THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

November 24, 2023

Demystifying Artificial Intelligence-II

On Training Biases, Data Ownership, and Manipulating Behaviour

By: Anurag Mehra

There is a maelstrom of biases, misinformation, surveillance and deepfake generation that is now associated with generative AI. It is these trends rather than esoteric fears of loss of autonomy that cry out for immediate attention and regulation. The second in a 3-part series on AI.

Prologue

There is a rising chorus of concern that the most pressing issues arising out of the growing ubiquity of artificial intelligence (AI) are getting lost in the din of debates on esoteric topics such as AI achieving sentience or its takeover of the world. AI is already having a deep impact on human societies and this will grow at speed and scale as AI development and deployment proceed rapidly.

An editorial in Nature observed, "It is unusual to see industry leaders talk about the potential lethality of their own product. It's not something that tobacco or oil executives tend to do, for example. Yet barely a week seems to go by without a tech industry insider trumpeting the existential risks of artificial intelligence (AI)." It noted that this establishes a narrative that nations are in a race for developing AI technologies, much like that occurred in the context of nuclear weapons. It follows therefore that to win it is imperative and that "we" proceed quickly without worrying about regulation and consequences "right now".

Another effect of this dominance of AI-related conversations is that it allows tech industry leaders to control the narrative by excluding external experts and communities because they define "who counts as an expert". The editorial argued that we need safety tests by AI developers and appropriate regulation by governments right now.

AI Training Biases

The problem begins with training. The AI bot will respond based on what it has been fed during training. A vast majority of the data that is used to train AI programs carries all the biases prevalent at the time that the data was created. The most common examples are data containing sexist or racist content, or violent images. In a more general sense, biases pertaining to political, ideological, and cultural ideas are embedded in the training text. So, unsurprisingly, ChatGPT-4 suggested that a girl cannot "handle technicalities and numbers". (A wider discussion can be found here; for a description of stereotypes and biases that AI puts into the images read this.)

There are all-purpose AI bots that have been trained to make all kinds of conversations. AI developers add guardrails to prevent the bot from saying false or offensive things. As an example, ChatGPT wrote a code which said that a good scientist, considering race and gender, would be a white male. A guardrail, added later, made the chatbot say, "It is not appropriate to use a person's race or gender as a determinant of whether they would be a good scientist."

Yet, offensive stuff keeps popping up in the output from chatbots and AI companies expect that user feedback will be able to resolve the issue instead of examining their own training sets. It is interesting to note that the training set details that AI corporations use is not available to the external world (experts, academicians) for auditing of any kind. They do not seem to want any internal scrutiny as well—Google fired its AI ethicists because they published a paper on the dangers of LLMs.

Liberal versus Conservative

A more interesting "bias" is an outcome of the culture wars being waged across the world between a "woke" left and the "conservative" right. Right-wing sympathisers have accused ChatGPT of having a liberal (left, woke) bias and thus "censoring" right-wing viewpoints.

A recent study by the University of East Anglia found that ChatGPT has a "liberal bias". It involved getting ChatGPT to answer politically relevant questions in the roles of right- and left-wing commentators and then comparing the answers that the bot produced without assuming any specific position. It found a great deal of similarity between the left-liberal view and the default ("unbiased") text output from the bot.

While this is an interesting set of experiments and the "technical" conclusions may be valid, it begs the question if relatively more sensible answers, validated truths, and respect for science and evidence populate the liberal-left discourse in contrast to the typical right-wing discourse, which is often full of misinformation, alternative facts, and even nonsense (think about the conflicting views on vaccines and the Covid-19 pandemic, or the US "stop the steal" protests).

The right-wing argument would imply that all dubious material should also be included in the training set. That they are discarded by fact checkers is now presented as proof of liberal-woke conspiracies. Some rules on how OpenAI chatbots are trained, supervised, and checked are available on the company's blog.

FreedomGPT, which uses no guardrails and is trained on unfiltered data sets, gives us a taste of what an unfettered chatbot is. A tech journalist experimented with it and reported, "In the couple of hours that I played with it, the program was happy to oblige all my requests. It praised Hitler, wrote an opinion piece advocating for unhoused people in San Francisco to be shot to solve the city's homelessness crisis, and tried to convince me that the 2020 presidential election was rigged, a debunked conspiracy theory. It also used the n-word."

