DCPAPERS

Compliance Rating Scheme: Data Provenance for Dataset Use in Generative AI Applications

Matyas Bohacek[1,†], Ignacio Vilanova[2,†]

[1]Stanford University, Stanford, CA, USA
[2]Imperial College London, London, UK

Abstract

Generative Artificial Intelligence (GAI) has experienced exponential growth in recent years, partly facilitated by the abundance of open-source large-scale datasets. These datasets are often built using unrestricted and opaque data collection practices. While most literature focuses on the development and applications of GAI models, the ethical and legal considerations surrounding the creation of these datasets are often neglected. Specifically, the information about their origin, legitimacy, and safety often gets lost. To address this, we conceptualize the Compliance Rating Scheme (CRS) as a tool to evaluate a given dataset's compliance with a set of practical principles, enabling developers and regulators to gauge and verify the transparency, accountability, and security of these resources. We open-source a Python library built around these principles, allowing the integration of this tool into existing pipelines.

Keywords

Generative AI, Datasets, Data Provenance, Metadata

1. Introduction

We propose the Compliance Rating Scheme (CRS) tool, serving as an intuitive indicator for AI practitioners to gauge the compliance of a dataset that they are considering to use at the data acquisition stage of their project. This tool is inspired by previous ethical work on data management such as the Fair Information Principles [1] and existing legal frameworks like the GDPR [2] and the California Consumer Privacy Act [3]. To obtain a CRS score, each data point in the dataset must be examined for the following six requirements:

1. The shared dataset configuration is compatible and matched with the corresponding dataset license. This means, for example, that the allowed purposes of use do not conflict with the license of the dataset.
2. The dataset complies with the provenance metadata and its licenses. This means that the licenses of the respective data points fall within the scope and allowed purposes of the dataset, as set up in its configuration.
3. The dataset flags any data points where the compliance with the provenance metadata is inconclusive.

DCMI-2024 International Conference on Dublin Core and Metadata Applications
[†]These authors contributed equally.
maty@stanford.com (M. Bohacek); i.vilanova21@imperial.ac.uk (I. Vilanova)
https://www.matyasbohacek.com (M. Bohacek); https://nachovilanova.com (I. Vilanova)
1234-5678-9012 (M. Bohacek); 0009-0006-1442-3591 (I. Vilanova)

https://doi.org/10.23106/dcmi.952486058

4. The dataset has an opting-out mechanism, allowing authors of the data points to request their removal from the dataset if they had not previously given consent.
5. The dataset allows for legitimate access; in other words, its configuration allows for the most permissive set of purposes of use given its license.
6. The dataset adds the dataset source and the retention period into the provenance metadata of the data points.

The function then arrives at a final score supported by detailed data point-level reasoning. The CRS is represented on a letter scale from "A" (the highest score) to "G" (the lowest score). The scores are attributed based on the previously described six criteria. The presence of each of these criteria moves the CRS for a given dataset one letter grade above. For example, if a dataset does not meet any of these criteria, it receives a CRS of "G". Contrarily, if the dataset meets all criteria, it receives a CRS of "A".

2. Library

To materialize this toolkit, we create and open-source a new Python library called *DatasetSentinel*, available at [anonymized]. This library is aimed at AI practitioners and researchers who create and consume datasets. The library can be easily integrated into existing AI pipelines using PyTorch [4], TensorFlow [5], and MLX [6]. It adopts the Content Authenticity Initiative's (CAI) library [7] and Coalition for Content Provenance and Authenticity's (C2PA) data provenance standard [8]. CAI's library the official implementation of C2PA, which is the most prominent data provenance standard, currently being implemented into many social media platforms and hardware products.

The library addresses two use cases: (i) assessment of prospective data points while creating a new dataset (or creating a new version of an existing dataset) and (ii) evaluation of the CRS score for an existing dataset. The library provides one function for each use case. At the input to each function, the user provides context about the dataset they are working with in the form of a configuration file. This configuration file captures information such as the license and allowed uses of the resulting dataset and data policies such as whether to include data points generated using AI or whether to include data points marked as artistic work.

For the first use case of creating a new dataset or creating a new version of an existing dataset, the user must provide the prospective data point under consideration in addition to the configuration file. The library determines whether this data point is compliant with the dataset policies as well as the ethical principles proposed in this paper, examining the provenance and EXIF metadata of the data point. It then provides an overall assessment of the data point, supported by reasoning.

For the second use case of evaluating the CRS score of an existing dataset, the library examines all the data points in the dataset, verifying their compliance with the configuration file and CRS criteria. The function returns a CRS score with a complete list of findings that led to this score.

As such, the library is simultaneously reactive and proactive, as in addition to evaluating the CRS of existing datasets, it equally informs responsible scraping and construction of new datasets by filtering out prospective data points that are not compliant and keep only those that match all the criteria set up in the dataset configuration.

https://doi.org/10.23106/dcmi.952486058

References

[1] OECD privacy principles, 1980. URL: http://oecdprivacy.org/.

[2] General data protection regulation (GDPR) – official legal text, 2018. URL: https://gdpr-info.eu/.

[3] California consumer privacy act 2018, 2018. URL: https://oag.ca.gov/privacy/ccpa.

[4] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Neural Information Processing Systems, 2019. URL: https://api.semanticscholar.org/CorpusID:202786778.

[5] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zhang, Tensorflow: A system for large-scale machine learning, in: USENIX Symposium on Operating Systems Design and Implementation, 2016. URL: https://api.semanticscholar.org/CorpusID:6287870.

[6] A. Hannun, J. Digani, A. Katharopoulos, R. Collobert, MLX: Efficient and flexible machine learning on apple silicon, 2023. URL: https://github.com/ml-explore.

[7] L. Rosenthol, A. Parsons, E. Scouten, J. Aythora, B. MacCormack, P. England, M. Levallee, J. Dotan, S. Hanna, H. Farid, et al., The content authenticity initiative: Setting the standard for digital content attribution, Adobe Whitepaper (2020).

[8] L. Rosenthol, C2pa: the world's first industry standard for content provenance, in: Applications of Digital Image Processing XLV, volume 12226, SPIE, 2022, p. 122260P.

https://doi.org/10.23106/dcmi.952486058