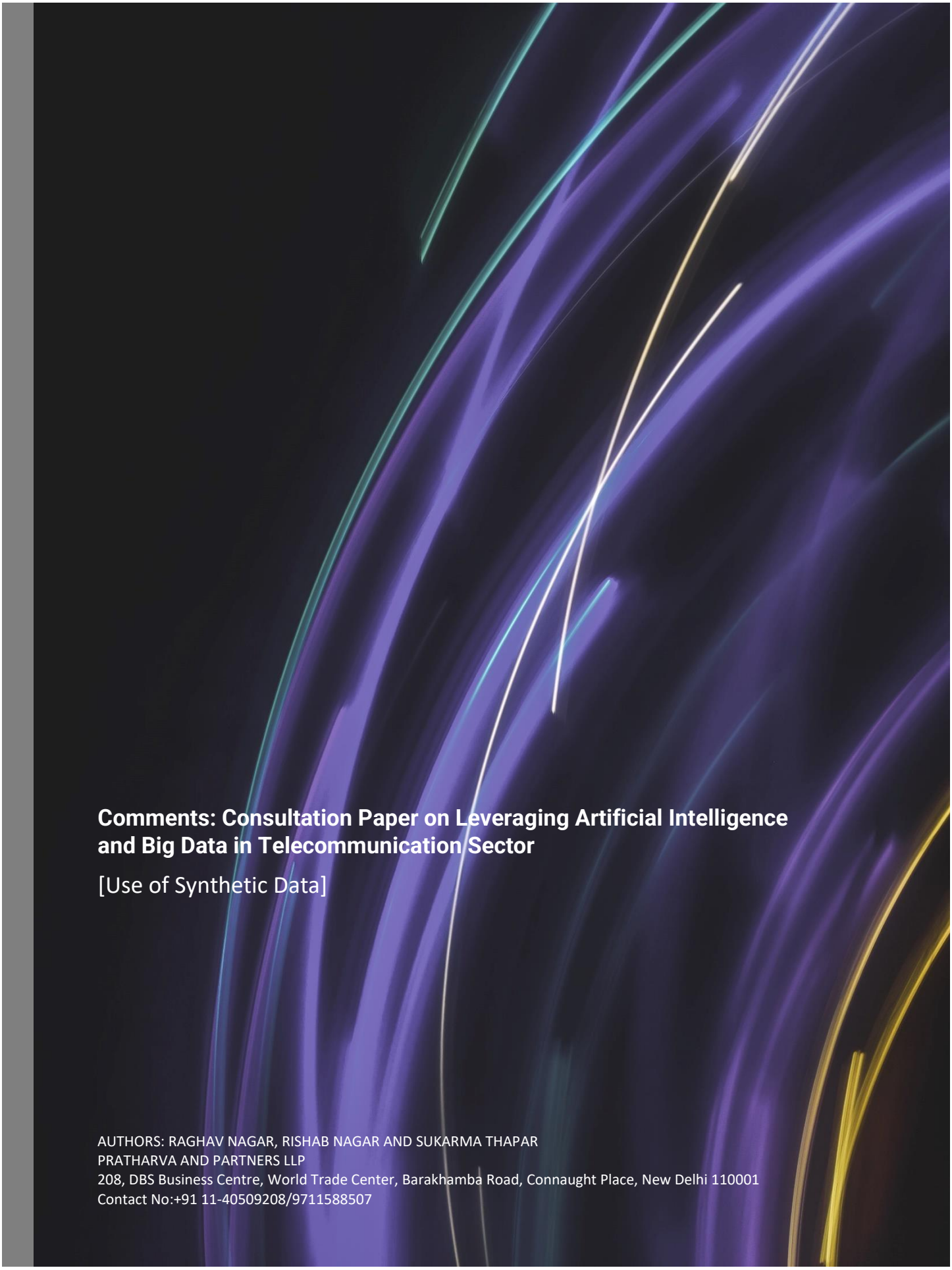

Comments: Consultation Paper on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector

[Use of Synthetic Data]

