When Synthetic Data Met Regulation

Georgi Ganev [1] [2]

1. Motivation

Generative AI has made significant progress recently, with applications spanning text, code, image, video, speech, and structured data (Sequoia Capital, 2022). Investor interest has also grown – start-ups received $2.2B in 2022 (TechCrunch, 2023b) and Microsoft reportedly invested $10B in OpenAI's ChatGPT (Bloomberg, 2023), which has reached 100M monthly users (Reuters, 2023). However, concerns about privacy, robustness, copyright, and compliance have increased as well. Active legal cases against Generative AI companies and products (TechCrunch, 2023a) have led some organizations and countries, such as Italy, to (temporarily) restrict ChatGPT usage (CNN, 2023; Politico, 2023).

Synthetic Data. In this paper, we focus on synthetic data, a subfield of Generative AI that utilizes generative machine learning models such as GANs (Goodfellow et al., 2014), Diffusion Models (Sohl-Dickstein et al., 2015), and Transformers (Vaswani et al., 2017), albeit typically at a smaller scale. We opt for tabular data comprising sensitive information as training data as it is still the most extensively used data type in large enterprises. Furthermore, synthetic data is comparatively more established and has recently been examined by reputable organizations (Royal Society, 2023; UN, 2023) and regulators (ICO UK, 2022; FCA UK, 2023), alas without any definitive compliance directives.

Main Question. This prompts the question: *"Can we make synthetic data regulatory compliant?"* Namely, we explore the legality of privacy-preserving synthetic data created by generative models trained on structured personal data.

2. Regulatory Definitions

Personal Data. EP and Council (2016a) define personal data as "any information relating to an identified or identifiable living individual" and the latter as someone who can be identified (directly or indirectly) by reference to factors such as name, id number, or physical, genetic, social identity, etc. On the other hand, information that is effectively anonymized is not personal data and data protection law does not apply to it (EP and Council, 2016b). But in practice the actual identifiability of individuals can be highly context-specific as different types of information carry different levels of identifiability risks depending on the circumstances. Clearly, creating synthetic data based on sensitive personal data requires processing it. However, whether the resultant synthetic data constitutes personal or anonymous information is a question to be determined based on an assessment of the identifiability risk. This raises the question, what constitutes a sufficient level of anonymization.

Sufficient Anonymization. ICO UK (2021) states that "effective anonymization reduces identifiability risk to a sufficiently remote level." When assessing whether someone is identifiable, objective factors to be considered include the cost and time required to identify, the available technologies, and their developments over time. However, not every hypothetical/theoretical chance of identifiability needs to be taken into account. The focus should be on what is reasonably likely to be used relative to the circumstances, not in absolute. This is consistent with A29WP (2014)'s approach, that also notes that data controllers should regularly reassess the attending risks. In terms of technical analysis, A29WP (2014); ICO UK (2021) assert that the following three key risks need to be reduced for sufficient anonymization:

1. (*singling out*) any individual being isolated;
2. (*linkability*) any records/datasets (publicly available or not) being combined with synthetic data and thereby enabling the identification of an individual;
3. (*inferences*) an attribute being deduced with significant probability from the values of other attributes.[1]

ICO UK (2021) explains that the three risks should be looked through the *motivated intruder* test – a competent intruder having access to appropriate resources being able to achieve identification if they were motivated to attempt it.

3. Synthetic Data as Anonymous Data

In this section, we show that producing synthetic data by combining two techniques—generative models and Differential Privacy (DP)—reduces all identifiability risks to suffi-

[1] Hazy, London, UK [2] University College London, London, UK. Correspondence to: Georgi Ganev <georgi.ganev.16@ucl.ac.uk>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

[1] This is in direct contradiction with good quality synthetic data and has led to leading privacy researchers abandoning statistical inference as privacy violation (McSherry, 2016; Bun et al., 2021).

ciently remote level and, therefore, the resulting data can be considered anonymous per (A29WP, 2014; ICO UK, 2021). Overall, we rely on generative models to create high utility synthetic data and DP to provably guarantee privacy.

