CSSPP Center for Science, Society, and Public Policy

Generative AI and the Future of Elections

R. Michael Alvarez, Frederick Eberhardt, and Mitchell Linegar
California Institute of Technology

July 21, 2023

POLICY BRIEF

Caltech Center for Science, Society, and Public Policy (CSSPP)

Generative AI and the Future of Elections

R. Michael Alvarez, Frederick Eberhardt, and Mitchell Linegar
California Institute of Technology
Center for Science, Society, and Public Policy*

July 21, 2023

Executive Summary

The rapid proliferation of Large Language Models and Generative AI introduces important new risks for the development and dissemination of misinformation in the 2024 elections. For example, we may see:

1. Highly realistic and hyper-targeted misinformation could mislead registered voters about where and when they can obtain and cast their ballots. Targeted campaigns could be developed to mislead voters about identification requirements, or how they can return vote-by-mail or absentee ballots.

2. The images and voices of popular celebrities or politicians could be used to mislead, confuse, or turnoff voters. These materials could be micro-targeted to selected voters in specific areas, perhaps spreading allegations of election fraud, in order to suppress turnout by sowing distrust.

3. Highly negative, misleading, and inflammatory negative materials could be generated about particular candidates (using their images and voices), and disseminated to persuadable voters in close elections.

LLMs have made it very easy for anyone to generate realistic content, in particular about politics and elections. Campaigns are now using LLMs and AI to produce campaign ads. Content can be generated quickly, and altered so quickly, so quickly that we fear it may be impossible for misinformation to be detected and mitigated using current approaches. Thus, we argue in this white paper that the risks of misleading AI-based misinformation campaigns in the 2024 elections is very high.

*https://lindeinstitute.caltech.edu/research/csspp

1 LLMs and Misinformation Risks in 2024

The past two decades have witnessed an immense transformation in political messaging, driven primarily by advancements in data science and digital technology. This revolution was initially propelled by the Obama campaign's use of data-driven micro-targeting, marking the beginning of a new era where personalization became a critical aspect of political messaging.[1] Complementing this shift has been an explosion in digital ad spending and the emergence of direct outreach to voters via texts and emails. Additionally, social media platforms have become pivotal stages for political conversation.

The rise of Large Language Models (LLMs) represents the next phase in this technological progression. Equipped with the capability to produce expansive volumes of text mirroring human-like writing, these AI models stand to further optimize the domain of political messaging. The swift rise of OpenAI's ChatGPT, demonstrated by its acquisition of 100 million monthly users merely two months post-release, reflects the potential these AI tools have to reshape the landscape of political communication and capture the public's imagination. As we continue to develop these technologies, it's crucial to understand their potential short-term and long-term implications for our political and social institutions, particularly with the 2024 U.S. presidential election already underway.

Given that unregulated LLMs have dramatically reduced the costs of generating highly realistic content generation, and since they can be used to develop hyper-targeted campaign communications, in this white paper we argue that LLMs pose significant risks for misinformation campaigns in the 2024 U.S. elections.

[1] Issenberg 2012.

2 Large Language Models and the Evolution of Digital Political Campaigning

The dawn of the digital age ushered in significant changes in political campaigning. This shift involves the emergence of data-driven micro-targeting, a strategy using large datasets to customize and personalize political messages for specific audiences.[2] Micro-targeting relies on detailed individual demographic and psychographic data, which prompted a surge in digital ad spending as political entities they could reach individual voters in the digital space.

Over the past twenty years, we've seen a major transformation in political messaging, largely driven by the advent of digital technology. The Obama campaign's innovative use of data-driven micro-targeting marked the beginning of an era where personalized messages became a key aspect of political communication. This change has been accompanied by a marked increase in digital ad spending and the rise of direct voter outreach through texts and emails. Additionally, social media platforms have become essential venues for political dialogue, providing a space where political ideas are shaped and shared.

In tandem with this trend, there was a pronounced shift towards direct voter outreach via digital channels such as emails and text messages, thereby revolutionizing the modalities through which political entities interacted with constituents. Concurrently, social media platforms ascended to prominence, becoming indispensable instruments for political communication. They emerged as global arenas where political narratives could be shared, debates held, and public sentiment molded. These platforms provided politicians, campaigners, and citizens with potent tools for engagement and discourse, dramatically reshaping the landscape of political communication.

