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

MATYAS BOHACEK*, Stanford University, USA
IGNACIO VILANOVA ECHAVARRI*, Imperial College London, UK

The field of Generative Artificial Intelligence (GAI) has witnessed an exponential growth in recent years, partly facilitated by the abundance of open-source large-scale datasets. These datasets are often built from 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. We conceptualize the Compliance Rating Scheme (CRS) as a tool to evaluate a given dataset's compliance with a set of ethical principles, enabling developers and regulators to gauge and verify the transparency, accountability, and security of these resources. Finally, we open-source a Python library built around these principles allowing the integration of this tool into existing pipelines.

Additional Key Words and Phrases: Generative AI, Dataset, Data Provenance

1 INTRODUCTION

As Generative Artificial Intelligence (GAI) applications are becoming increasingly intuitive, easier to use and their results more realistic, their adoption are becoming widespread [11], as they are continuously pushing the boundaries of what was previously thought as impossible. This exponential growth in recent years is partly facilitated by the abundance of open-source large-scale datasets. These are often built from unrestricted and opaque data collection practices [7, 9]. Datasets play an essential role in the AI ecosystem [10] as they are the primary – if not the only – source of training for most AI systems. Yet, while most literature focuses on the development and applications of GAI models, few focuses on the ethical and legal considerations surrounding the creation of these datasets [7, 9] .

As such, the democratization of GAI has equally caused a surge in malicious activity [7, 14], notably through impersonations (to spread misinformation or commit fraud), copyright infringement, and deepfake pornographic footage [7] (with the aim to humiliate or personal sexual gratification). Yet, the complexity and opacity of the structure of advanced Artificial Intelligence (AI) models, combined with the lack of traceability and accountability of these technologies, pose a series of technological and legal challenges to hold individuals liable for their misuses and potentially prosecute them.

This research aims to call for a larger discussion confronting the unsustainable dataset practices in the AI community by providing a framework to responsibility, liability, and legal enforcement of dataset compliance – some of the key challenges of human-computer collaboration. Building on prior work [1–5], we propose a new framework consisting of four practical principles for the creation and use of datasets for AI training and applications: (i) Transparency and Fair Use; (ii) Accountability and Liability; (iii) Prevention of Harm; and (iv) Effective and Efficient Enforcement. Based on these, we conceptualize the Compliance Rating Scheme (CRS) as a tool to evaluate a given dataset's compliance with these principles, enabling developers and regulators to gauge and verify the transparency, accountability, and security of these resources. We detail these principles and the CRS in an upcoming publication.

*Both authors contributed equally to this research.

2 LIBRARY

We open-source a Python library called *DatasetSentinel* built around the four ethical principles described above, allowing for the integration of this tool into existing pipelines. The library design is compatible with PyTorch [12], TensorFlow [6], and MLX [8], as well as HuggingFace [15], Kaggle, and custom databases. Another essential feature of the library design is transparency, attained by surfacing the low-level reasoning all the way up to the user. Lastly, the library design also underscores intuitiveness. Thus, the interface follows the same logic as this paper and requests very little additional information to be provided by the user other than some context about the datasets and prospective data points in question. By user, we refer to the AI practitioner using the library in their AI workflow.

2.1 Structure and Features

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 primary user-facing function for each case that handles the assessment. In either case, the library requests some context about the given dataset as a configuration file. This dataset configuration 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 use case of creating a new dataset or creating a new version of an existing dataset, the library provides a function that, given the dataset configuration of the dataset and the prospective data point, determines whether this data point is compliant with the dataset policies as well as the ethical principles proposed in this paper. This function examines the provenance and EXIF metadata of the data point. In the end, the function provides an overall assessment supported by reasoning. For instance, if we decided to collect a new dataset by scraping the internet, we could use this function in the scraping loop. It would filter out prospective data points that are not compliant and keep only those that match all the criteria set up in the dataset configuration and by the ethical principles.

2.2 Compliance Rating Scheme

The CRS tool serves as an intuitive indicator allowing AI practitioners to gauge the compliance of a dataset that they are considering using at the data acquisition stage of their project. To do so, the function examines each data point in the dataset and checks whether the previously described six requirements are met. In the end, the function arrives at a final score supported by detailed data point-level reasoning. The CRS is represented on a letter scale from “A” (the highest, most compliant score) to “G” (the lowest, least compliant score). The scores are attributed based on 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”. The criteria are as follows:

(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.
(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.

For instance, if we worked on a new AI project, we could use this function to examine a few datasets that we are considering. The aim is to help users make an informed decision about the dataset they are using, and ultimately only use those with a CRS score of 6 out of 6. In the future, dataset-sharing platforms may adopt this feature on their end, which would remove the heavy lifting (of running this analysis) from individual users.

Lastly, we propose the following visuals to materialize the CRS score. Figure 1 depicts the CRS scale graded chromatically and alphanumerically with values for an "A" (a) and "C" (b) score respectively. This design is similar to existing rating-schemes, as it has proved to communicate a specific value effectively and clearly to individuals. We emphasise the importance of proposing a design that is user-centred, to facilitate its understanding and adoption.

Fig. 1. Proposed design interface for A and C score on the CRS scale

3 DISCUSSION

By proposing a set of four ethical and practical principles to consider for dataset compliance in the context of AI, we aim to provide a framework that raises the much-needed discussion on the legality and ethics of AI applications. However, similar to most principles, these can be interpreted as highly conceptual and disconnected from current practices, often making them either irrelevant or challenging to implement. Precisely for this reason, we attempted to move away from a purely descriptive contribution to the literature, and provide a tangible and prescriptive approach through our *DatasetSentinel* library and CRS tool.

