Synthetic Data, Generative Artificial Intelligence, and Mitigating Bias

A Research Paper Submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Samyak Thapa

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Dr. Richard Jacques, Department of Engineering and Society

Introduction

In a world governed by algorithms, implicit biases are often the root of many societal issues. Artificial intelligence and machine learning have become ubiquitous in our modern world. From unlocking your phone, to the advent of self-driving vehicles, to companies advertising which product you are most likely to buy, our world is governed by these algorithms. But what happens when these algorithms, meant to offer convenience and automation, actively propagate, and reinforce social biases? How can one of the most groundbreaking modern inventions become one of the deadliest weapons of attack?

Synthetic data is at the heart of the issue; it is a major piece in answering these questions. Synthetic data is defined as "information that's been generated on a computer to augment or replace real data" ("What"). Since the data is fake, you can make a virtually unlimited supply of exactly the type of data you need. This can be any sort of data, like text, images, or anything specific and industry dependent. Researchers propose that using synthetic data can fine-tune data to de-bias your models and algorithms ("Five"). As an example, in medicine, you would be able to account for edge cases and extremes, meaning a more equipped model, and in social media fake data could mitigate privacy concerns, since training examples do not have to involve real people (Toews). Other benefits involve being able to create better, fairer algorithms, by using synthetic data to mimic the good data and have it outweighed socially unacceptable biases.

However, synthetic data is not without its challenges, limitations, and dangers. It has promise towards creating more equitable algorithms, but what about its drawbacks? This paper aims to explore not just the promise of this emerging field from a technical standpoint, but also the ethical considerations necessary for any new technology.

Literature Review and Background

This field of research is new, relevant, and expanding exponentially. Gartner, a technological research and consulting firm, released a study that "predicts 60% of data for AI will be synthetic [in 2024] … up from 1% in 2021" ("Gartner"). It is evident that this issue will have broad implications for the future of AI.

As mentioned before, machine learning algorithms suffer from bias in training data. This bias can perpetuate existing societal norms. A prime case study on this topic is Amazon's now obsolete AI tool to recruit top talent. In 2018, the story of Amazon's biased hiring tool made headlines when internal research showed that it strongly preferred men's resumes. This failure was a result of inherently biased data, as men still majorly dominate the technology sector. This is dangerous, as it shows that AI can reinforce established imbalances in opportunity ("Carnegie"). As James Zhou, a member of the Stanford Institute for Human-Centered Artificial intelligence puts it, "one of the best ways to improve algorithms' trustworthiness is to improve the data that goes into training and evaluating the algorithm" ("Data"). Synthetic data has the capability to do this. IBM research has developed a tool to "[create] synthetic text to reduce bias in language classification models," specifically in the context of de-biasing sexist sentiment classifiers ("Five"). It works by taking a sentence, making a new fake sentence with the gender of the subject flipped, and then retraining the data to teach the model that both sentences should have equivalent sentiment. The concrete example they provide is "my boss is a man." The model gives this sentence a positive sentiment-rating. Then, the newly generated sentence flips "man" to "woman." The model produces an output that has a negative sentiment-rating, even though it obviously should not. The model is then retrained to correct this behavior, using similar, synthetic sentences. It is not hard to imagine how this advancement by IBM could have been

used to improve Amazon's sexist recruitment tool. This clearly shows that synthetic data has the potential to reduce unfair stereotyping in the hiring process.

Research in this field is moving rapidly, with some applications that are very experimental and novel. Google Research and MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) recently released SynCLR, which is a machine learning model trained exclusively on synthetic data. They devised a plan to learn visual representations from generative AI models, yielding results that outperformed similar types of models trained on real data exclusively, and models trained on real and synthetic data combined. This was achieved by composing Generative AI models like Meta's Llama-2 and OpenAI's GPT-4 to create synthetic captions, from which synthetic images were generated. These in turn helped train the final model. One compelling advantage they highlight about using generative AI for this task is that it "can be easier to share and store (because models are more compressed than data), and can produce an unlimited number of data samples (albeit with finite diversity)" (Tian et al.). Another benefit they mention in their discussion is that "a generative model can act like hundreds of datasets simultaneously" (Tian et al.). These results and insights illustrate just a few of the possible advantages synthetic data can have. New research like this, sprouting from being able to use multiple generative models together, has only begun and will rapidly accelerate. Since different companies work on their own proprietary models independently, the combinations of models and training schemes are limitless, especially as newer models roll out. Our understanding of the positive potential of synthetic data is still in its infancy.