Today, we are on the brink of yet another transformation. The emergence of LLMs heralds a new epoch, bringing tools that can augment and enhance the strategies developed over the past two decades. LLMs, such as OpenAI's GPT models, possess the capability to generate vast volumes of text. They can tailor this content to specific audiences, thereby providing powerful mechanisms to further refine micro-targeting and direct voter outreach.

With the advent of LLMs, we are encountering several unique characteristics that distinguish them from traditional digital strategies. These include the ability to generate content that is highly personalized to individual voters, a degree of subtlety that poses challenges for traditional monitoring methods, a comprehensive reach owing to their digital nature, and a heightened persuasiveness owing to their human-like text generation. Notably, these models can produce tailored content at a velocity and scale that significantly surpasses human capabilities.

Navigating this evolving digital landscape necessitates careful consideration of both the potential benefits and the associated risks of these advancements. Recently, the Caltech Center for Science, Society, and Public Policy hosted a discussion of these risks,

[2] Wylie 2019.

"Shaping the Future: The Societal Implications of Generative AI."[3] This white paper summarizes what we learned about these risks for Election 2024.

On one hand, these technologies hold the promise of fostering democratic dialogue and engaging voters more effectively. On the other hand, they raise the risk of amplifying misinformation and further polarizing political discourse. It is, therefore, paramount that we delve into these prospective long-term impacts, aiming to harness the advantages while mitigating the negatives, as we chart our course into the future of political campaigning.

3 Current Approaches to Misinformation

Efforts to combat misinformation in the digital realm adopt various strategies with varying levels of success. A few common tactics employed by platforms include banning users or groups, removing content, reducing its visibility, or appending content warnings. The ever-evolving landscape of technology, especially with the rise of LLMs, often undermines these approaches.

Banning individual users shows limited success, owing to the relative ease with which new accounts can be created, potentially even utilizing LLMs.[4] The banning of groups might be more effective, but is still restricted by the fluid nature of online communities.

The typical process that is used by social media platforms to enforce their policies regarding the use of their platforms to spread falsehood, misinformation, and deceptive content, usually have two different components. One component regards the detection of the problematic content. The detection step typically uses some combination of algorithms, which act as an initial filter, and human moderators. The detection process typically uses one of four processes, involving combinations of algorithms, humans, and community detection:

1. A moderator evaluates and decides on the content directly.
2. An automated process screens content before a moderator makes the final decision.
3. Purely algorithmic moderation that blocks only the most egregious and obvious offenders, such as posts caught by keyword filters or apparent spam.
4. The community reports suspicious content, which then undergoes automated screening, followed by human moderation.

Each of these formats has its own set of challenges. The first method often proves too slow due to the sheer volume and pace of social media content. The second and third methods open the door to strategic classification problems. The fourth method is

[3] For more about this event, see https://www.caltech.edu/about/news/generative-ai-regulation-kevin-roose.

[4] See for example, Sanderson et al. 2021.

reactive and often allows for some degree of harm before enough community reports trigger moderation. Misinformation disseminated at scale can cause significant damage before it is flagged and moderated.

A significant challenge with content moderation is the strategic classification problem, as discussed by Brückner and Scheffer 2011. This paradigm involves a decision-making entity (a classifier) and the entities subjected to its decisions. As these classified entities dynamically adjust their behavior in response to the classifier's strategies, thus reducing the effectiveness of the classifier. This is consistent with other theories such as Goodhart's Law and the Lucas Critique, indicating that any measure or model's effectiveness diminishes when it becomes the target of strategic adaptation (Goodhart 1984; Lucas 1976).

The adversaries developing and distributing misinformation may be well-resourced and highly strategic, working to mask their identities and hide the evidence of their activities. While there has been some research on methods to detect problematic online behavior in the context of the strategic classification problem (see for example, recent Caltech research, Srikanth et al. 2021), sustained investment in the maintenance, refinement, and evolution of classifiers is paramount. Mechanisms for funding such investments could include the implementation of a usage tax on technologies that utilize these classifiers.