AUTHORS: RAGHAV NAGAR, RISHAB NAGAR AND SUKARMA THAPAR
PRATHARVA AND PARTNERS LLP
208, DBS Business Centre, World Trade Center, Barakhamba Road, Connaught Place, New Delhi 110001
Contact No:+91 11-40509208/9711588507

Re: Suggestions/feedback on the Consultation Paper on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector

The present note refers to Q.17 framed by Telecom Regulatory Authority of India as part of the Consultation Paper on "Leveraging Artificial Intelligence and Big Data in Telecommunication Sector", framed as below:

"Q.17. Whether the authority or body or institution as suggested in response to Q.11 may also be entrusted with the task to manage and oversee collection, cataloguing and storage of data? Whether such authority or body or institution need to be entrusted to generate and make available ***synthetic data****? Please justify your response with rationale and suitable examples, if any."*

In absence of a legal framework regarding protection of personal data and non-personal data, vesting a body with powers to oversee collection, cataloguing and storage of data may be a tall ask. One aspect to be addressed is the role of synthetic data in ensuring data privacy, i.e., protecting privacy of user data. Entrusting a body or institution with responsibility to generate and make available synthetic data requires an insight into the privacy concerns involved in generation of such data. Our brief comments on the same is as following:

i. In essence a 'fake' data as opposed to a 'real' data, synthetic data can enable machine learning by offering datasets with the similar statistical attributes as the original data. While de-identification takes away the utility, synthetic data is considered a viable privacy preserving technology for use in government sectors, health sector and software projects by offering a reasonably representative data and shareable data while ensuring privacy at the same time.

ii. A simulation of real-world individuals can be generated without actual reference to their personally identifiable information. This method employs data anonymization at the first stage and then carries out creation from the anonymized data.

iii. The de-anonymization/identification of personal information is the primary area of friction and concern in formulation of privacy laws and regulations. Synthetic data as a subset of anonymized data can do away with the possibility of identification as it does not refer to any particular real-world individual. It is viewed as offering indistinguishable results when compared with analysis undertaken on original data.

iv. Even though synthetic data does not have a 1:1 relationship with real world data, the data is generated based on mobilization of the original dataset which can comprise of personal as well as non-personal data. However, it cannot be viewed as “silver bullet” data.

v. The fidelity of synthetic data vests on the statistical similarity it offers to original data and thus synthetic data must be vetted after generation to check if it overfits the original data and refers to personal data. Addition of differential privacy measures is suggested on synthetic data to avoid risk of data leakage.

vi. By using additive techniques in place of subtractive ones, synthetic data aids the data-privacy problem by producing statistically nearly-identical replica of original data. This does not change the fact that synthetic data reshapes the original data and raises concerns over the status of protection on the original dataset.

vii. Given the privacy-security risks, the system of check for verification of overfit of synthetic data as discussed above requires that a law be put in place to protect the use of personal data.

viii. The authority or body or institution as described by the subsection 5.3.4 of the consultation paper must ascribe the standards for formulation of a synthetic data. Aspect of “leakage” must be addressed by the body before mobilizing the original datasets.

ix. We suggest the formulations of rules to process original data, which includes personal data. We also suggest that a deeper study into the impact of using synthetic data on privacy of personally identifiable information must be carried out.

References:

1. Synthetic data and data protection, by Frederico Marengo, Synthetic data and data protection | Qubit Privacy.
2. Synthetic data, by Robert Reimann, Synthetic Data | European Data Protection Supervisor (europa.eu).
3. Synthetic Data: Applications in Data Privacy and Machine Learning, by Lulu Tan, Synthetic Data: Applications in Data Privacy and Machine Learning | by Lulu Tan | Towards Data Science.
4. Privacy and Synthetic Datasets, by Steven M. Bellovin,Preetam K. Dutta and Nathan Reitinger, 22 STAN. TECH. L. REV. 1 (2019), ARTICLE (stanford.edu).
5. What Synthetic Data Means to Information Privacy, CaseGuard, Synthetic Data, Privacy, Computer Programs, Technology (caseguard.com).
6. Reducing AI bias with Synthetic data, by Alexander Watson, Reducing AI bias with Synthetic data | Towards Data Science.