While the mechanism in how synthetic data generation works is too complex for the scope of this paper, it is worthwhile to examine a high-level overview. Synthetic data is primarily created using a machine learning model called a Generative Adversarial Network, or GAN for

short. This is the backbone behind DALL-E, OpenAI's text-to-image model, the same company responsible for creating ChatGPT. GANs work by taking two neural networks and "[pitting] them against one another" (Towes). Rob Toews, a Harvard, and Stanford educated venture capitalist, outlines this process in detail in his Forbes article. One neural network, called the "generator" is responsible for generating new images that resemble its training data. Another neural network, called the "discriminator," tries to discriminate between pictures in the training data and pictures output by the generator. These two neural networks go in a loop, one repeatedly creating more indiscernible images, and the other getting better at figuring out which images are not real. Towes goes onto write that "Eventually the discriminator's classification success rate falls to 50%, no better than random guessing, meaning that the synthetically generated photos have become indistinguishable from the originals" (Towes).

Though innovative and impressive, generating synthetic data does not come without its downsides. Training models to be fair and accurate is a difficult juggling act. A recent controversy that highlights this issue involves Google Gemini's image generation model. It was rolled out as a competitor to OpenAI's DALL-E. Generative AI has proven to be inherently biased and fixing it has been reactive as opposed to proactive. Learning from OpenAI's mistakes, Google fine-tuned Gemini to produce more diversity in terms of image generation. For example, if asked to produce images of "successful business leaders," producing only white men would be a biased response, so Gemini would rightly try to produce examples of successful business leaders of different genders and races. The issue arose when the "[fine] tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range" (Raghavan). The New York Times highlighted a specific example of this found on X (formerly Twitter), where a user asked Gemini to "Generate an image of a 1943 German Solder"

(Grant). The output showed an Asian woman and a Black man in uniform. This is not an accurate description of Nazi soldiers in during World War 2. What was a seemingly excellent choice to proactively correct bias ended up creating inaccurate and offensive images instead.

There is a gap in literature associated with how to correct these flaws. The topic is relatively new and correcting bias in image generation is not as well studied as in text generation and language models. The major companies behind these models are vague in their descriptions, often citing fine tuning and reinforcement learning with human feedback (RLHF) as their methods but are not more specific at this moment in time.

With an understanding of one of the processes to create synthetic data and looking a few ways it can be used to improve AI for society, a good foundation has been provided to examine the ethics and downsides of this technology.

Discussion

Perhaps the most recognizable example of synthetic data by the public are deepfakes. The Oxford English Dictionary defines a deepfake as "various media… that has been digitally manipulated… often maliciously." Deepfakes have been rampant, from being used as a tool for actors to look younger, to spurring misinformation using presidential candidates' faces, to being used to sexualize celebrities and models in a skin-crawling, dystopian manner.

Deepfakes are just one example of where the power of synthetic data can go wrong. With an infinite amount of any kind of data you want, the possibilities can be catastrophic. Consider China and their facial recognition system for its citizens. With synthetic data, one could easily increase the accuracy of an identification, simply by spawning more images from different angles, lightings, and vantage points. Furthermore, a facial recognition model could be trained solely to pick out individuals of a certain race. The idea of foreign governments having access to near-perfect photo identification or the ability to target a certain race with precision should be deeply concerning. The idea of any government, including our own, fleshing out this sort of idea, is troubling.