Once problematic content has been discovered, the next step regards how the platform will deal with the the content, in order to mitigate or prevent further spread of this content on their platform. Current mitigation and prevention strategies primarily focus on:

1. Countering individual fake claims or news stories via fact-checking or tagging content as "disputed" or "rated false." However, the impact of these efforts is often short-lived[5] and can be diminished by counterarguments in the form of opinion pieces[6].

2. Issuing general warnings about potential misinformation, which while fostering skepticism, can also decrease the perceived accuracy of legitimate news.[7]

3. Banning pages or groups that enable the spread of misinformation. Although this slows the spread of misinformation, it does not necessarily contain it.[8]

A hands-off approach to misinformation, analogous to achieving "herd immunity," has been proposed by some, believing that competitive forces among misinformation purveyors will self-regulate its impact. This reasoning assumes that consumers have a certain preference for "fake news," and saturation will eventually lead to diminishing effects. However, when fake and real news compete directly for attention, fake news, by virtue of its sensationalist, attention-grabbing nature, generally holds an edge over real news on an article-to-article basis. As such, an entirely laissez-faire approach will likely amplify the reach and impact of misinformation. Further research into these dynamics,

[5] Carey et al. 2022

[6] Nyhan, Porter, and Wood 2022

[7] Clayton et al. 2020

[8] Thèro and Vincent 2022

including the role of algorithms in shaping consumption patterns, is needed to effectively counter misinformation.

4 Large Language Models and the 2024 Election

What are some of the harms that misinformation fueled by LLMs can produce, especially for the upcoming American federal elections in 2024?

1. Highly realistic and hyper-targeted misinformation could mislead registered voters about where and when they can obtain and cast their ballots. Targeted campaigns could be developed to mislead voters about identification requirements, or how they can return vote-by-mail or absentee ballots.

2. The images and voices of popular celebrities or politicians could be used to mislead, confuse, or turnoff voters. These materials could be micro-targeted to selected voters in specific areas, perhaps spreading allegations of election fraud, in order to suppress turnout by sowing distrust.

3. Highly negative, misleading, and inflammatory negative materials could be generated about particular candidates, and disseminated to persuadable voters in close elections.

These examples help to underscore the risks that unregulated LLMs pose for the 2024 federal elections in the United States. LLMs have made it very easy for anyone to generate realistic content, in particular about politics and elections. We are already seeing campaigns use LLMs and AI to produce campaign ads.[9] Content can be generated quickly, and altered quickly, so quickly that we fear that the strategic classification problem may make it impossible for misinformation to be detected and mitigated using current approaches.[10] Thus, we believe that the risks that misleading campaigns will be mounted in the 2024 elections is very high.

5 Regulating Large Language Models

Addressing the challenges associated with LLMs presents two primary paths of legislative intervention: a) restricting LLMs themselves, or b) mitigating the harms they can potentially generate. Each path, however, harbors unique complications and consequences.

[9] See for example the recent Ai-generated digital ad, https://youtu.be/kLMMxgtxQ1Y.

[10] As an exercise, we used free or inexpensive off-the-shelf generative AI tools to make two fake news spots. These each took minimal time and resources, and are examples of the type of content that can easily be developed and which could be quite realistic if posed to social media. See https://drive.google.com/file/d/1FLrrn91371j2NaFVvFzZjQh67aPxbdMZ/view?usp=sharing and https://drive.google.com/file/d/1QIWRl9jb2EDNbl_Sy8olDJe_f0461IQg/view?usp=sharing.

5.1 Restricting Large Language Models

An immediate response might involve censoring the ″foundational″ models, i.e., the core trained parameters, such as those found in the publicly released LLaMA or proprietary models like GPT-4. However, this strategy might prove ineffective and yield undesirable outcomes.

The abundance of open-source models and data —- unequivocally a boon for the scientific and technological community —- implies that restrictions placed on foundational models would be short-lived at best. Given the public accessibility of code and data, it is inevitable that, given time or resources, entities would train uncensored versions of the models.

Beyond training new models, entities could ″prompt engineer″ their way around restrictions. A telling example follows the release of ChatGPT, which sparked a wave of internet discussions on how to manipulate the model into assuming a new persona, dubbed ″DAN″ (an acronym for ″Do Anything Now″). Consequently, the ostensibly sanitized model could be prompted to generate explicit, offensive, or otherwise harmful content.

