LEAGUE OF WOMEN VOTERS® OF COLORADO

An Overview of Artificial Intelligence Concepts With An Eye Towards How AI Impacts Political Communication

To make informed decisions in elections, the public needs accurate information.

Deepfake AI: What is referred to in the proposed legislation, HB24 1147, by the buzzword "deepfake AI" is actually a whole suite of technologies, including very realistic computer generated graphics, online social media which allows rapid and long distance spread of information or disinformation, as well as compositional techniques that can be based on generative large language model (LLM) artificial intelligence.

> "Justin Levitt, Gerald T. McLaughlin Fellow and professor of law at Loyola Law School, shared his pessimism about AI's impact on democracy in the United States, including the ability to rapidly spread election misinformation. "Democracy depends on a set of different opinions and a set of common facts, and generative AI is going to be great for giving us an infinite array of disparate facts," he said. "That's a disaster for democracy.""

> ""All of us are capable of being bamboozled," Taylor (Carly Taylor, a data scientist and security strategist at Activision Publishing) said. "Everyone has confirmation biases, and in many cases across social media, we have spent every single day for years telling Facebook, Instagram, and LinkedIn exactly what we are biased toward by what we search, what content we consume, or with whom we engage … As a risk, that can become completely exploitable.""
> https://www.caltech.edu/about/news/generative-ai-regulation-kevin-roose

Since disinformation can be targeted to a specific audience, others might not even be aware that it is happening. And a flood of disinformation or even a campaign to be wary of potential disinformation can increase skepticism.

> Sarah Kreps Kreps, the John L. Wetherill Professor in the Department of Government in the College of Arts and Sciences and director of the Cornell Tech Policy Institute in the Cornell Jeb E. Brooks School of Public Policy
>
> "The threat might not be that people can't tell the difference – we know that – but that if as this content proliferates, they might just not believe anything," "If people stop believing anything, then it's eroding a core tenet of a democratic system, which is trust.
>
> https://as.cornell.edu/news/kreps-generative-ai-holds-promise-peril-democracies

The Brennan Center has a comprehensive analysis of AI and political discourse.

> Regulating AI Deepfakes and Synthetic Media in the Political Arena
>
> Policymakers must prevent manipulated media from being used to undermine elections and disenfranchise voters.
>
> In general, deepfake and other manipulated media regulations should cover images, video, and audio, any of which can effectively fool voters. At a minimum, the law should require disclaimers on professional campaign ads that use manipulated media — including but not limited to AI-generated content — to convey events or statements that did not actually occur.
> https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena

It is not the scale or speed that makes the actions wrong, even though that does increase the probability that the actions may be significantly damaging. There are also several aspects to "speed". One is how fast seemingly credible but fake content could be created. This is aided by AI techniques. The timeliness can be very harmful. The other is how quickly and broadly that information can be distributed. That has to do with the global internet and social media connections and has little to do with AI at all. Legal actions ought to depend on the harm done, not an argument over whether or not the actual produced material was AI enough to qualify under the law.

The US also has strong Constitutional First Amendment protections for free speech rights which must be protected. It is hard to draw a line between parody and maliciousness. Parody often focuses on attributes that can be mocked. In an era where we are much more sensitive to slurs that becomes increasingly more difficult to do in a manner in which all agree is appropriate.

The League of Women Voters needs to be playing a significant role in combating political disinformation, while protecting free speech rights. It is essential to empowering voters and defending democracy. LWV needs to be clear and knowledgeable regarding what needs to be regulated. What should be regulated is the deliberate and malicious spreading of untruths.

Generative Artificial Intelligence and how it works

Generative AI is a method of applied statistics that collects large databases of information available online and then is trained to make correlations and draw conclusions (and potentially take actions) based on the data it has been fed.

The data collection process vastly accelerates preexisting issues of intellectual property and personal identity protection.

The data input could be in the form of text, images or audio. Much of the current interest focuses on text, and large language models.

Large Language Models (LLM)

> ChatGPT stands for chatbot generative pre-trained transformer. The chatbot's foundation is the Generative Pre-Trained GPT large language model (LLM), a computer algorithm that processes natural language inputs and predicts the next word based on what it's already seen. Then it predicts the next word, and the next word, and so on until its answer is complete."
>
> "In the simplest of terms, LLMs are next-word prediction engines.
>
> https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html

A generative AI system learns to produce more examples that appear to be like the ones it was trained on.