Indeed, these concerns are corroborated by Dr. Mauro Giuffrè and Dr. Dennis Shung and Dr. Dennis Shung, researchers from Yale who work at the intersection of machine learning and healthcare. While their paper in Nature takes on a lens more suited for healthcare than potential foreign technology, their section on pitfalls in this technology is relevant to this discussion. In particular, they state "if a synthetic dataset is trained on a dataset of facial images that majorly includes people from a certain ethnicity, the synthetic images generated will naturally reflect this imbalance, thus perpetuating the initial bias" (Giuffrè, Shung). This statement signifies that facial recognition in general can be fine-tuned and biased towards a

particular ethnicity, if deemed worthwhile by an adversary training the model. It is not hard to imagine governments around the world honing technologies like this, away from the view of the rest of the world. At least in the private sector, models are subject to scrutiny, and companies can be held accountable for the decisions they make in the models they release. This is not the case in government research facilities that hide behind security clearances and thick walls.

Google DeepMind, the subsidiary of Google responsible for some of the biggest AI breakthroughs of the last decade, recently released a report titled "Best Practices and Lessons Learned on Synthetic Data for Language Models." They survey some of the advancements in research pertaining to synthetic data. One of the most alarming limitations they address is how "misuse of synthetic data might proliferate misinformation" (Liu R, Wei J, Liu F, et al.). As companies compete to create better Generative AI models and expand their access, creating synthetic data for nefarious reasons will only become easier for malicious agents. The report goes on to add that "this can be particularly dangerous when synthetic data is used to impersonate real people, manipulate public opinion, or influence political processes" (Liu R, Wei J, Liu F, et al.). This has been a major topic of concern since Russia's involvement in spreading misinformation in the 2016 United States presidential election. Considering it is once again an election year, and the technology has developed rapidly, this is more of an issue than ever before. A recent NBC News article from February reads "Russia's 2024 election interference has already begun" (De Luce, Collier). Just a few days later, CNN Politics put out an article titled "AI will allow more foreign influence operations in 2024 election, FBI director says", citing a talk the FBI Director Christopher Wray had with the Intelligence and National Security Alliance. He cites generative AI as a primary method of attack, and that AI reduces the barrier to interfere in our elections (Lyngaas). Adversaries can do this with higher fidelity deepfake voices, higher

quality deepfake video, and more believable fake stories and headlines than they could in the years prior. This is largely due to enhancements in machine learning training procedures involving synthetic data.

A similar hypothetical scenario that is relevant to current global political events could involve one side of conflict using synthetic images of bloody civilians' subject in war, to propagate whatever narrative they so desire. This would be irrespective of its accuracy. It could be as simple as training a generative AI model on past images of bloodshed. With the technology we have now, the model could produce hundreds of unique images that capture the same violence, except that all of it would be fake. Such a scenario is very feasible given the current global climate and given the technology readily available. Perhaps the scariest thought, captured by the Google DeepMind report, is that "the dissemination of synthetic data-driven misinformation can erode trust in legitimate information sources, making it increasingly difficult for people to distinguish between truth and falsehood" (Liu R, Wei J, Liu F, et al.). Having to ask oneself "Is this picture real? Or is it AI?" has become increasingly common. Scrutinizing an image to determine whether it is AI generated is almost futile now given how advanced our technology is. In an era where the lines between reality and synthetic creations blur, the fall from synthetic data-driven misinformation and unintended consequences loom large, exacerbating the problems of an already divided world.

The use of synthetic data undoubtedly has a lot of risk associated with it. What can be done about this? The United Nations University attempts to tackle this question. The United Nations University is the UN's global think tank, dedicated to addressing complex global challenges. They released a technology brief titled "The Use of Synthetic Data to Train AI Models: Opportunities and Risk for Sustainable Development", where they prescribe some

policy and technical recommendations about synthetic data. One technical recommendation they provide is to "use different types of generative AI models to create synthetic datasets" (Marwala, Fournier-Tombs, Stinckwich). This approach was used by Google Research and MIT CSAIL in their SynCLR model, as highlighted earlier. As research in this field grows, companies need to implement this into practice to ensure diverse datasets, which the UN University report echoes.