5.2 Regulating Firm Behavior

An alternative path involves targeting firms that disseminate content directly to consumers, such as social media and traditional media companies. Content, being directly observable, is easier to regulate than the complex intricacies of model training. Moreover, this approach more directly links the intervention to the potential harm.

In this context, the problem isn't necessarily the generation of misleading content by LLMs; rather, the issue lies with misleading content itself. As we have acknowledged, these challenges predate LLMs, and an exclusive focus on LLM-generated content won't solve the broader problem of misinformation, but may introduce additional issues.

Legislation holding firms accountable for downstream harms resulting from their models might seem appealing initially. Still, it will likely prove infeasible due to the difficulty in attributing a particular piece of content to a specific model, especially given the close similarity between LLM and human-generated content. Such legislation may result in these firms choosing not to release their trained models, effectively creating a ″digital moat″ around LLMs. However, this would not prevent motivated entities from training uncensored versions of the models. Instead, it would force the general public to pay private firms for individual text outputs, thereby fostering artificial monopolies, imposing additional costs on the public, and stifling innovation.

6 Utilizing Large Language Models to Countering Misinformation

LLMs, despite posing certain challenges, can be powerful allies in our battle against misinformation. Harnessing their potential for public good, however, requires a comprehensive reassessment of some prevailing assumptions surrounding media consumption and a reconsideration of how we can optimally employ these advanced technologies.

Commonly accepted beliefs in media consumption tend to underscore that factual news is invariably beneficial, that fake news has an inherent appeal that outstrips real news, that opinion pieces hold a higher engagement value compared to news-based content, that the propagation of fake news invariably leads to polarization, that fake news is intrinsically harmful, that consumers have a robust capacity to discern real news from fake, and that individuals are mindful of the differences between fact-based and opinion-based content. However, the validity of these assumptions is not universal and can vary based on situation-specific or demographic-specific factors. For instance, it might be plausible to assume that the inherently "dry" nature of news appeals to only a select group of individuals with high intellectual curiosity. Alternatively, it could be the case that existing methods of news dissemination and the stylistic choices employed do not optimally engage potential readers. Hence, tailoring news to cater to individual tastes and packaging it in an appealing format might foster better engagement and a more informed public.

In the context of these diverse interests and tastes among news consumers, not all news carries equal weight or relevance for everyone. Therefore, an effective strategy could involve the use of LLMs to generate content that is more engaging than what individuals would typically consume, consequently elevating the general level of public information. This could also provide an additional line of defense against the influence of unverified or misleading information. Utilizing LLMs, we can navigate the content space with unprecedented precision and flexibility, thereby creating content that is informative, engaging, and tailored to individual preferences and comprehension levels.

In addition to their role in crafting personalized content, LLMs can also significantly enhance automated processes, thereby reducing the workload on human moderators and promoting more efficient content moderation. For instance, in situations where definitive content classification is challenging, LLMs can generate low-risk "warning labels" that inform readers about potential controversies associated with the content. This intervention can foster a degree of informed skepticism among readers and help prevent the uncritical acceptance of potentially misleading information.

Deploying LLMs strategically in the creation of engaging, personalized, and comprehensible content stands as a potential boon for enhancing the public's overall informational level. This stands as a robust countermeasure against the proliferation of misinformation. Although LLMs present certain challenges in this endeavor, the significance of their potential as a tool for public good, when effectively harnessed, cannot be understated.

7 Conclusion

As we navigate the landscape of digital political campaigning in the age of Large Language Models (LLMs), the insights from this discussion provide vital touchpoints for future endeavors. The evolution of digital political campaigning, while opening new horizons, has also underscored the challenges posed by misinformation and the increasing polarization of public discourse. However, the advent of LLMs heralds unprecedented potential to address these issues while refining and enhancing the efficacy of political communication.