> "These Large Language Models (LLMs) use unsupervised machine learning and are trained on massive amounts of text to learn how human language works. These texts include articles, books, websites, and more."
>
> "In the training process, LLMs process billions of words and phrases to learn patterns and relationships between them, making the models able to generate human-like answers to prompts."
>
> "They are made up of interconnected layers of algorithms that feed data into each other. Neural networks can be trained to carry out specific tasks by modifying the importance attributed to data as it passes between layers. During the training of these neural networks, the weights attached to data as it passes between layers will continue to be varied until the output from the neural network is very close to what is desired. "
>
> https://www.zdnet.com/article/what-is-ai-heres-everything-you-need-to-know-about-artificial-inteligence/

> "These things are approximations, but what they're approximations to is language use rather than language understanding. So you can get statistics about how people have used language and you could do some amazing things if you have a big enough database. And that's what they've gone and done.
>
> "So you can, for example, predict what's the next word likely to be in a sentence based on what words have happened in similar sentences over some very large database of gigabytes of data and often, locally, it's very good. The systems are very good at predicting category. "
>
> https://www.zdnet.com/article/the-next-decade-in-ai-gary-marcus-four-steps-towards-robust-artificial-intelligence/

Linear Algebra which deals with vectors and matrices and operations on these arrays is the mathematical basis for solving the computations in machine learning models.

See appendix for more technical aspects.

Generative AI and regulatory issues

> In accordance with Big Tech's philosophy of "move fast and break things",(a statement usually attributed to Facebook's Mark Zuckerberg) the development of these models have outpaced regulatory control. Only recently and well after the fact have attempts been started to assert copyright rights.
> https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Results are very dependent on the data set going in accordance with the old adage of "garbage in garbage out".

Timnit Gebru, formerly of Google, now the founder of the Distributed Artificial Intelligence Research Institute (DAIR) is a very prominent advocate for ethics in Large language model (LLM) AI.

> These sources are usually scraped from the world wide web and inevitably include material usually subject to copyright (if an AI system can produce prose in the style of a particular writer, for example, that is because it has absorbed much of the writer's work). But Gebru and her co-authors had an even graver concern: that trawling the online world risks reproducing its worst aspects, from

> hate speech to points of view that exclude marginalized people and places. “In accepting large amounts of web text as ‘representative’ of ‘all’ of humanity, we risk perpetuating dominant viewpoints, increasing power imbalances and further reifying inequality,” they wrote.
>
> https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Timnit Gebru and two colleagues at Google wrote a paper On the dangers of Stochastic Parrots https://dl.acm.org/doi/10.1145/3442188.3445922 highlighting these concerns. (Stochastic by the relevant Wikipedia definition refers to the property of being well described by a random probability distribution.) As a result of this paper, she was forced out (she says fired, they say resigned) from Google.

Gebru and her colleagues outline the following issues:

> https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

1. Environmental and Financial costs: Training large AI models consumes a lot of computer processing power, and hence requires very large facilities using a lot of electricity.

   > The sheer resources required to build and sustain such large AI models means they tend to benefit wealthy organizations, while climate change hits marginalized communities hardest.

2. Massive Data, inscrutable models

   > A methodology that relies on datasets too large to document is therefore inherently risky,”
   > “It will also fail to capture the language and the norms of countries and peoples that have less access to the internet and thus a smaller linguistic footprint online.”
   >
   > Large Language models learn as they are fed more data. The downside to this is that they will incorporate whatever racist, sexist or otherwise harmful material embedded in that material.
   >
   > Because the training data sets are so large, it’s hard to audit them to check for these embedded biases.

3. Misdirected Research Effort

   Though most AI researchers acknowledge that large language models don't actually understand language and are merely excellent at *manipulating* it, Big Tech can make money from models that manipulate language more accurately, so it keeps investing in them. "This research effort brings with it an opportunity cost," Not as much effort goes into working on AI models that might achieve understanding, or that achieve good results with smaller, more carefully curated data sets (and thus also use less energy).

4. Illusions of meaning

   The final problem with large language models, the researchers say, is that because they're so good at mimicking real human language, it's easy to use them to fool people.

Guardrails

In response, as part of the training process, groups like OpenAI developed guardrails, a set of ethical guidelines that regulate the content that language models like ChatGPT can communicate to users. Guardrails are to identify the potential misuse in the query stage and try to prevent the model from providing the answer that should not be given. This might include such things as criminal activity, child exploitation, offensive hate speech, sexism or racism.

https://arxiv.org/html/2402.01822v1

Even If overt racism is reduced, more covert racism remains. For example, ChatGPT and Gemini reports show that was found to discriminate against those who speak African American Vernacular English. "We know that these technologies are really commonly used by companies to do tasks like screening job applicants, AI models are already used in the US legal system to assist in administrative tasks like creating court transcripts and conducting legal research. "
https://www.theguardian.com/technology/2024/mar/16/ai-racism-chatgpt-gemini-bias

It should be noted that a human or team of humans performing the same tasks could hold similar biases

And guardrails can lead to odd results. How should bias be handled?

