Muhammad Hassan Nishat

PHL377

12/4/2023

Ethical Implications behind Synthetic Data

Machine learning technology has been exponentially growing in both popularity as well as in its ability within the last few years. With the release of ChatGPT and countless other AI tools to the public, AI is now more accessible than ever. One potential issue that arises with the continuing growth of AI, is the scarcity of data. AI algorithms are trained on massive amounts of data and sometimes the most difficult part of creating these AI algorithms is finding the data for the algorithm to be trained on. Without a large, detailed, and fully complete data set, AI algorithms cannot be made with high efficiency. As AI becomes more and more popular, more and more data sources will be needed to train these algorithms. This issue has been noticed by the AI industry and an interesting solution has been proposed. A proposed way to get around a lack of data, is to generate your own data using AI itself. Synthetic data refers to a large set of data created by AI for the purpose of training other AI. Synthetic data can be generated in the form of text, audio, or even visual data. Generating data with AI is a relatively new practice but is starting to be more regularly adopted. Although using synthetic data to train algorithms initially seems like a good solution to the scarcity of data, some ethical concerns can arise from using it. In this paper, I will argue that the use of synthetic data, in the current way it is generated, in training AI algorithms is unethical. Firstly, I will explain how synthetic data is generated so that there is a better understanding as to why I believe it causes ethical issues. Then I will go on to cover a few different ethical violations I believe to persist through the use of AI algorithms trained on synthetic data.

To understand why synthetic data is cause for ethical concern, it is important to understand how synthetic data is generated. There are two widely used algorithmic techniques in generating synthetic data, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). GANs are made up of two neural networks called the generator and the discriminator. The generator is fed random noise and learns to generate synthetic data from the noise. Real data as well as the synthetic data is then fed to the discriminator algorithm which attempts to classify the real and synthetic data. If the discriminator succeeds in classifying the synthetic data, the generator is updated to improve its output. If the discriminator fails in classifying either the real or synthetic data, it is also updated to increase its efficiency. Through this process the generator is pushed to output the most convincing results. VAEs consist of an encoder, which takes real data and compresses it into its important features and fills in the gaps with noise. This compressed image is passed onto a decoder which learns to reconstruct the real data from the compressed image. Once this algorithm is trained, it can be used to create detailed data when provided with a scarce and easy to obtain dataset. These kinds of synthetic data generating techniques have been particularly popular with computer vision solutions. One example of its implementation could be training software for self-driving cars. To achieve self-driving software, a large number of visual data is needed on all of the different scenarios that a car would need to be able to handle while driving. By training GANs or VAEs on a small amount of visual data of driving scenarios, they can be used to create more of these scenarios for the self-driving software to be trained on.

I believe that these kinds of synthetic data generation do not suffer from any ethical concerns that don’t already exist for machine learning and AI, such as the issues behind black box models and biased training data. These issues are legitimate concerns; however, it could be

argued that they can be addressed by those using these algorithms by ensuring the training data is unbiased and providing transparent explanations for the decisions made by the algorithms. Where I feel a new set of ethical concerns arise however, is with the potential of generating synthetic data with more generalized AIs such as large language models. The two implementations of synthetic data generation have been mostly used for specialized cases and a new algorithm would need to be trained for each use case. For example, an algorithm trained to create visual driving scenarios cannot be used to generate data to train an algorithm on identifying a specific object. The newer more generalizable image generating AI such as DALL-E is able to create images that could be used to train all kinds of AI algorithms. These kinds of generalized data generations are starting to be used in training AI as they provide a higher level of convenience and are suited to more use cases. I believe that the convenience of generalizability is also the main cause for ethical concern.

These newer generative models such as ChatGPT and DALL-E are probabilistic models that are trained on an extremely large data set of text or images taken from across the internet. The algorithms are able to take in a text prompt from the user and can determine what is most likely to be the response based on all of its training data. I believe that this causes a greater issue with bias than already exists with most standard AI algorithms. As these models are trained on a non-specific type of data, its results when asked to produce specific data are affected by a much different kind of bias. Not only is there a bias present within the real data that is trying to be replicated, but an unknown number of biases may be affecting the results from completely irrelevant datasets. For example, if ChatGPT was asked to create synthetic data by completing a survey a number of times for a study, it would be impossible to know what caused the algorithm to respond the way it does. Some information it was trained on that has nothing to do with

answering the survey could affect the algorithm and alter the results. For this reason, I believe that synthetic data, especially data created from generalizable AIs are cause for ethical concern.

Now that I have shown why believe the generation of synthetic data is potentially unethical, I will also discuss some of the ethical concerns that come with the use of synthetic data assuming that the data itself is created without bias. According to the article "Data Scarcity? Generative AI to the Rescue", written by Geoff Livingston, there are five areas of ethical consideration needed before the use of synthetic data in training AI. Firstly, Livingston mentions bias, which I will assume is not a concern for this section of the paper. The other issues Livingston brings up that he believes need to be addressed for ethical use of synthetic data are privacy, transparency, validation, and governance. Livingston states synthetic data should be private, meaning that we should not be able to trace synthetic data back to the identities of people in the real world. As well as this, synthetic data should be transparent in how it is generated and should be accurate. Finally, Livingston states that a clear and uniform set of governing policies should be enforced in the use of synthetic data. To address privacy, I do not believe that it should be an issue for synthetic data as there is no one real person to trace from the synthetic data. Transparency, as I have discussed above, is a difficult issue depending on how the data is generated. Models such as GANs or VAEs can be transparent by putting them through explainability softwares which simplifies the decisions made by the neural nets and provide explanations for how the results were derived. Technologies such as ChatGPT and DALL-E may be more of a challenge to make transparent due to the relatively large and broadness of the data they are being trained on, however, I believe as time goes on the explainability, and transparency of such models will continue to improve. Validation is another concern that is linked to bias. Synthetic data in its current form is quite accurate and is continuing to become more and more

accurate, I believe the issue is if we are able to reach a point at which synthetic data can be generated without any unwanted bias and still be accurate. Governance is an issue that I believe will be solved in time. As the use of synthetic data becomes more and more popular, it is the job of research institutions to ensure their ethics boards have a clear set of guidelines surrounding the use of synthetic data. Given these set of guidelines, I believe that synthetic data can be used ethically. However, I believe that the current way in which it is generated, may be cause for concern due to the immeasurable amount of unwanted bias that may be present.

In conclusion, synthetic data is data that can be generated in large amounts by AI, for training AI. Synthetic data can be generated through multiple ways, but it seems as though a useful way to create data could come in the form of recently made popular technologies such as ChatGPT and DALL-E as they are general and are ready to create data for any use case. I believe that this could cause potential ethical concerns as generating specific training data from AI that has been trained on a more general dataset could lead to unwanted biases in the generation of data that could then be passed onto the algorithm being trained on that synthetic data. Looking past the potential concern of unwanted bias, I believe that synthetic data can be implemented ethically and could serve as a useful solution to the growing problem of scarce data.

Works Cited

Amamou, Walid. “Creating Synthetic Data with CHATGPT Using NER.” *Medium*, UBIAI NLP, 8 Mar. 2023, https://medium.com/ubiai-nlp/entity-based-synthetic-data-generation-with-chatgpt-6344a28f0739.

Livingston, Geoff. “Data Scarcity? Generative AI to the Rescue.” *Evalueserve*, https://www.evalueserve.com/blog/data-scarcity-generative-ai-to-the-rescue/#:~:text=Many%20industry%2Dspecific%20or%20domain,%2C%20and%20consumer%2Dpacked%20goods.

“What Is Synthetic Data in Machine Learning and How to Generate It.” *V7*, https://www.v7labs.com/blog/synthetic-data-guide#h6.