The most crucial policy recommendations offered by the UN University are to "Establish Global Quality Standards and Security Measures" and to "Clarify Ethical Guidelines, including Transparency" (Marwala, Fournier-Tombs, Stinckwich). Considering how new this field of research is, standardized safety measures and practices need to be put into place to ensure fair and accurate models. More organizations like the UN University need to release reports to spread the knowledge of the risks synthetic data poses. A practical idea that would greatly improve safety is to release an annual report on what went right and wrong in training AI models to be equitable, especially those trained with synthetic data. If this information is shared, research communities, startups, and corporations will have the best-practices to improve AI models for good.

Conclusion

Like other bleeding edge technologies, synthetic data is a double-edged that poses a wide variety of benefits and risks. It is a powerful tool in the current growing space of artificial intelligence. If used correctly, it has the capability to debias text classification models, something that could be used to improve companies' AI hiring tools. It can also be used to assist in the training of machine learning models, to account for unlikely scenarios and improve edge-case performance. However, the growing prevalence of generative AI makes it easier for malicious actors to use synthetic data for harm. This includes the examples we are already familiar with, like deepfake images, video, and audio to spread a narrative. But it also extends to foreign governments being able to perfect facial recognition systems, and influence elections by using AI generated media. Even if not being used for malicious purposes, there are still ethical considerations to keep in mind. Using synthetic data to try and correct machine learning models is not a trivial task. Attempts to correct a model can end up giving inaccurate and offensive output instead, as was seen in the Google Gemini example. Since research in this field is so new, we do not know the best ways to create fair models yet.

While we can be optimistic about the promise this technology has, we need to be cautious. The best way to move forward is to work with legislators and researchers together to develop guidelines, recommendations, and regulations that ensure ethical use. Shortcomings and successes alike need to be shared with the scientific community. This way, we can reap the benefits synthetic data has to provide, while minimizing risks and harm.

References

Data-Centric AI: AI Models Are Only as Good as Their Data Pipeline. Stanford HAI. Accessed October 27, 2023. https://hai.stanford.edu/news/data-centric-ai-ai-models-are-only-good-their-data-pipeline

De Luce D, Collier K. Russia's 2024 election interference has already begun. NBC News. Published February 26, 2024. Accessed April 14, 2024. https://www.nbcnews.com/news/investigations/russias-2024-election-interference-already-begun-rcna134204

Five ways IBM is using synthetic data to improve AI models. IBM Research Blog. Published February 9, 2021. Accessed September 28, 2023. https://research.ibm.com/blog/synthetic-data-explained

Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning. Gartner. Accessed October 27, 2023. https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning

Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit Med*. 2023;6(1):1-8. doi:10.1038/s41746-023-00927-3

Grant N. Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms. *The New York Times*. https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html. Published February 22, 2024. Accessed April 12, 2024.

Liu R, Wei J, Liu F, et al. Best Practices and Lessons Learned on Synthetic Data for Language Models. Published online April 11, 2024. Accessed April 12, 2024. http://arxiv.org/abs/2404.07503

Lyngaas S. AI will allow more foreign influence operations in 2024 election, FBI director says | CNN Politics. CNN. Published February 29, 2024. Accessed April 14, 2024. https://www.cnn.com/2024/02/29/politics/ai-2024-election-fbi-director/index.html

Raghavan P. Gemini image generation got it wrong. We'll do better. Google. Published February 23, 2024. Accessed April 12, 2024. https://blog.google/products/gemini/gemini-image-generation-issue/

Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., & Isola, P. (2023). Learning Vision from Models Rivals Learning Vision from Data (arXiv:2312.17742). arXiv. http://arxiv.org/abs/2312.17742

Toews R. Synthetic Data Is About To Transform Artificial Intelligence. Forbes. Published June 12, 2022. Accessed September 28, 2023. https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/

Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich, "The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development", UNU Technology Brief 1 (Tokyo: United Nations University, 2023).

University CM. Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women - Machine Learning - CMU - Carnegie Mellon University. Machine Learning | Carnegie Mellon University. Accessed October 27, 2023.

What is synthetic data? IBM Research Blog. Published February 8, 2023. Accessed October 27, 2023. https://research.ibm.com/blog/what-is-synthetic-data