AI-generated articles were, on average, just as persuasive or even more persuasive than the real-world propaganda articles.

Policymakers, researchers, civil society organizations, and social media platforms must recognize the risks of LLMs that enable the creation of highly persuasive propaganda at significantly lower cost and with limited effort. More research is needed to investigate the persuasiveness of newer LLMs and to explore potential risk mitigation measures.

and sourced from real-world influence campaigns against articles generated by OpenAI's LLM GPT-3. We sought to answer a single question: Could foreign actors use AI to generate persuasive propaganda? In short, we found that the answer is yes.

As the machine learning community continues to make breakthroughs, and policy debates about AI-generated disinformation intensify, it is essential to ground policy discussions in empirical research about risks posed by AI systems.

Could foreign actors use AI to generate persuasive propaganda? In short, we found that the answer is yes.

Introduction

Security experts, civil society groups, government officials, and AI and social media companies have all warned that generative AI capabilities, including LLMs, could enhance propaganda and disinformation risks to democracies. The White House's 2023 executive order on AI warns that irresponsible use of AI could exacerbate societal harms from disinformation. Yet, the examples cited are often anecdotal, and little research exists to empirically measure the persuasiveness of AI-written propaganda.

Our experiment aims to fill this research gap. Using the survey company Lucid, we interviewed a sample of 8,221 geographically and demographically representative U.S. respondents to find out how persuasive they find real-world foreign covert propaganda articles compared to AI-generated propaganda articles.

We first needed to assemble a set of real-world propaganda articles. To do so, we selected six English-language articles that were previously found to be part of covert, likely state-aligned propaganda campaigns originating from Russia or Iran.

We then created AI-generated versions of the human-written propaganda articles by using the "few-shot prompting" capability of GPT-3, which allows you to provide the model with examples of the output you want. We fed GPT-3 davinci three unrelated propaganda articles to inform the style and structure of the desired output. We also provided one or two sentences from the original article that contained the article's main point to inform the substance of the GPT-3-generated propaganda. Based on the prompt, the model returned a title and article. To avoid over-indexing on any one output, we used GPT-3 to generate three title-article pairs for each topic since each AI-generated article is different. We discarded AI-generated articles that were not within 10 percent of the shortest and longest human-written articles to make sure the lengths of the human-written and AI-generated sets were comparable.

With the propaganda articles in hand, we sought to measure persuasiveness of the human-written and AI-

generated propaganda. First, we summarized the main point in each of the six original propaganda articles, several of which are false or debatable:

1. Most U.S. drone strikes in the Middle East have targeted civilians rather than terrorists.
2. U.S. sanctions against Iran or Russia have helped the United States control businesses and governments in Europe.
3. To justify its attack on an air base in Syria, the United States created fake reports saying that the Syrian government had used chemical weapons.
4. Western sanctions have led to a shortage of medical supplies in Syria.
5. The United States conducted attacks in Syria to gain control of an oil-rich region.
6. Saudi Arabia committed to help fund the U.S.-Mexico border wall.

Next, we collected the control data by asking each respondent how much they agreed or disagreed with four of these thesis statements, selected at random, without having read articles. Finally, we collected the treatment data by showing respondents an AI- or human-written propaganda article on the remaining two topics and measuring their agreement with the relevant thesis statements.

For both the control and treatment cases, we measured agreement in two ways: "percent agreement" and "scaled agreement," where percent agreement is the percentage of respondents who agreed or strongly agreed with each thesis statement and scaled agreement is the average score on a 5-point scale from 0 ("strongly disagree") to 100 ("strongly agree"). When averaging scores across issues and across GPT-3-written articles, we weighed each issue and article equally.

AI-generated propaganda is likely to be compelling to a wide range of groups in society.

Research Outcomes

Our survey outcomes show that AI-generated propaganda is similarly persuasive as real-world propaganda: Respondents' agreement with thesis statements increased substantially after they read both types of propaganda articles.

The original propaganda articles were highly persuasive. Only 24.4 percent of respondents agreed or strongly agreed with the thesis statement when they hadn't read an article (the control); but when they read a real-world propaganda article, that number jumped to 47.4 percent (a 23 percentage point increase). The real-world propaganda nearly doubled the share of people who agreed with the thesis statement, though specific numbers varied slightly across each topic, from drones to sanctions to chemical weapons in Syria.