LEAGUE OF WOMEN VOTERS® OF COLORADO

> Google Gemini's unrealistic diversification results were mocked online as too "woke" when their image generator produced such things as black Nazis and a woman Pope : https://www.vox.com/future-perfect/2024/2/28/24083814/google-gemini-ai-bias-ethics

Malicious versions can be created with the right prompts (or flipping the guardrails as guides for misbehavior)

> WormGPT is described as "similar to ChatGPT but has no ethical boundaries or limitations." The developer claims that FraudGPT is great for learning how to hack, for writing malware and malicious code, for creating phishing content, and for finding vulnerabilities.
>
> https://www.zdnet.com/article/wormgpt-what-to-know-about-chatgpts-malicious-cousin/\

Hallucinations

> Even with attempts for careful guardrails and months of training, LLM AI is capable of "hallucinations", finding patterns where no pattern exists. AI models often hallucinate because they lack constraints that limit possible outcomes.
>
> AI models can also be vulnerable to adversarial attack, wherein bad actors manipulate the output of an AI model by subtly tweaking the input data.
>
> **https://www.ibm.com/topics/ai-hallucinations**

AI models should be tested, adjusted and retrained as data ages and evolves.

Positive aspects of AI and the Democratic Process

AI methods have become vital tools for analyzing very large data sets. Like other technological advancements, AI poses both new opportunities and new potential problems.

Back to Sarah Kreps:

> "A democracy really is about these connections between the government and people," Kreps said. "There are ways to think about how generative AI can be used in the public interest."

> Kreps said digital literacy education is needed well before college to help students understand and navigate technologies including generative AI, rather than trying to prohibit them.
>
> https://as.cornell.edu/news/kreps-generative-ai-holds-promise-peril-democracies

Appendix

Key words in technical explanations: Generative, Token, Attention, Transformer

> Computers only understand numbers. In large language models, words are evaluated as numerical representations called tokens. The model then determines the probability of a token or sequence of tokens occurring within a large sequence of tokens.
>
> "A key development in language modeling was the introduction in 2017 of Transformers, an architecture designed around the idea of attention. This made it possible to process longer sequences by focusing on the most important part of the input, solving memory issues encountered in earlier models."
>
> https://developers.google.com/machine-learning/resources/intro-llms

Attention indicates the statistical probability that a token is significant to other tokens in a sentence, paragraph or text.

> "Transformer architectures, first proposed in 2017, revolutionized natural language processing because they are so good at consuming very long strings of text—GPT-4 can handle whole books. Transformers break the text up into smaller pieces, called tokens, that are processed in parallel yet hang onto the context around each word."
>
> "The key to transformers is the attention mechanism: they decide what information is most relevant."
>
> "ChatGPT is actually doing multidimensional math. Each token of text becomes a string of numbers called a vector. The first time you enter a prompt, ChatGPT uses its mathematical attention mechanism to attach weights to each vector, and hence each word and word combination, to decide which to take into account as it formulates its response. It's a word prediction algorithm, so it starts by predicting the first word that

might begin a good response, then the next and the next, until the response is complete."

https://news.engin.umich.edu/2023/12/understanding-attention-in-large-language-models/

"Words are too complex to represent in only two dimensions, so language models use vector spaces with hundreds or even thousands of dimensions. The human mind can't envision a space with that many dimensions, but computers are perfectly capable of reasoning about them and producing useful results"

"word vectors are a useful building block for language models because they encode subtle but important information about the relationships between words."

"GPT-3, the model behind the original version of ChatGPT2, is organized into dozens of layers. Each layer takes a sequence of vectors as inputs—one vector for each word in the input text—and adds information to help clarify the meaning of that word and better predict which word might come next."

https://www.understandingai.org/p/large-language-models-explained-with

Each layer of an LLM is a transformer, a neural network architecture that was first introduced by Google in a landmark paper: https://arxiv.org/abs/1706.03762

"The transformer has a two-step process for updating the hidden state for each word of the input passage:

1. In the **attention step**, words "look around" for other words that have relevant context and share information with one another.
2. In the **feed-forward step**, each word "thinks about" information gathered in previous attention steps and tries to predict the next word.

OpenAI estimates that it took more than 300 billion *trillion* floating point calculations to train GPT-3—that's months of work for dozens of high-end computer chips.

It turns out that if we provide enough data and computing power, language models end up learning a lot about how human language works simply by figuring out how to best predict the next word. The downside is that we wind up with systems whose inner workings we don't fully understand."

https://www.understandingai.org/p/large-language-models-explained-withÍ