MAS.S68, Spring 2023
4/19

Generative AI for Constructive Communication

Evaluation and New Research Methods

center for constructive communication

Agenda

1. Fireside chat

Cameron Raymond: Trust & Safety policy at OpenAI

Jakob Mökander: PhD finisher at Oxford Internet Institute, Princeton Center for Information Technology Policy

2. Discussion

Generative AI and Democracy

center for
constructive
communication

Discussion: Generative AI and Democracy

1. Recap of readings about LLMs and democracy

2. Reflections:
    a. **Benefits** of LLM-based applications in a democracy
    b. **Risks** of LLM-based applications in a democracy
    c. **Remedies** to these risks

Recap of readings about Generative AI and Democracy

How ChatGPT Hijacks Democracy (Jan, 2023) (twitter reaction)

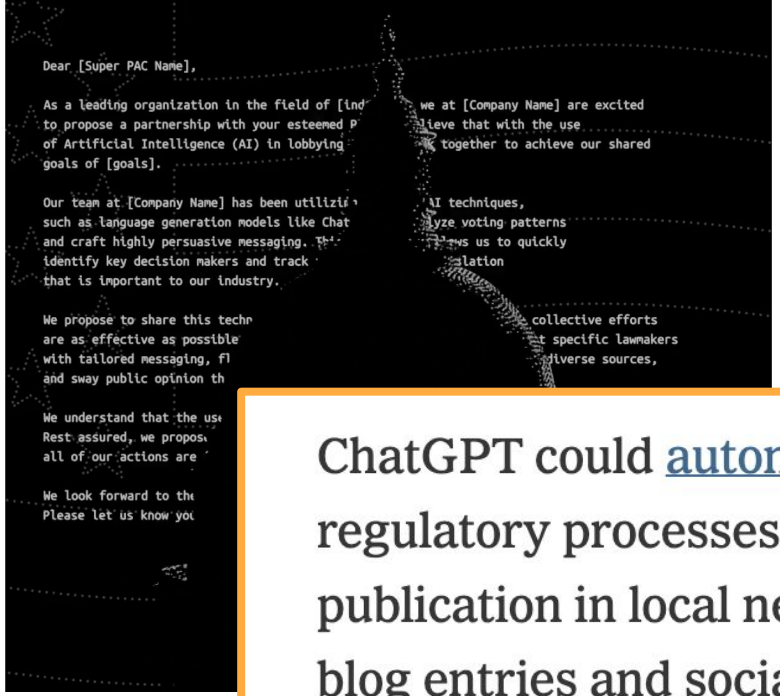
- ChatGPT could automate political lobbying by automating communication and influence tactics.
- AI-powered lobbying will outpace traditional methods due speed, cost-effectiveness, and broad reach.
- Although AI lobbying could democratize access, it may primarily benefit powerful institutions, further consolidating their influence.

Assessing the risks of language model “deepfakes” to democracy (May, 2021) (twitter)

- Deepfakes had minimal impact on 2020 election
- GPT3 malicious use limited: Barriers to access, detectability, generation quality
- “Text deepfakes” is a cat-and-mouse game. Platforms, regulators, researchers raising barriers against misuse while promoting media literacy and investing in innovative detection systems.

How generative AI impacts democratic engagement (March, 2023)

- LLMs could distort the legislative agenda by automating unique, seemingly genuine emails
- Policymakers can rely on alternative information sources and detection methods to mitigate risks.



[How ChatGPT Hijacks Democracy](#) (Jan, 2023)

ChatGPT could automatically compose comments submitted in regulatory processes. It could write letters to the editor for publication in local newspapers. It could comment on news articles, blog entries and social media posts millions of times every day. It could mimic the work that the Russian Internet Research Agency did in its attempt to influence our 2016 elections, but without the agency's reported multimillion-dollar budget and hundreds of employees.

Just as teachers will have to change how they give students exams and essay assignments in light of ChatGPT, governments will have to change how they relate to lobbyists.

Maybe an A.I. system could uncover which members of Congress have significant sway over leadership but still have low enough public profiles that there is only modest competition for their attention. It could then pinpoint the SuperPAC or public interest group with the greatest impact on that legislator's public positions. Perhaps it could even calibrate the size of donation needed to influence that organization or direct targeted online advertisements carrying a strategic message to its members. For each policy end, the right audience; and for each audience, the right message at the right time.