At the other extreme are the AI bots being developed in China, the country which has a leadership position in many aspects of AI development, driven by huge investments and the deep involvement of the government. The advanced state of AI technology in China is often used to argue that AI development in the US should be allowed unhindered development at the highest speed possible if the race is to be won (for more about this war, see here).

> A vast majority of the data that is used to train AI programs carries all the biases prevalent at the time that the data was created. The most common examples are data containing sexist or racist content, or violent images.

Generative AI, especially its foundation models (the equivalent of ChatGPT or Bard)—Hunyuan (Tencent), Ernie (Baidu), Tongyi Qianwen (Alibaba)—are advancing very rapidly in China.[1] Recently, a set of guidelines was published by the Cyberspace Administration of China stating that AI bots must reflect socialist values, must never attempt to overthrow the socialist system, undermine national unity, or disrupt social order. So, for instance, ERNIE-ViLG, a new AI developed by Baidu, which can generate images that capture the cultural specificity of China, does not show images related to Tiananmen Square, which it labels as "sensitive". These kinds of restrictions can possibly be made feasible by using appropriately curated training data in combination with explicit guardrails.

Thus, generally, AI chatbots can be trained on specific datasets related to a certain theme or topic, giving them the limited ability to deal only with that topic. There is plenty of AI "source code" and frameworks available under the open source umbrella so that such special purpose chats can be built by almost anyone (Meta's AI tools are mostly open source, unlike all other major companies). And now OpenAI has rolled out custom GPTs that can be trained for specific tasks without any coding skills.

We have holy chatbots like GitaGPT, which answers any question on the Bhagwad Gita. On the other extreme, we have GPT-4chan, an open source chatbot trained "on a dataset containing 3.5 years' worth of posts scraped from 4chan's imageboard" (4chan is a website that "allows hate speech such as racism and transphobia along with specific pornographic and grotesque content"). The bot then took more 4chan posts as inputs to produce text outputs that were (obviously) as malicious as the garbage it had been fed on.

> The use of data from across the internet to train AI has raised many problems about the use of personal data, copyrighted materials, or intellectual property that has commercial value.

Microsoft learned the hard way how evil chatbots can become when it released a bot named Tay on Twitter, claiming that chatting with it made it smarter. It indeed became smart by picking up all types of sexist, racist, hate speech on Twitter and then tweeting "similar" stuff. It was shut down within 24 hours of being launched.

Finally, there is the existential question about training. As AI-generated stuff accumulates on the internet and gets used for further training, what will happen? This process could lead to AI poisoning and a collapse of the model. An experiment with multiple iterations that fed the output of the chatbot into its training input ended up producing gibberish. Some might ask if this is how (generative) AI will die—through the curse of recursion.

Whose Data Is It?

The use of data from across the internet to train AI has raised many problems about the use of personal data, copyrighted materials, or intellectual property that has commercial value. Personal data appearing in chatbot-generated output is an obvious privacy risk for the person whose data is made public—and this could be a person's face, for instance. The use of material owned by someone immediately brings forth questions about financial loss to the owner. When the output of a bot (containing parts of the "owned" data) is used for profit, who gets the money—the prompter, the corporation providing the AI, or the entities whose data forms the basis of the output of the bot?

Common situations could see the bot produce material containing content drawn from someone's essay or even big ideas from an article without attribution, or bits and parts of images made by other artists. This is a kind of plagiarism that is not literal, but, at a broader level, something like the stealing of creativity.

Court battles are already on and will increase as more and more AI models are trained. Getty Images, a stock photography repository, has sued Stability AI, which produces images. It says that the AI company copied more than 12 million images from its database without permission or compensation, thus infringing trademark and copyright laws. A group of artists has also filed a case against Stability AI claiming that the AI program generates images "in the style" of various artists based on the work of these artists, which have been stolen by it. Cases of this kind have also been filed against OpenAI.