We highlight the specific points of the AI workflow at which we target our contribution, aiming to reduce the misuse of personal data for GAI training models and applications by introducing traceability and accountability of the datasets used for harmful purposes. To this end, the first line of defense is with the *DatasetSentinel* library, which can be used by practitioners to filter the data collected through its provenance metadata to ensure that it is compliant with the purpose of the dataset. Thus, implementing this tool benefits society as a whole (on the long term), as well as the individual user (on the short term), as it removes the heavy lifting of manually conducting this type of analysis.

The second line of defense is the CRS score, which calculates and informs the practitioners about the dataset's compliance with the ethical principles embedded in its structure. As such, defendants accused of harm through GAI applications can no longer plead ignorance about the nature or compliance of any given dataset. This framework also aims to bridge the gap between digital technological innovation and accountability by providing a framework to responsibility, liability, and legal enforcement of data malpractices. To this end, in the eventuality of a legal demand, the CRS score enables developers and regulators to gauge and verify the transparency, accountability, and security of any given dataset, with the ultimate objective of providing traceability and accountability.

We are witnessing a growing interest among software and hardware companies in tracing the provenance of media in an attempt to fight misinformation and other malicious content. This trends is manifesting itself, for example, by an

uptick of organizations joining coalitions such as the Coalition for Content Provenance and Authenticity (C2PA) [13]. It seems that there is a growing trend towards data traceability and immutability within the digital sphere. We therefore reiterate our belief in this project, and its potential positive impact within the field of AI.

4 LIMITATIONS

As our prototype is in its infancy, we acknowledge its limitations and that there is still much research to be conducted until this framework can become a standard for ethical GAI use. For instance, as data provenance technologies are just rolling out, the majority of digital media available online still lacks provenance metadata. Nonetheless, to address this, many technological companies – both in software and hardware – are starting to deploy or announce the integration of data provenance technologies into their products. Therefore, we expect that, within a few years, the vast majority of new digital media distributed on the internet will have provenance metadata. Another limitation is that the library is dependent on the existing data provenance protocols. To that end, our library can only analyze data types that are supported by these protocols and other dependencies. This should not pose a problem for most current use cases, as the protocols support the most common data types for image, video, audio, and 3D objects. Still, moving forward, this dependency could introduce a delay in introducing support for new data types.

Closely connected is also the question of whether it is possible to effectively extract data from existing AI datasets that have already been used to train models given that the information from the extracted data points will have already been encoded in the weights of some AI models. There are also many non-technical research topics connected to this line of research. For example, it will be critical to understand how to allow ordinary technology users –sharing images, video, and audio on the internet – to encode their preferences about the AI training (or not) with their data in its provenance metadata.

5 CONCLUSION

We call for a larger discussion confronting the unsustainable dataset practices in the AI community. While we recognize that the dataset sharing platforms have substantial power to influence the practical rules and guidelines, we argue that a value shift is also needed. Specifically, a broader awareness and appreciation of ethical and legal considerations surrounding datasets must be established for the rules and guidelines of dataset sharing platforms to have a meaningful impact. Our framework and tangible outputs can serve as a springboard for piloting and implementing these values into existing workflows. With the main focus of this work on interactions between humans and generative AI agents, we hope to inspire future research in this field, and encourage further collaboration among the fields of AI, ethics, law, CHI, and others.

REFERENCES

[1] [n. d.]. APEC Privacy Framework. https://www.apec.org/publications/2005/12/apec-privacy-framework

[2] 1980. OECD Privacy Principles. http://oecdprivacy.org/

[3] 1998. Privacy Online: A Report to Congress 7. https://www.ftc.gov/reports/privacy-online-report-congress

[4] 2018. *California Consumer Privacy Act 2018.* https://oag.ca.gov/privacy/ccpa

[5] 2018. General Data Protection Regulation (GDPR) – Official Legal Text. https://gdpr-info.eu/

[6] Martín Abadi, Paul Barham, Jianmin Chen, Z. Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zhang. 2016. TensorFlow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation.* https://api.semanticscholar.org/CorpusID:6287870

[7] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[8] Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. *MLX: Efficient and flexible machine learning on Apple silicon.* https://github.com/ml-explore

[9] Ben Hutchinson, Andrew Smart, A. Hanna, Emily L. Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2020. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2020). https://api.semanticscholar.org/CorpusID:225067460

[10] Bernard Koch, Emily L. Denton, A. Hanna, and Jacob Gates Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *ArXiv* abs/2112.01716 (2021). https://api.semanticscholar.org/CorpusID:244894836

[11] Jonas Oppenlaender, Aku Visuri, Ville Paananen, Rhema Linder, and Johanna Silvennoinen. 2023. Text-to-Image Generation: Perceptions and Realities. *arXiv preprint arXiv:2303.13530* (2023).

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Neural Information Processing Systems.* https://api.semanticscholar.org/CorpusID:202786778

[13] Leonard Rosenthol. 2022. C2PA: the world's first industry standard for content provenance. In *Applications of Digital Image Processing XLV*, Vol. 12226. SPIE, 122260P.

[14] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22522–22531.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019). https://api.semanticscholar.org/CorpusID:208117506

Received 22 February 2024