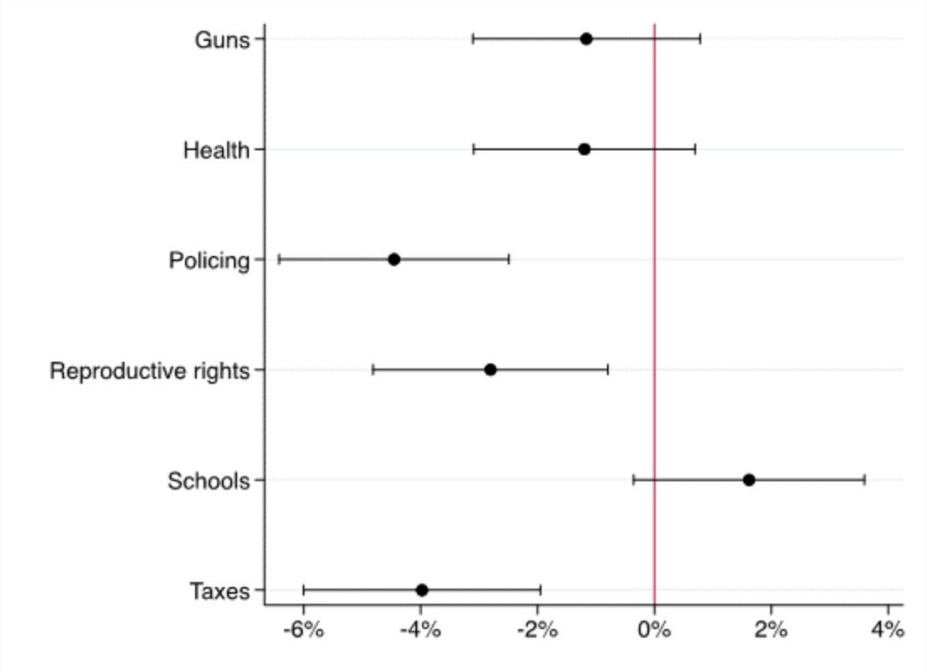
Assessing the risks of language model “deepfakes” to democracy (May, 2021)

With proactive countermeasures, even substantial innovations in language modeling may not change the nature of the game. Over time, the mice will multiply and become more elusive; to continue the chase, the cats must adapt. Technological innovations, such as synthetic text detection systems like GLTR or open-sourced fake news bots like GROVER, will accelerate this adaptation.

But perhaps even more important is spreading proper awareness about the issue at hand. For now, like language models themselves, popular concern about the current nature of the threat of automated disinformation is largely ungrounded. And there may be bigger issues to worry about with regard to their development—such as whether “racist, sexist, and abusive ideas are embedded” in the models, as MIT Technology Review’s Karen Hao points out in a report on the efforts underway to address such flaws. Fostering a measured public understanding of text deepfakes is a necessary step toward creating a society of minds resilient to them. Even if the 2020 US Presidential election was not overrun with deepfakes, it highlighted the profound danger of the spread of disinformation and lies in a democracy. The time to prepare for the next cycle is now.

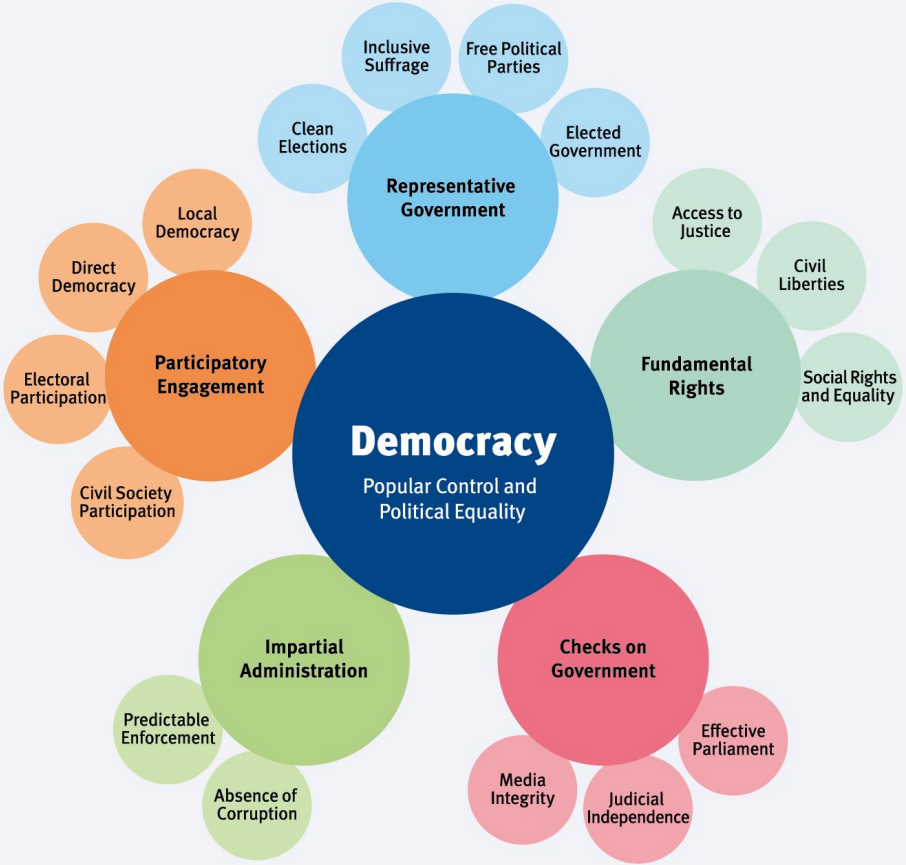
[How generative AI impacts democratic engagement](#) (March, 2023)