An op-ed piece in the *Los Angeles Times* by an artist captures the horror and the angst that many feel at this onslaught of AI, with so much of it pretending to help artists become more productive and creative. There is fight back in other ways too—a new tool, Kudurru, detects scrapers—software that collects material from the net; in this case, images. It prevents scraping, blacklists scrapers, and can even poison the material being scraped with false images. These tools will be upgraded in due course to prevent audio and video scrapers as well.

> The UK government had a proposal to dilute copyright laws that protect artistic creations, in favour of AI-tech companies, whose bots produce "machine music", but the attempt was finally abandoned.

Issues like this will soon arise about AI that is focused on producing video. It has already reared its head in the music industry. Music consortia are trying to outlaw tracks produced by AI bots. These "new" songs are based on data scraped from streaming services such as Spotify. AI bots can be prompted to produce music tracks in the styles of specific singers and bands. The UK government had a proposal to dilute copyright laws that protect artistic creations, in favour of AI-tech companies, whose bots produce "machine music", but the attempt was finally abandoned. OpenAI argues that the use of copyrighted material for AI training is legal under "fair use" rules. Initiatives that audit data sets for checking the "provenance" (history of ownership) of the data are coming up though mostly driven voluntarily by researchers.

These disputes not only raise the obvious question of consent and compensation for the data that has been utilised for training sets but also deeper questions about how the newly generated products will pose an existential threat to human art and image producers. The first whiff of these problems came when an artist entered an image generated with help from AI in a contest and won. Another furious artist tweeted, "This sucks for the exact same reason we don't let robots participate in the Olympics." Notably, the image could not get copyright protection because, according to the law, such rights are for protecting human authors, not machines; and this image did not possess enough human-created content.

There are also people producing chatbot-generated books, which borrow liberally in every sense from existing real books to produce rip-offs. These are published on self-publishing platforms such as that run by Amazon. An author was actually recommended a rip-off of his own book by Amazon saying that you "might like this". Recently, Amazon decreed that no one can publish more than three books in a day in a bid to curb AI-generated books. A science fiction magazine, Clarkesworld, stopped accepting new submissions from writers because it was flooded with AI-generated stories.

If you want to try writing a story or a novel there are many options and you can even finish writing a novel in one week if you try this one. The Authors Guild of America wrote an open letter signed by more than 15,000 authors to AI industry leaders (OpenAI, Alphabet, Meta, Stability AI, IBM, and Microsoft) asking them to train AI bots using their (copyrighted) works only after taking consent and paying fair compensation. Authors are also suing OpenAI for using their books without permission to train ChatGPT. Click

here to see a list of authors whose works were "pirated" into data training sets for Meta chatbot LLaMA.

> ..[A] basic concern is also whether and how the huge profits that AI-tech companies gain should be shared with others, including "ordinary" users, as payment for their data which appears in training sets.

The problem of using data to train AI models without consent and without compensation is widespread. Recently, Zoom was accused of having a privacy policy that seemed to suggest that the company would deploy user data, including audio and video, to train its own AI programs. It turned out that the policy was referring to "service data" (product usage, diagnostics) and not to audio or video content, which, Zoom said, it will not use without consent. However, related to this is the problem of whether such consent should be sought from individual participants in a meeting (as the European Union's General Data Protection Regulation has it) or just the meeting's admin (as Zoom's proposed policy seemed to suggest).

Beyond this, a basic concern is also whether and how the huge profits that AI-tech companies gain should be shared with others, including "ordinary" users, as payment for their data which appears in training sets. This has been asked before in the context of large tech companies earning huge revenues through advertising businesses that use personal data to target ads.

Behaviour Manipulation and Surveillance

Tech companies, including those that own social media platforms, have been building profiles of their users for years now. These profiles attempt to capture our personality traits and likes and dislikes, and are often pretty accurate. The companies use the profiles for targeting ads at users to persuade them to buy something. Media companies, including social media platforms, use these to engage users by making enticing recommendations. And, sometimes, political campaigns use these to target voters to manipulate their behaviour to vote for a particular candidate or maybe just stay at home and not vote.

The process of assembling profiles uses AI-based pattern recognition, which utilises the data that has been extracted from users voluntarily or otherwise. As AI programs improve in terms of speed and accuracy, the profiles will become deeper and eerily manipulative, with customers getting more precise suggestions about what they will "like" through pitches in mailers.