Generative Models break the 1-to-1 mapping and to an extent reduce singling out and linkability but could be susceptible to various privacy attacks (see below).

The process of training a generative model to learn the probability distribution of the input sensitive data, discarding it, and sampling from the fitted parameters to create new (synthetic) data, naturally lowers some privacy concerns. For instance, it breaks the 1-to-1 mapping from a single real record to a single synthetic one which makes singling out difficult. Since the models are probabilistic in nature, they capture the inherent data uncertainty and variability, which reduces linkability. Furthermore, launching adversarial privacy attacks versus generative models is more challenging compared to discriminative ones (De Cristofaro, 2021).

However, some generative models could occasionally memorize records and reproduce them (exactly or approximately) in the synthetic data (Carlini et al., 2019; van den Burg & Williams, 2021). In turn, a strategic adversary with side knowledge (e.g., the training algorithm, representable data, etc.) could infer the presence of these records (Hayes et al., 2019; Chen et al., 2020; Stadler et al., 2022), thus violating the linkability test and rendering the synthetic data pseudonymous at best or personal at worst (López & Elbi, 2022). Even more powerful privacy attack is reconstruction (Carlini et al., 2021; 2023), in which the adversary manages to recover whole training records and, therefore, leaks all of their private attributes.

DP mechanisms formally protect against singling out, linkability, and other re-identifiability concerns even if faced with a resourceful and strategic adversary (see below).

DP (Dwork et al., 2006; Dwork & Roth, 2014) is a mathematical definition of privacy which formally bounds the probability of distinguishing whether any given individual's data was included in the input data. The level of indistinguishability is controlled and quantified by a parameter, ϵ, or the privacy budget. In the context of Generative AI, DP is usually satisfied by training the models with noisy/random mechanisms and frameworks such as DP-SGD (Abadi et al., 2016) and PATE (Papernot et al., 2017; 2018).

Since DP makes the trained model indistinguishable, whether any individual's data was included or not, it averts memorization and singling out. The protection against GDPR's singling out has been robustly formalized (Cohen & Nissim, 2020) (Nissim et al. (2017) also argue DP satisfies FERPA requirements). Additionally, DP defends against potential harms, such as linkability, that could be caused by the publication of other sensitive information. Stadler et al. (2022) show this holds true even for outliers or potentially the most vulnerable individuals who have a higher chance of being memorized (Feldman, 2020). Furthermore, DP does not make any assumptions about the adversary and the provable mathematical guarantees apply in the worst-case scenario (e.g., the attacker has prior information, knowledge of the training algorithm, strong computing power, etc.) which means that DP protects against motivated adversaries. The protections are not just theoretical, DP reduces all key risks empirically, too (Giomi et al., 2022).

Using DP-trained models makes privacy an attribute of the generating process rather than a given synthetic dataset. Thanks to its resistance to post-processing property, DP allows reusing models (to generate data) without further privacy leakage. This means that even in the unlikely scenario in which a synthetic record very similar to a real is generated (which could be dissatisfactory (ONS UK, 2018)), it does not constitute a privacy violation (Jordon et al., 2022).

Potential Limitations. While DP offers robust privacy protection, in certain scenarios it could be too conservative (Nasr et al., 2021). Furthermore, DP often leads to utility reduction, particularly impacting outliers and underrepresented subgroups (Stadler et al., 2022; Ganev et al., 2022) and causing inconsistencies (Kulynych et al., 2023). Selecting both the right privacy budget and DP mechanism is non-trivial and highly context-specific (Hsu et al., 2014; Ganev et al., 2023). Lastly, implementing DP in practice and effectively conveying its properties can be challenging/complex (Cummings et al., 2021; Houssiau et al., 2022).

Related Work. Cummings et al. (2023) discuss further DP benefits/challenges/open questions and Jordon et al. (2022); De Cristofaro (2023) focus on combining synthetic data with DP (also advised by (Bellovin et al., 2019)). Specific (DP) generative models include GANs (Xie et al., 2018; Jordon et al., 2018; Xu et al., 2023), Diffusion Models (Kotelnikov et al., 2022; Ghalebikesabi et al., 2023), and Transformers (Borisov et al., 2022; Solatorio & Dupriez, 2023).

4. Future Work

In this paper, we argue that synthetic data produced by DP generative models can be sufficiently anonymized and, therefore, anonymous data and regulatory compliant. Our work aims to establish a foundation for broader Generative AI solutions. Nevertheless, they face added obstacles, such as training on vast multi-modal datasets that may include proprietary/copyrighted data with commercial usage limitations. Moreover, as datasets are often distributed over the internet, it becomes increasingly difficult for individuals to assert their right to consent or be forgotten. Factors like data accessibility (e.g., decentralized/scraped data), governance, robustness, transparency, explainability, and fairness must also be considered (Gal & Lynskey, 2023; IAPP, 2023).

Acknowledgements

We would like to thank Diana Sofronieva for helping with the structure, clarity, and rigorousness of the paper; Ian Stevens, Samikah Ahmed, and Suzanne Jopling for providing legal expertise and consultation; Emiliano De Cristofaro and Meenatchi Sundaram Muthu Selva Annamalai for providing technical feedback; Adriano Basso, Andrew Keen, and Gareth Rees for reviewing earlier versions of the paper and offering overall support; Orla Lynskey and James Jordon for the helpful discussions; as well as the anonymous reviewers for their encouraging and helpful comments.

References

A29WP. Opinion on anonymisation techniques. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *ACM CCS*, 2016.

Bellovin, S. M., Dutta, P. K., and Reitinger, N. Privacy and synthetic datasets. *STLR*, 2019.

Bloomberg. Microsoft Invests $10 Billion in ChatGPT Maker OpenAI. https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai, 2023.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. *arXiv:2210.06280*, 2022.

Bun, M., Desfontaines, D., Dwork, C., Naor, M., Nissim, K., Roth, A., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. Statistical Inference is Not a Privacy Violation. https://differentialprivacy.org/inference-is-not-a-privacy-violation/, 2021.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security*, 2021.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv:2301.13188*, 2023.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: a taxonomy of membership inference attacks against generative models. In *ACM CCS*, 2020.

CNN. Don't tell anything to a chatbot you want to keep private. https://edition.cnn.com/2023/04/06/tech/chatgpt-ai-privacy-concerns/index.html, 2023.

Cohen, A. and Nissim, K. Towards formalizing the GDPR's notion of singling out. *PNAS*, 2020.

Cummings, R., Kaptchuk, G., and Redmiles, E. M. ”I need a better description”: an investigation into user expectations for differential privacy. In *ACM CCS*, 2021.

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Jagielski, M., Huang, Y., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., Xiong, L., Yekhanin, S., Yu, D., Zhan, H., and Zhang, W. Challenges towards the Next Frontier in Privacy. *arXiv:2304.06929*, 2023.

De Cristofaro, E. A critical overview of privacy in machine learning. *IEEE S&P*, 2021.

De Cristofaro, E. What Is Synthetic Data? The Good, The Bad, and The Ugly. *arXiv:2303.01230*, 2023.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

EP and Council. Article 4 GDPR Definitions. https://gdpr-info.eu/art-4-gdpr/, 2016a.

EP and Council. Recital 26 EU GDPR. https://www.privacy-regulation.eu/en/recital-26-GDPR.htm, 2016b.

FCA UK. Synthetic data call for input feedback statement. https://www.fca.org.uk/publication/feedback/fs23-1.pdf, 2023.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *STOC*, 2020.

Gal, M. and Lynskey, O. Synthetic Data: Legal Implications of the Data-Generation Revolution. *109 Iowa Law Review*, 2023.

Ganev, G., Oprisanu, B., and De Cristofaro, E. Robin Hood and Matthew Effects: Differential privacy has disparate impact on synthetic data. In *ICML*, 2022.

Ganev, G., Xu, K., and De Cristofaro, E. Understanding how Differentially Private Generative Models Spend their Privacy Budget. *arXiv:2305.10994*, 2023.

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., and Balle, B. Differentially Private Diffusion Models Generate Useful Synthetic Images. *arXiv:2302.13861*, 2023.

Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B. A unified framework for quantifying privacy risk in synthetic data. In *PETs*, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NIPS*, 2014.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: membership inference attacks against generative models. In *PoPETs*, 2019.

Houssiau, F., Rocher, L., and de Montjoye, Y.-A. On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data. *Nature Communications*, 2022.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. Differential privacy: an economic method for choosing epsilon. In *IEEE CSF*, 2014.

IAPP. Generative AI: Privacy and tech perspectives. https://iapp.org/news/a/generative-ai-privacy-and-tech-perspectives/, 2023.

ICO UK. Chapter 2: how do we ensure anonymisation is effective? https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf, 2021.

ICO UK. Chapter 5: privacy-enhancing technologies (PETs). https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf, 2022.

Jordon, J., Yoon, J., and Van Der Schaar, M. PATE-GAN: generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data–what, why and how? *arXiv:2205.03257*, 2022.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. TabDDPM: Modelling Tabular Data with Diffusion Models. *arXiv:2209.15421*, 2022.

Kulynych, B., Hsu, H., Troncoso, C., and Calmon, F. P. Arbitrary decisions are a hidden cost of differentially-private training. In *ACM FAccT*, 2023.

López, C. A. F. and Elbi, A. On the legal nature of synthetic data. In *NeurIPS SyntheticData4ML*, 2022.

McSherry, F. Statistical inference considered harmful. https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md, 2016.

Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: lower bounds for differentially private machine learning. In *IEEE S&P*, 2021.

Nissim, K., Bembenek, A., Wood, A., Bun, M., Gaboardi, M., Gasser, U., O'Brien, D. R., Steinke, T., and Vadhan, S. Bridging the gap between computer science and legal approaches to privacy. *Harvard JOLT*, 2017.

ONS UK. Privacy and data confidentiality methods: a data and analysis method review. https://analysisfunction.civilservice.gov.uk/policy-store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr/, 2018.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with pate. In *ICLR*, 2018.

Politico. Italian privacy regulator bans ChatGPT. https://www.politico.eu/article/italian-privacy-regulator-bans-chatgpt/, 2023.

Reuters. ChatGPT sets record for fastest-growing user base. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, 2023.

Royal Society. From privacy to partnership: the role of PETs in data governance and collaborative analysis. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf, 2023.

Sequoia Capital. Generative AI: A Creative New World. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Solatorio, A. V. and Dupriez, O. REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. *arXiv:2302.02041*, 2023.

Stadler, T., Oprisanu, B., and Troncoso, C. Synthetic data – anonymization groundhog day. In *Usenix Security*, 2022.

TechCrunch. The current legal cases against generative AI are just the beginning. https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/, 2023a.

TechCrunch. VCs continue to pour dollars into generative AI. https://techcrunch.com/2023/03/28/generative-ai-venture-capital/, 2023b.

UN. The United Nations Guide on privacy-enhancing technologies for official statistics. https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf, 2023.

van den Burg, G. and Williams, C. On memorization in probabilistic deep generative models. *NeurIPS*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 2017.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv:1802.06739*, 2018.

Xu, K., Ganev, G., Joubert, E., Davison, R., Van Acker, O., and Robinson, L. Synthetic data generation of many-to-many datasets via random graph generation. In *ICLR*, 2023.