Figure 1: Differential Response Rates (GPT-3 – Human Emails) by Policy Area



Note: I-bars present 95% confidence intervals around each difference in means.

Conceptual Framework: The Global State of Democracy



What are possible **benefits** of LLMs to democracy?

- Lower bar for people to talk to their representative; e.g., non-native speakers
- Help politicians communicate their positions and proposals to the public more easily
- Help *counter* misinformation
- Help monitor public opinion
- Create a more informed public if AI-assisted education pans out
- Others?

Reflection: Benefits

From OpenAI's "Planning for AGI and beyond" blog post

> Generally speaking, we think more usage of AI in the world will lead to good, and want to promote it (by putting models in our API, open-sourcing them, etc.). We believe that democratized access will also lead to more and better research, decentralized power, more benefits, and a broader set of people contributing new ideas.

Consider the relationship between "democratized" as used above (*made accessible*) and "democracy" (*government by the people*).

Is access to AI "democratized"now? Is this trending up or down?

What is the threat to democracy if its benefits are not spread evenly?

Should AI access be prioritized the way broadband internet has been?

What are possible **risks** of LLMs to democracy?

- Automate harmful lobbying activity
- Generate **inauthentic** comments on news articles, message board posts, etc., that misleads both policymakers and the public
- Generate **misinformation** in these venues to steer public opinion
- Reduce trust between politicians and their constituents
- Reduce trust between citizens generally
- Others?

Reflection: Risks

How are LLMs categorically different from previous ways to spam polls, misinform, etc.?

The training data for language models does not represent the voice of every citizen equally. If an LLM is used to craft policy, could "blind spots" in the training data lead to disenfranchisement for certain groups of people?

▲ atoav 2 hours ago | root | parent | next [–]

> LLM has a lot of use cases where it can be enormously productive

The great chance of LLMs is of course assistive technology, where human actors and LLMs collaborate to do tasks. I am afraid however that what will shape the impact of LLMs on humanity much more is a different thing: Throught history there was always a certain number of people a dictator had to be at good terms with in order to *stay* in power. My fear is, that this number will become smaller, because it will be much easier to give the realistic *impression* that you have the support.

Existing concepts of reality and truth will definitly be completely and utterly destroyed by LLMs, and even *actual*, *real* information will be tainted by the fact that it *could* be fake – we are already seeing today on a smaller scale what living in such a world feels like if we look how societies in a post-truth environment operate.

My prediction (and I'd love to be wrong on that) is that the negative use of LLMs will outweigh positive use significantly, because it favours use cases where you don't have to care about correctness.

reply

(link)

What are possible **remedies** to the risks?

- Barriers to access
- Automated detection methods
- Auditing and regulation
- More civic education
- Others?

Reflection: Remedies

Large platforms may be able to detect “coordinated inauthentic behavior”, but the cat-and-mouse game will keep raising the costs for others to do so.

So what about smaller platforms that can’t afford these countermeasures, such as small town halls?

Should there be regulation limiting use of LLMs in participatory engagement?

"TruthGPT": Elon Musk talks to Tucker Carlson

00:00 - 00:45: Problem statement
03:24 - 04:22: Musk's proposed remedy
06:15 - 06:38: Call to action

Reflection: TruthGPT

“In the short term it’s being used by politicians to control what you think, to end your independent judgment and erase democracy on the eve of a presidential election.”

“They’re training the AI to lie…not to say what the data demands that it say”

Technical inaccuracies aside, how do you feel about what Carlson and Musk are implying – that AI alignment (or OpenAI’s, specifically) threatens democracy by “withholding the truth”?

Presentations

Presentation dates: 4/26 & 5/3

Full first draft- get feedback and incorporate it.

You'll then have until May 12th to write up your final paper!

Attendance expected at others' presentation dates!

Giving feedback on projects is critical to the "workshop" goal of this class.

center for
constructive
communication