Addressing the perils that come with this evolution, we have identified two potential regulatory paths: restricting LLMs themselves or intervening to prevent the harms associated with them. The former, although seemingly straightforward, leads us down a path of ineffectiveness and undesirable outcomes, largely due to the open-source nature of models and data. In contrast, placing restrictions on firms that direct content towards consumers proves more feasible and efficient. This not only addresses the pressing issue of misinformation, but also ensures a more direct link between interventions and the prevention of harm.

Turning our attention to the potential of LLMs as a tool to combat misinformation, we delved into the assumptions regarding the consumption of media and its effects on public opinion. By creating more engaging and informative content, tailored to the preferences of the individual, we can raise the informational level of the populace, thus providing a buffer against misinformation. This explorative approach to content generation, coupled with more efficient moderation, can transform the manner in which information is disseminated and consumed.

In this context, the digital revolution of political campaigning and the rise of LLMs is not just a new challenge, but also an opportunity. It serves as a reminder that the same tool can be either a potent weapon or a force for good, depending on how it is wielded. By ensuring regulatory oversight, promoting transparency, and embracing innovation, we can leverage the potential of LLMs to create a more informed and engaged electorate, ensuring a healthier democratic discourse for the future.

References

Brückner, Michael, and Tobias Scheffer. 2011. "Stackelberg Games for Adversarial Prediction Problems." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 547–555. KDD '11. San Diego, California, USA: Association for Computing Machinery. ISBN: 9781450308137. https://doi.org/10.1145/2020408.2020495. https://doi.org/10.1145/2020408.2020495.

Carey, John M., Andrew M. Guess, Peter J. Loewen, Eric Merkley, Brendan Nyhan, Joseph B. Phillips, and Jason Reifler. 2022. "The Ephemeral Effects of Fact-Checks on COVID-19 Misperceptions in the United States, Great Britain, and Canada." *Nature Human Behavior,* https://doi.org/10.1038/s41562-021-01278-3. https://doi.org/10.1038/s41562-021-01278-3.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. "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* 42 (4): 1073–1095. ISSN: 15736687. https://doi.org/10.1007/S11109-019-09533-0.

Goodhart, C. A. E. 1984. "Problems of Monetary Management: The UK Experience." In *Monetary Theory and Practice: The UK Experience,* 91–121. London: Macmillan Education UK. ISBN: 978-1-349-17295-5. https://doi.org/10.1007/978-1-349-17295-5_4. https://doi.org/10.1007/978-1-349-17295-5_4.

Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns.* Crown.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1:19–46. ISSN: 0167-2231. https://doi.org/https://doi.org/10.1016/S0167-2231(76)80003-6. https://www.sciencedirect.com/science/article/pii/S0167223176800036.

Nyhan, Brendan, Ethan Porter, and Thomas J. Wood. 2022. "Time and Skeptical Opinion Content Erode the Effects of Science Coverage on Climate Beliefs and Attitudes." *Proceedings of the National Academy of Sciences of the United States of America* 119 (26): e2122069119. ISSN: 10916490. https://doi.org/10.1073/PNAS.2122069119/SUPPL_FILE/PNAS.2122069119.SAPP.PDF. https://www.pnas.org/doi/abs/10.1073/pnas.2122069119.

Sanderson, Z., M.A. Brown, R. Bonneau, J. Nagler, and J.T. Tucker. 2021. "Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform." *Harvard Kennedy School (HKS) Misinformation Review* 24 (4).

Srikanth, Maya, Anqi Liu, Nicholas Adams-Cohen, Jian Cao, R. Michael Alvarez, and Anima Anandkumar. 2021. “Dynamic Social Media Monitoring for Fast-Evolving Online Discussions.” In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,* 3576–3584. KDD ’21. Virtual Event, Singapore: Association for Computing Machinery. ISBN: 9781450383325. https://doi.org/10.1145/3447548.3467171. https://doi.org/10.1145/3447548.3467171.

Thèro, Héloise, and Emmanuel M. Vincent. 2022. “Investigating Facebook’s Interventions Against Accounts That Repeatedly Share Misinformation.” *Information Processing and Management* 59:102804. https://doi.org/10.1016/j.ipm.2021.102804. https://doi.org/10.1016/j.ipm.2021.102804.

Wylie, Christopher. 2019. *Mindf**ck: Cambridge Analytica and the Plot to Break America.* Random House.