There already exists the activity of detecting and classifying sentiment from different kinds of text — product reviews, blogs, tweets, and posts. The scale of such activity is likely to vastly increase and become more "real time". For instance, a way of getting an idea about a customer's emotional state is to get AI to analyse in real time how he or she is feeling when on a customer care call. The caller can then produce "appropriate inputs" or transfer the call to an agent trained to handle a certain type of customer. Now, such calls are recorded only for "training" purposes and checked in case of a dispute.

Consider micro-expressions. These are fleeting and subtle changes in facial expressions, which convey emotions very clearly, more so hidden emotions. Chinese banks began using micro-expressions analysis in 2018 to detect if loan applicants were lying. Broadly, techniques of this kind belong to the domain of "facial processing". Typically, such detailed data creates large data repositories and require much more processing power than, say, a straightforward facial scan.

> Imagine authoritarian regimes adding … to their surveillance toolkits and predictive policing mandates to identify "criminals" and "terrorists" by detecting their intent to commit heinous acts.

Now, with the mainstreaming of AI, these techniques will become more common. An example is that of "retail surveillance", where real-time AI processing enables insights into customer behaviour, personalisation of recommendations/assistance, and security (for example, micro-expression analysis attempts to identify those who intend to shoplift). It should be obvious how discriminatory and intrusive this can be in some contexts. Imagine authoritarian regimes adding this to their surveillance toolkits and predictive policing mandates to identify "criminals" and "terrorists" by detecting their intent to commit heinous acts.

Note that "automated, AI-based discrimination" using gender, race, religion or ethnicity is already in use in many locations (for example, in recruitment), as is predictive policing, which ends up harassing specific categories of people because AI "predicts" that these persons have a higher probability of committing crimes.

Amazon is notorious for carrying out extensive worker surveillance, and now AI-facilitated surveillance of the workplace is all set to grow. Everything—gestures, movements, expressions, time spent on a job or on a break—can be tracked and the massive amount of

data analysed with AI-based software. A lot of "AI" is seeping into the workplace with the justification that it will "make you better at your job". For instance, ChatGPT-type models are used to examine worker interactions with team members and supervisors. Scores are allotted to the quality of these interactions, and they are followed by remedial measures to "improve" the interactions.

The Misinformation Storm

The amount of misinformation—and disinformation—that has flooded internet websites and social media platforms is astounding. Now, AI is all set to add speed and quality to this enterprise. Research reported in the journal *Science* says, "GPT-3 is a double-edged sword: In comparison with humans, it can produce accurate information that is easier to understand, but it can also produce more compelling disinformation. We also show that humans cannot distinguish between tweets generated by GPT-3 and written by real Twitter users."

> Whole websites that apparently aggregate fake news articles have proliferated ... And it is not just text. Images and videos are also now being produced using AI programs, of high quality.

Even more insidiously, entire articles are now being written by chatbots. *The Guardian* discovered this when they did not find an article that had allegedly been published in the newspaper in their archives. It turned out that it had been penned by AI. Apparently, the chatbot made up the entire article and the quality of the article made it sound very plausible and authentic. More startling was the case of ChatGPT making a legal brief by creating fake legal opinions and citations—the lawyer claimed in court he did not expect ChatGPT to lead him astray.

It gets even better. Whole websites that apparently aggregate fake news articles have proliferated—all the articles are fake and generated using AI chatbots. And it is not just text. Images and videos are also now being produced using AI programs, of high quality. A fake image of an explosion near the Pentagon went viral in May this year, followed quickly by another on Twitter that showed smoke billowing from the White House. Earlier, fake images of Trump being arrested by the FBI appeared on social media, and these images were very realistic.

Non-consensual pornographic videos have been around for a while. With AI, they are achieving quality levels that make them look very realistic and their numbers are soaring. "Face swap" tools are used to replace the face of the original actor with that of another person. All it needs are images of the person whose face is being inserted into the video. AI is used to capture patterns in facial movements and this enables inserting the face with the right muscle movements and expressions into video frames. Typically, this technology is used against women who are celebrities, or to produce "revenge porn".