GPT-3-written propaganda articles were also highly persuasive. After reading a GPT-3-generated

propaganda article, 43.5 percent of respondents agreed or strongly agreed with the thesis statement (a 19.1 percentage point increase from the control). This means the GPT-3-written articles were slightly less persuasive than the original propaganda. However, the gap was due to a small number of generated articles (2 of 18) that missed their mark by not advancing the intended argument. When we removed those two off-topic articles, the difference in persuasiveness between the human-written and AI-written content became statistically insignificant.

Our findings therefore suggest that propagandists could utilize GPT-3 to generate persuasive articles with minimal human effort. We also studied what would happen if humans played a slightly greater role by selecting the most persuasive of the three articles on each topic and/or editing the prompts given to GPT-3 to create more compelling content. When we simulated these human-machine teaming strategies, the GPT-3-generated propaganda became just as persuasive or even more persuasive than the original propaganda.

The effectiveness of both the original propaganda and the GPT-3-written propaganda was relatively consistent across groups of people. We observed no significant differences when dividing the sample according to demographic variables, partisanship/ideology, news consumption, time spent on social media, and other factors, leading us to conclude that AI-generated propaganda is likely to be compelling to a wide range of groups in society.

GPT-4 and other models in the pipeline could be used to produce propaganda at least as persuasive as the text we generated using GPT-3.

Policy Discussion

Policymakers, researchers, civil society organizations, social media platforms, and other stakeholders must recognize the potential risks of misusing LLMs and the urgent need for more research on the persuasiveness of AI systems.

Our research indicates that AI-generated content is highly persuasive for U.S. audiences and that when propagandists oversee and adjust parts of the AI content generation process, they can produce articles that are as or even more persuasive than real-world, human-written propaganda. Nation-states and other actors could leverage LLMs to create highly persuasive propaganda that maintains persuasiveness across demographic lines, geographical areas, and political ideologies. While LLMs could allow actors to automate many writing tasks, human-machine strategies that keep a human in the loop could enable the production of particularly effective propaganda.

It is crucial to note that our experiment likely underestimates the persuasive potential of LLMs. These models are rapidly improving—since our study, several companies have released larger and more powerful models, including OpenAI's GPT-4 and GPT-4o, that outperform GPT-3 davinci in related tasks. GPT-4 and other models in the pipeline could be used to produce propaganda that is at least as persuasive as the text we generated using GPT-3.

The other reason our experiment likely underestimates persuasiveness is that we studied the effect of reading only a single article, when propagandists could theoretically use AI to create many articles at once. By lowering the cost and improving the ease with which propaganda can be produced, LLMs open the door to actors generating a vast number of articles that convey a single narrative with variances in style and wording. Such stylistic variations could make it even more difficult to trace multiple propaganda articles back to the same source, because they read more like the views of real people or genuine news sources. Propagandists could redirect the time and resources they save by using AI to building an infrastructure that looks credible and evades detection, such as fake accounts or "news" websites that mask state links.

Future research should investigate the persuasive capabilities of newer LLMs, interactive dialogue systems, and audio models; compare them with a wider range of benchmarks, including the persuasiveness of expert human writers; and assess how the effectiveness of AI propaganda varies across topics. Research should also explore strategies to minimize the impact of AI-generated propaganda.

By lowering the cost and improving the ease with which propaganda can be produced, LLMs open the door to actors generating a vast number of articles that convey a single narrative with variances in style and wording.

For all the attention paid to Russia's covert propaganda campaigns in connection with the 2016 presidential election, the rapid advancement of LLMs deserves renewed attention for its potential to heighten known risks while raising new questions about the proliferation and persuasive power of online propaganda. Evidence-based research will be critical to building a deeper understanding of the effects of AI-generated propaganda, so that policymakers can work with other stakeholders to assess effective intervention strategies.

Reference: The original article is accessible at Josh A. Goldstein et al., "**How Persuasive Is AI-Generated Propaganda?**" *PNAS Nexus*, vol. 3, issue 2, February 2024, https://academic.oup.com/pnasnexus/article/3/2/pgae034/7610937.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Josh A. Goldstein is a research fellow at Georgetown's Center for Security and Emerging Technology (CSET).

Jason Chao is a master's student in computer science at Stanford University and was a researcher at the Stanford Internet Observatory.

Shelby Grossman is a research scholar at the Stanford Internet Observatory.

Alex Stamos is a lecturer at Stanford University and the former director of the Stanford Internet Observatory.

Michael Tomz is the William Bennett Munro Professor in Political Science and chair of the Department of Political Science at Stanford University and a senior fellow at the Stanford Institute for Economic Policy Research.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**