Synthetic Data for AI Training

P B

V D

J F

K N

J P

S S

California State University, Fullerton

CPSC 315, Professional Ethics in Computing

Professor Harnick-Shapiro M.A

April 21, 2020

Executive Summary

In this paper we discuss the applications, ramifications, and future of synthetic data relative to its usage and creation. This is broken down by topics with the following:

- What is Synthetic Data and How is it Used?
- Synthetic Data and Copyright
- ACM Code of Ethics and the Implications for Synthetic Data
- Risks of Synthetic Data
- Current Regulation and News on Synthetic Data
- Modern Cases of Synthetic Data

Our goal is to use the the ACM code of ethics as well as the ethical viewpoints to look at the current level of regulations as well as the current use cases of synthetic data in order to present the reader with knowledge about where and why synthetic data is being used, as well as the risks and benefits using synthetic data has, especially as it relates to ethical ideals. This is done through looking at studies on the viability and reliability of synthetic data, and looking at the companies of Hazy and twentyBN. We will start by explaining what synthetic data is, then explaining where and why it is used. We will follow this with explaining the risks, both technical and ethical, and then analyze that relative to the ACM code of ethics and copyright law, looking at data ownership and privacy concerns.

Abstract

Artificial computer-generated data that mimics real data is called synthetic data. Synthetic data is not measured and collected from actual real-world situations or properties. Instead, the data copies user-specified parameters that remove private identifying aspects such as social security numbers and addresses so that the data becomes “anonymized”. However, as technology rapidly advances and improves, the gap between synthetic data and authentic data decreases. In fact, it ensures privacy which is the main reason for its rapid growth. However risks, regulations, copyright, and ethical use are one of the many concerns regarding synthetic data. These concerns need to be addressed to clear the roadblocks that are affecting the growth of synthetic data in AI training.

Introduction

As the modern era begins to rely on AI and data mining inorder to boost efficiency, the use and overall need of synthetic data has propelled its value alongside AI technology and data privacy relative to data mining. This creates new issues on the ethics of synthetic data currently being overlooked as it is spread far and wide as a powerful and necessary tool.

New developments in AI are only possible due to recent advancements in deep neural networks which allow us to generate highly accurate and shockingly realistic data. In order to respect and protect customer privacy, some of the largest organizations in finance, telecommunications, and healthcare are already relying on synthetic data (Siwicki, 2020). With such rising interest and popularity, it is worth researching the different specifications of synthetic data. Throughout our research, we will look deeper into the details of how synthetic data is used, the issue of copyright, ownership and fair use of the created data, current regulations, and the ethical risks and implications on the future of AI training.

What is Synthetic Data and How is it Used

Synthetic data can be both extremely helpful and useful for a multitude of things such as building machine learning models to predict cancer, data mining, and fraud detection systems. Synthetic data allows the user to generate valid and useful datasets that provide information into scenarios that may otherwise be unavailable or scarce to the user. Before getting too far ahead, what is synthetic data? Synthetic data is defined byMcGraw-Hill Dictionary of Scientific and Technical Terms as "any production data applicable to a given situation that are not obtained by direct measurement" (Synthetic Data, 2009, McGraw-Hill Education)

The best way to really understand what synthetic data really is, is through an example. Imagine a company wants to build a machine learning model that can accurately predict the weather. To make these predictions, the project lead on the project decides that they want to make these predictions by just asking for 2 inputs: the location and the date. The team of developers could spend the next couple of days or weeks searching for a dataset that has these two features in a large quantity, but it is not guaranteed to exist and could take too long, so the project lead decides that the team will synthetically create the data and train the machine learning model based off of that. This means that the team will programmatically generate data in a high enough volume that includes the needed features -- data and location.

It can be easy to think that synthetic data is just random made up data, but it is important to understand that this data was generated using specific algorithms and previous models that were designed specifically to mimic real data (Deng, Robert H.; Bao, Feng; Zhou, Jianying

2002). In fact, many researchers use synthetic data to test their developed frameworks as the synthetic data is “the only source of ground truth on which they can objectively assess the performance of their algorithms”(Jackson, C., Murphy, R., & Kovacevic, J., 2009).

Now that it is understood what synthetic data is, when would it be used? Well, as alluded to in the previous statement as well as in the weather example, synthetic data is used when either someone wishes to test their algorithms on objective data or the data available to them did not fit the criteria deemed necessary for the project. In other cases, data might be available that somewhat resembles what is needed, but it may be too time consuming to “clean” or remove data points that contain missing or strange values. It all boils down to the same thing that most things in the computer science world do -- truth, feasibility and convenience.

Modern Cases of Synthetic Data

There are several overarching use cases for synthetic data we noticed through research. The first use involves protecting data privacy by using synthetic data for data mining rather than actual customer data. Data mining can be used for both testing systems through simulation from synthetic data, as well as predicting market changes and consumer patterns through synthetic data. In both cases, using real data could lead to potential privacy violations, and both cases also require large volumes of data, as well as benefit from repeat testing simulating with different data sets. This also is a strong reason why in the field of privacy data synthesizing techniques have a strong market value. Another use case is for generating large data sets quickly and cheaply, no matter how obscure that data may be. Especially with the development of AI, newer or smaller companies who wish to use deep learning to train their AI systems require large data sets that they may not have access to due to their small size or recent history. A third artistic use case can also be observed as well.

TwentyBN is a data company that sells data sets to other companies. It started generating data sets through “crowd acting”, however this led to scalability issues which were supplemented through using Unity to create similar synthetic data (Westphal, 2018). Originally TwentyBN would ask civilians to take videos of mundane things (Westphal, 2018). By reaching out to a lot of people, the company could gather a video data set of the necessary type (one example might be room schematics) (Westphal, 2018). This helps create data sets where they do not already exist, as there was previously no demand for that exact data set. In order to help supplement the process of generating these data sets, the company began to create algorithms to mimic the

videos they originally had generated through crowd acting inside the Unity environment (Westphal, 2018). The entire approach leads to using synthetic data to do lower level deep algorithm training, when the system being trained is new and does not need precise real video, and then shifts this training from synthetic to real data as the system progresses. Software trained on solely synthetic data would have a chance to only be able to work properly for synthetic data, and wouldn't be capable of handling real world data as it never received such samples to train from. The accuracy on independent test datasets increased by 6% in comparison to training on pure real data without pre-training" (Westphal, 2018). While the viability of these data sets relative to completely real world data sets is still in the air, with the aid of synthetic data it becomes possible to generate larger data sets faster than in the real world (not sourced, common sense…).

Hazy is a synthetic data company that creates synthetic data entirely through deep learning. According to Hazys website it does this inorder to "[generate] smart synthetic data that's safe to use and actually works as a drop in replacement for real data science and analytics workloads". The website also notes that Chief Scientific Director Anthony Finklestien commented "Privacy preserving synthetic data, of the kind being pioneered by Hazy, is massively important with potential to reshape the data economy and beyond." This is a very different use case from TwentyBN. Instead of not possessing a data set to begin with, there is real world data to being with. The problem is that this data is tied to the individuals (customers) who produced that data, and in certain cases may be tied to private or otherwise confidential information. In order to data mine or otherwise manipulate the data without having to worry

about privacy concerns, one could create a similar data set synthetically that isn't tied to any real world people or private information.

This leads to the final use case of synthetic data, checkups. Google's Deep Dream was "designed to help us understand how neural networks work and what each layer has learned"(Alexander Mordvintsev, 2015). This is an extension of deepmind AI capable of generating synthetic data by essentially running in reverse through Deep Dream (Alexander Mordvintsev, 2015). The program can take an input of either text or a picture, and generate an accompanying picture or text. If inputting a picture, the software can generate a line of text describing that picture. By doing the opposite and giving it a line of text and asking for an image, the AI will generate synthetic data of its own. This has led to an unintended effect of creating pictures with strange patterns and colors, as seen in Figure 1.1. Google's Deep Dream software is currently open source.

Figure 1.1

This figure was created using the DreamScope App, which utilizes Google's Deep Dream software. Figure 1.1 can be considered a piece of synthetic data that was generated off of the original picture, real world data, with the purpose of artistic appeal. Synthetic data is created by computers and has an air of mystery into how it goes about doing so.

As synthetic data becomes more and more developed and widespread certain companies will grow the number of use cases and the overall emphasis their business model has on synthetic data like customers of Hazy, while newer companies will begin to pick up synthetic data in order to compete with larger corporations that possess their own data libraries like customers of TwentyBN. Using synthetic data to create art being a bonus on top of that with other possibilities still yet to arise with how synthetic data can be applied and furthered.

Synthetic Data and Copyright

Synthetic Data introduces some unique copyright issues. Creating synthetic data helps developers avoid copyright problems in using copyrighted datasets. Most datasets can be protected under copyright. Can synthetic data be copyrighted? Should the AI that created that dataset hold the copyright?

AI systems and their databases can be copyrighted to the developers that have created them. In 1991, the Copyright Society published an article titled Copyright Protection For Artificial Intelligence Systems, stating that "Our thesis is that AI systems face- and should face- no more difficulty than traditional computer programs in obtaining copyright protection" (Goldberg, 1991, p. 57). Goldberg and Carson state that all AI systems can be broken down into two basic pieces(Goldberg, 1991, p. 61). One being the program that runs it, and the other being the database(s) it uses (Goldberg, 1991, p. 62). The US copyright law covers computer programs under literary works, and protects the "expression of original ideas" not the code itself (Title, 1976). The database that the AI system uses is protected with copyright under literary works as well. Databases are considered as compilations, as long as it is an original grouping of data. Copyrighting AI systems is unique because copyright law does not extend to systems, as systems are considered an idea. Ideas cannot be protected under copyright law. AI systems are considered an expression of an idea, and that can be protected under copyright (Goldberg, 1991, p. 69). Goldberg and Carson state, "Copyright is especially appropriate for AI, where rapid technological advances would threaten to make any new statutory scheme obsolete almost as soon as it was enacted" (Goldberg, 1991, p. 74). Trade secrets would be too restrictive to new AI

systems, and not allow other developers to use the system with their own datasets. Patents may also be too restrictive for a developer's liking, but they still want their ideas protected. Protecting AI systems and their data allow for the greatest amount of protection and flexibility of use.

Using synthetic data eliminates copyright claims since the AI creates its own data to train with. In some cases, using copyrighted data for AI training is considered Fair Use as long as the copyrighted data is not used to create a copy of what was used (Levendowski, 2017, p. 582). Many large companies, such as Google, will release their AI toolkits as open source. However, they do not release the datasets used to create the toolkit (Levendowski, 2017, p. 582). Synthetic data would allow for each developer to generate and use their own copyright free set of data. This data can also be better quality, since it is not trained from biased sources (Levendowski, 2017, p. 583). The higher quality and variety of data being used will make for higher quality AI systems.

Since AI systems can be copyrighted, synthetic data created by an AI should be protected under copyright as well. The dataset created is new, and not part of any existing dataset. The expression of the AI training is also protected under copyright as well. The copyright would be held by the developers or the company that created the AI system that creates the data. But, if the AI created the data, could it hold the copyright to that data?

In the Journal of Internet Law, the article "Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artifical Intelligence Systems", written by Ana Ramalho brings up several arguments for and against AI being able to protect its own work (Ramalho, 2017). First, the copyright system is not prepared to grant legal protection to non-human entities (Ramalho, 2017). Authorship in all references of copyright law, both in the

United States and throughout the world, has in common “a human being who exercises subjective judgement in composing the work and who controls the execution” (Ramalho, 2017). Synthetic AI may have created its own dataset, but the developers who wrote the code are the ones who set the AI on its path. It is their algorithms that cause the AI to start on its work, and who thought up the expected goal of the training. Secondly, the copyright office has also declined to register “works produced by nature animals, plants and neither by “machines or mere mechanical processes that operate randomly or automatically without any creative input or intervention from a human author”” (Ramalho, 2017). Synthetic data would not fall under this registration, because it can be considered under creative input and intervention from its developers. The developers would be the ones to hold the copyright for the AI system and its self-created dataset. The AI system is a tool the developers use to create the dataset, even if they didn’t specifically enter the data. Lastly, extending copyright to an AI system itself can also raise questions about what other legal standings AI should be given (Ramalho, 2017). If the AI system creates the synthetic data and owns it, should that AI also own the profits from the use of the data? Could users get financial judgements against the AI for damages from mistakes in the data? Giving AI it’s own copyright for its synthetic data could open up a whole host of legal problems. These legal problems are unnecessary when the copyright can be given to the developers who created and control the AI that created the synthetic data.

Although synthetic data has some unique copyright issues, many of them can be solved by keeping the data under copyright by the developers who created the AI. Copyright protection is already given for AI systems and their databases that express original ideas. Since the legal

and ethical consideration for AI to own its own copyright have not been worked out, synthetic data would remain copyrighted to the developers who created it.

ACM Code of Ethics and the Implications for Synthetic Data

While using synthetic data eradicates third-party copyright claims, it lies in the gray area when it comes to the ACM Code of Ethics. The ACM Code of Ethics lists seven general ethical principles. According to the Code of Ethics, a computing professional should: contribute to society and to human well-being, avoid harm, be honest and trustworthy, be fair and take action not to discriminate, respect the work required to produce new ideas, respect privacy, and honor confidentiality (ACM Code of Ethics and Professional Conduct, 2016). Since synthetic data is generated by the AI itself, it is impossible to predict the outcome of using that data to train models. This could lead to harm to humans and/or society depending on the usage of the data and could potentially violate the ACM Code of Ethics.

Avoiding harm is another general principle of the Code of Ethics. The Code of Ethics refers to "harm" as negative consequences, including unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. "Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions" (ACM Code of Ethics and Professional Conduct, 2016).

While it is true that synthetic data does not infringe on any individual's privacy, it is still important to understand the risks of the accuracy of the predictions made by the model that was trained by that data. In many cases, models are trained by using 60-80% of the known data. The rest of the 20-40% of the known data is used to test the model and calculate the model's average accuracy. But when the known data is synthetically generated by AI, this accuracy could be

jeopardized. This means that the model that was trained may be unreliable, and if the model is used for any predictions that would affect society, then it could have negative consequences.

Since synthetic data can have positive effects as well, we can apply some of the workable ethical theories to it. Kantianism is an ethical theory that claims motives and universal rules are important aspects in judging what is wrong and right. According to Kantianism we can argue that if the end goal of using synthetic data for training models is intended to have positive consequences then it is morally ethical.

Current Regulation and News on Synthetic Data

We fear that one day, just as sci-fi movies have predicted, robots might take over the world and enslave humanity. Some in the tech industry are saying that the technology is not advanced enough (Modius, 2012). However, some argued that artificial intelligence technology needs to be regulated to help manage and hopefully mitigate the risks (Jee, 2019).

Since AI training depends on a vast amount of data, we have seen the popularity of synthetic data rise due to its ability to meet the requirement of AI training (Nisselson, 2018). Synthetic data is not real and is mostly generated artificially by a computer program. There are many uses of a(Macaulay, 2019). There is plenty of data generating on this planet every minute. Smaller companies, especially startups, cannot collect the data from the real world to train prototypes to make models because synthetic data is not as risky and costly as original data(Marr,2018).Therefore, synthetic data is a great substitute of original data.

Data privacy is another reason for using synthetic data. Privacy can be enhanced by removing the key information from data. However, this limits data utility. Instead, it can synthesize data that replicates a protected dataset for analytical purposes but is less likely to reveal any private or sensitive information. (Wible, 2019, p.349). According to an article in MIT news, it is very challenging to produce high-quality synthetic data especially if the system is complex. However, in that same article, Koperniak goes on to suggest that synthetic data is also a valuable tool for educating students (Koperniak, 2017, para.10). Therefore, we can say that it is very reasonable to worry about using synthetic data in AI training and create more laws and regulations. However, regulations beyond those that already exist can be a barricade to the development of AI that could be overwhelmingly beneficial.

A report published on Norway's local government website states that The Freedom of Information Act regulates how public data should be made available for reuse. Since 2012, Norway has required government agencies and digital services to make data in machine-readable formats. (Ministry of Local Government and Modernisation, 2019, p.14). The primary take away from the report was that government agencies should arrange for data to be available in the long term with an emphasis on data integrity, authenticity, usability, and reliability.

Therefore, regulations that make it more expensive to develop AI or prevent certain uses could significantly delay the progress being made towards projects in those respective fields -- especially by smaller business owners -- as it adds more loops that these companies may simply not be able to afford to jump through.

Risks of Synthetic Data

As synthetic data and forms of AI such as machine learning begin to merge, the pace of change will also increase rapidly. According to the NewVantage Partners' annual Big Data Executive Survey published in early 2017, "88.5% of top executives surveyed believe that AI is first among all new capabilities that will have a disruptive impact on their companies" (Bean, 2017). Major corporations such as American Express, Capital One, Disney, Ford Motors, General Electric, and JPMorgan Chase were represented in this survey. With big corporations believing that AI will pose possible risks of disruption or displacement, the traditional ways of doing business may dramatically change. Corporations could be faced with the problem of responding effectively to such dynamic changes or run the risk of lagging behind in the competitive marketplace. Some corporations like American Express are already taking a step ahead and proving their capabilities by recreating themselves and offering new products and services. This introduces new paths for innovation. These large corporations need to learn and adapt to sudden shifts by taking advantage of big data and the emerging capabilities of machine learning. Not only will the increased use of synthetic data for AI development affect the nation's biggest corporations, it may also pose long-term safety risks.

In some cases, synthetic data will not always be the perfect solution. Most datasets are too complex to fabricate correctly (Peng, T., & Telle, A., 2018). Additionally, synthetic data will tend to be generated with biases from the original data. There is a possibility that a model could create very unrealistic expectations if trained with synthetic data because it may have performed better than a model trained with realistic data. As a result, "the models built from synthetic data

can be used to make biased and unreliable decisions" (Gonfalonieri, 2020) . The bias element in AI influences strong debate around the topic of whether or not AI is ethical. There is a potential risk that the decisions resulting from the model could be biased towards a particular religion, a section of society, race, and even gender as well. For instance, if a model was built upon training data that was based on a historic data set showing more money was lent to a particular gender, then the lending decision outcome of the model will most likely be the same. Hence, machine learning algorithms and its decisions could cause real-world harm and affect people's lives. One way to overcome this problem is to build tools that could detect and remove bias in machine learning models. However, this could be a very long and tedious process. The challenge is in assuring the accuracy of synthetic data and making sure that its statistical properties are similar to real data.

Conclusion

In this paper, we have attempted to explain multiple areas of synthetic data generation such as the use cases, how it is covered in copyright laws, the ethics of using and generating data, the risks, and regulation regarding the use of synthetic data. In particular, we looked at when and why synthetic data is used when it comes to training machine learning models, how synthetic data generation allows you to avoid copyright infringement, where synthetic data lies on the ethical scale, and how synthesizing data can be risky when training machine learning models.

Although synthetic data has the possibility to create flawed machine learning models and is ethically in a gray zone, the expressiveness and openness that is created by the use of synthetic data allows the furtherment and development in multiple fields and is generally a good metric to test on -- so long as it is used in a responsible and informed way.

References

Acm.org. 2016. *ACM Code Of Ethics And Professional Conduct*. [online] Available at: <http://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct> [Accessed 14 March 2020].

Bean, R. (2017, May 8). How Big Data Is Empowering AI and Machine Learning at Scale. Retrieved from https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/

Deng, R. H., Feng, B. H., & Zhou, J. H. (2002). A Synthetic Fraud Data Generation Methodology . Information and Communications Security, 265–277. Retrieved from https://books.google.com/books?id=GL1sCQAAQBAJ

Goldberg, M.; Carson, D.O. (1991). Copyright Protection for artificial intelligence systems. Journal of the Copyright Society of the U.S.A., 39(1), 57-75.

Gonfalonieri, A. (2020, January 6). Do You Need Synthetic Data For Your AI Project? Retrieved from https://towardsdatascience.com/do-you-need-synthetic-data-for-your-ai-project-e7ecc2072d6b

Jackson, C., Murphy, R., & Kovacevic, J. (2009). Intelligent Acquisition and Learning of Fluorescence Microscope Data Models. IEEE Transactions on Image Processing, 18(9), 2071–2084. doi: 10.1109/tip.2009.2024580

Jee, C. (2019, February). Artificial-intelligence development should be regulated, says Elon Musk. Retrieved May 1, 2020, from

https://www.technologyreview.com/2020/02/19/906156/artificial-intelligence-development-should-be-regulated-says-elon-musk/

Koperniak, S., & Institute for Data. (2017, March 3). Artificial data give the same results as real data - without compromising privacy. Retrieved April 13, 2020, from http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. Washington Law Review, 93(2), 579–630

Macaulay, Thomas. "What Is Synthetic Data And How Can It Help Protect Privacy?" *Techworld*, 1 Oct. 2019, www.techworld.com/data/what-is-synthetic-data-how-can-it-help-protect-privacy-3703127/.

Marr, B. (2018, November 6). Does Synthetic Data Hold The Secret To Artificial Intelligence? Retrieved from https://www.forbes.com/sites/bernardmarr/2018/11/05/does-synthetic-data-hold-the-secret-to-artificial-intelligence/#21b4c94b42f8

Ministry of Local Government. (n.d.). The National Strategy for Artificial Intelligence. Retrieved April 13, 2020, from https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=4

Modis, T. (2012). Why the Singularity Cannot Happen. *The Frontiers Collection Singularity Hypotheses*, 311–346. doi: 10.1007/978-3-642-32560-1_16

Mordvintsev, Alexander, et al. “DeepDream - a Code Example for Visualizing Neural Networks.” Research Blog, Google, 1 July 2015, web.archive.org/web/20150708233542/googleresearch.blogspot.co.uk/2015/07/deepdream-code-example-for-visualizing.html.

Nisselson, E. (2018, May 11). Deep learning with synthetic data will democratize the tech industry. Retrieved May 1, 2020, from https://techcrunch.com/2018/05/11/deep-learning-with-synthetic-data-will-democratize-the-tech-industry/

Peng, T., & Telle, A. (2018). A tool for generating synthetic data. Proceedings of the First International Conference on Data Science, E-Learning and Information Systems - DATA 18. doi: 10.1145/3279996.3280018

Ramalho, A. (2017). Will Robots Rule the (Artistic) World?: A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems. Journal of Internet Law, 21(1), 1–25.

Recent News. (n.d.). Retrieved from https://www.exactdata.net/recent-news.html

Siwicki, B. (2020, March 6). Is synthetic data the key to healthcare clinical and business intelligence? Retrieved from https://www.healthcareitnews.com/news/synthetic-data-key-healthcare-clinical-and-business-intelligence

Synthetic Data. (2009). McGraw-Hill Dictionary of Scientific and Technical Terms. New York: McGraw-Hill Education.

"Synthetic Data That Works." Synthetic Data by Hazy: Differential Privacy Meets Utility., hazy.com/.

Title 17 of the United States Code (2016)

Westphal, S. (2018, January 12). Using synthetic data for deep learning video recognition. Retrieved from https://medium.com/twentybn/using-synthetic-data-for-deep-learning-video-recognition-49be108a9346

Wible, B. (2019, April 26). Synthetic data, privacy, and the law. *Science Magazine*, *364*(6438), 348–349. doi: 10.1126/science.364.6438.348-g