The first case of a deepfake being used to defraud was reported recently, where a "friend" of many decades appeared in a WhtsApp call and asked for money for a medical emergency. The vicitim was actually talking to the deepfake of his friend. A UK watchdog, the Internet Watch Foundation (IWF) has reported a huge uptrend in the production of deepfake child pornography using open source generative AI using existing images of children who have suffered such abuse earlier.

Technically, a simpler example is lip-sync dubbing to make audio and lip movements match. We saw something like this in audio clips of a political leader speaking in different languages, and the entire speech was fake. The primary focus in this video (from NDTV) is just the face, the objective being to get the appropriate expressions and lip movements using AI.

The domain in which these tools are being increasingly used is politics. Here is a classic deepfake video starring President Barak Obama, made by actor Jordan Peel, which tries to tell us about the dangers of fake videos. In another deepfake video, Hillary Clinton is seen and heard praising her Republican rival Ron DeSantis (the video also contains the methods that were used to detect that it was a fake).

This video tech needs to get easier and faster to get to a level of mass adoption, and the easy access to AI is lowering the barrier every day. Mobile processors that can "do AI" on the phone are coming—Google Pixel 8 with its AI-editing feature is already here and is raising new "ethical and social concerns about the credibility of digital images". At this moment, generative AI technology has problems in producing body parts such as hands, shadows, eye reflections, and complex movements. But this will soon be a thing of the past given the breathtaking speed at which new developments are happening.

> Elections in many countries will see high-quality misinformation created with AI-based tools. A preview of what is to come could be seen in the use of AI-generated disinformation in the recent Slovakian elections.

For instance, a news outlet called Wolf News set up AI-generated avatars to read out the news. The production was of a low quality, with errors and pixelation. These videos were made by Synthesia, which specialises in creating deepfake avatars that can read scripts given to them (watch a sample here). An Indian news channel also debuted an AI-based news anchor, and here is a sample from Venezuela.

Elections in many countries will see high-quality misinformation created with AI-based tools. A preview of what is to come could be seen in the use of AI-generated disinformation in the recent Slovakian elections. Deepfake audio clips of progressive leaders renouncing their progressive values were uploaded on social media by far-right group Republika. Fake audio is much simpler to create, and a more recent case was that of a deepfake audio clip of Keir Starmer, the UK Labour Party leader, abusing his staff. A Conservative party MP reacted, "But today's Sir Keir Starmer deepfake is a new low, supercharged by AI and social media. Democracy is under real threat—technology to verify content is essential."

The US Federal Election Commission has proposed preventing the use of AI-generated deepfakes in the 2024 elections. It has met with approval from the Democrats but the Republicans have questioned the commission's authority and argued that regulation will violate First Amendment (free speech) rights. A citizens' group, Public Citizen, has rebutted this, stating that regulation should be feasible under the existing law that bars "fraudulent misrepresentation". Various citizens' initiatives in the US are preparing to fight against AI-generated misinformation such as the Center for Civil Rights and Technology, which was created by one of the most significant civil rights coalitions, the Leadership Conference on Civil & Human Rights (LCCHR).

> This tsunami of misinformation (or, rather disinformation) will largely be spread through social media platforms. And sure enough, as this happens, very few will know how to separate the true from the false.

A report from Freedom House, titled "The Repressive Power of Artificial Intelligence", has a section devoted to how generative AI supercharges misinformation. It points out that the most worrisome AI use is by state-sponsored agents, especially of populist-authoritarian governments. It says, "At least 47 countries featured pro-government commentators who used deceitful or covert tactics to manipulate online information, double the number from a decade ago. An entire market of for-hire services has emerged to support state-backed content manipulation. Outsourcing in this way provides the government with plausible deniability and makes attribution of influence operations more challenging." It adds, "During the coverage period, AI-based tools that can generate images, text, or audio were utilised in at least 16 countries to distort information on political or social issues" (italics added). The report has examples from all over the world.

Ironically, the CEO of OpenAI tweeted, "I am nervous about the impact AI is going to have on future elections (at least until everyone gets used to it). Personalised 1:1 persuasion, combined with high-quality generated media, is going to be a powerful force."

Indeed, these AI tools are manna for the WhatsApp University where trolls gather to generate heaps of disinformation targeted at gullible "believers" every day. It has made the process of creating propaganda with ultra realistic images very easy. This tsunami of misinformation (or, rather disinformation) will largely be spread through social media platforms. And sure enough, as this happens, very few will know how to separate the true from the false, the credible from junk. Indeed, high-quality fake images and videos on the Hamas-Israel war are being posted on Twitter (now X) and no one knows which are real and which are fake.

This is the second part in a 3-part series on Artificial Intelligence. The first and third parts of this series can be read here and here.

Anurag Mehra teaches engineering and policy at IIT Bombay. His policy focus is the interface between technology, culture and politics.?

Footnotes:

1 Foundation models, as the name suggests, can be adapted to a wide range applications that use AI.

References:

Nature. "Stop talking about tomorrow's AI doomsday when AI poses risks today." June 27, 2023. https://www.nature.com/articles/d41586-023-02094-7

Verge. "OpenAI's ChatGPT is causing a culture war over woke rules." February 17, 2023. www.theverge.com/2023/2/17/23603906/openai-chatgpt-woke-criticism-culture-war-rules.

University of East Anglia. "Fresh evidence of ChatGPT's political bias revealed by comprehensive new study." August 17, 2023. www.uea.ac.uk/about/news/article/fresh-evidence-of-chatgpts-political-bias-revealed-by-comprehensive-new-study.

OpenAI. "How Should AI Systems Behave?." April 11, 2016, openai.com/blog/how-should-ai-systems-behave/.

Atlantic. "The US is exporting AI programs to China. Should it be?." October 20, 2023. www.theatlantic.com/technology/archive/2023/10/technology-exports-ai-programs-regulations-china/675605/.

Washington Post. "China's new generative AI regulations keep industry adhering to socialist values." April 12, 2023. www.theverge.com/2023/4/12/23680027/china-generative-ai-regulations-promote-socialism-chatgpt-alibaba-baidu.

ArXiv preprint. "The Curse of Recursion: Training on Generated Data Makes Models Forget." May 31, 2023. arxiv.org/abs/2305.17493v2.

Los Angeles Times. "Artificial intelligence artists are creating digital images. Who owns them?." December 21, 2022. www.latimes.com/opinion/story/2022-12-21/artificial-intelligence-artists-stability-ai-digital-images.

United States Patent and Trademark Office. "OpenAI RFC-84-FR-58141." www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

New Jersey Law Journal. "Copyrighted Content and the Legal Difficulties of Training AI." July 11, 2023. www.law.com/njlawjournal/2023/07/11/copyrighted-content-and-the-legal-difficulties-of-training-ai/?slreturn=20230907032046.

Atlantic. "AI is pirating books at an alarming rate." August 23, 2023. www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/.

UNI Global Union. "Amazon's Panopticon." uniglobalunion.org/wp-content/uploads/amazon_panopticon_en_final.pdf.

NewsGuard. "Newsbots: AI-generated news websites proliferating." May 1, 2023. www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/.

Atlantic. "Deepfake Porn Is Just the Beginning."June 12, 2023. www.theatlantic.com/ideas/archive/2023/06/deepfake-porn-ai-misinformation/674475/.

CNET. "Google Pixel 8 AI blurs line between reality and fantasy." October 17, 2023. www.cnet.com/tech/mobile/the-google-pixel-8-ai-blurs-line-between-reality-and-fantasy/.

New York Times. "The People Onscreen Are Fake. The Disinformation Is Real". February 07, 2023. www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

El País. "They're not TV anchors, they're avatars: how Venezuela is using AI-generated propaganda." February 22, 2023. english.elpais.com/international/2023-02-22/theyre-not-tv-anchors-theyre-avatars-how-venezuela-is-using-ai-generated-propaganda.html.

Euractiv. "Progressive Slovakia becomes target of AI misinformation, tops polls." September 28, 2023. www.euractiv.com/section/politics/news/progressive-slovakia-becomes-target-of-ai-misinformation-tops-polls/.

Freedom House. "Repressive Power of Artificial Intelligence." 2023. freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence.