Data Transparency Evaluation Using Quantitative Measures

Vritika Goyal
Textile Engineering, IIT Delhi, IMT Atlantique
Delhi, India/ Rennes, France

Vasumathi Muthaiyan
CSNE, IMT Atlantique
Rennes, France

Volker Tachin
Architecture and Engineering for the Internet of Things, IMT Atlantique
Rennes, France

***Abstract*—The focus of our research is to look for ways to improve data transparency in general applications. We face an issue of data insufficiency in the fields of artificial intelligence and machine learning and that affects the end product which doesn't have the same result with all users. It works well with certain users depending on the input data the product is modeled after. The source of the data also plays a vital role and it is crucial that the users are made aware of that information. Data bias and discrimination also have serious real-world effects and are especially harmful towards ethnic minorities and/or people of colour. We aim to increase the level of data transparency and reduce bias and discrimination in machine learning and facial recognition software. We came across numerous methods to achieve this goal over the course of our research and we have made recommendations to improve upon these methods.**

Keywords—Data interpretation, Fairness, Discrimination, Prejudice, Bias

I. INTRODUCTION

In recent years, there has been a significant increase in the amount of data being collected and analyzed in various fields like finance, healthcare, social media, and government. Due to this growth, concerns have been raised about the interpretation and fairness of data in decision-making processes. Issues related to discrimination and prejudice have been highlighted as major challenges that need to be addressed. One way to address these concerns is to increase data transparency and the methods used to analyze it. To do this, an understanding of what data bias is and how it affects data transparency is required. We should also know how data transparency, bias, and discrimination are related.

Data transparency refers to the ability to access, understand, and utilize data in a meaningful and effective manner. It is a critical aspect of data-driven decision-making and is particularly important in applications such as healthcare, finance, and government, where decisions can have significant consequences for individuals and society as a whole.

We can evaluate data transparency through the use of transparency indices, which are quantitative measures that assess the availability, accessibility, and quality of data [2]. These indices typically use a set of predefined criteria to evaluate data transparency, such as the availability of data in machine-readable formats, the presence of metadata that describes the data, and the availability of documentation that explains the data sources and methodology used in data collection. Transparency indices can be used to compare the degree of data transparency across different applications, sectors, or countries, providing a useful benchmark for assessing progress over time [3].

Another approach to evaluate data transparency is through the use of data quality metrics, which are quantitative measures that assess the accuracy, completeness, and consistency of data. These can be applied to specific datasets or to the overall data environment, providing valuable insights into areas where data quality may be lacking. For example, data quality metrics may be used to identify areas where data is missing or incomplete, where data sources may be unreliable or biased, or where data is not updated in a timely manner.

Our focus is on data bias and discrimination in Artificial Intelligence models. Facial recognition software are biased towards people of colour or a minority group. The different biases occurring in AI are systemic, automation, selection and implicit biases, to name a few. In medicine, textbooks often describe symptoms observed in Caucasian people rather than including the symptoms observed in people of other races too. This is a good example of systemic bias. Since there's a lot of discrepancy in data that have major applications in our daily lives, this poses a serious risk.

In addition to transparency indices and data quality metrics, other quantitative measures may be used to evaluate data transparency in specific applications. For example, in the field of healthcare, measures such as the percentage of patients who are able to access their medical records online, the degree of interoperability between different electronic health record systems, and the extent to which patient data is shared across different providers and organizations can be used to evaluate

the degree of data transparency in the healthcare system. Similarly, in the financial sector, measures such as the availability of data on corporate governance and executive compensation, the transparency of financial reporting, and the quality of credit ratings can be used to assess the degree of data transparency in financial markets.

One of the challenges of evaluating data transparency using quantitative measures is that there is no one-size-fits-all approach. The specific measures used will depend on the particular application and the nature of the data being evaluated. Furthermore, the interpretation of these measures can be subjective and may depend on the evaluator's perspective. For example, what may be considered a high degree of data transparency from the perspective of a government regulator may be viewed as inadequate by a civil society organization advocating for greater transparency and accountability [9].

Despite these challenges, quantitative measures can provide valuable insights into the degree of data transparency in specific applications. In particular, this research project will focus on the use of data interpretation, fairness, discrimination, and prejudice as key indicators of data transparency. By analyzing these indicators in different contexts, we hope to develop a set of quantitative measures that can be used to evaluate the degree of data transparency in various applications. Ultimately, the results of this research project will contribute to the development of best practices for promoting transparency and fairness in data analysis.

We discuss various methodologies of determining and reducing data transparency and data bias in section 2 and mention the encountered limitations in the existing methods for enhancing transparency in section 3. Followed by those are our discussions related to the research in section 4 and section 5 summarizes our conclusions.

II. METHODOLOGIES

A. AUDIO

AUDIO stands for Auditing Framework of Outlier-Mining-as-a-service [2]. It is a system that verifies if the data mined using an outsourced service is accurate. A small number of artificial tuples was introduced into the data and that creates artificial outliers and non-outliers that weren't present in the initial dataset. If these outliers are present in the server's output, then we can infer that the result is valid. There are two types of servers that return an invalid answer; semi-honest and malicious server. The semi-honest server runs the mining algorithm properly but might return incorrect outliers due to bugs or human error. The malicious server returns the wrong answer intentionally and it tries to escape verification if it knows those details beforehand. A completeness verification is done to check if the server has returned all the original outliers and a correctness verification is done to return the correctness probability.

B. Automated Annotations

Data annotation is another method which seeks to provide a framework by which organisations can annotate their datasets in an automated fashion [6], as a complement to datasheets or model cards.

This method is however limited, especially as the size of the datasets increases, due to manual entry. Because this method is quantitative, it is possible to generate these annotations by automatic computation. The annotations need to be organised in a hierarchical key-value pair manner, while also following a predefined standard, to prevent metadata silos due to variations in organisational annotations. It would be crucial to maintain some annotations with wide applicability, as this would help with standardization and speed up the annotation process. The main components of this reference architecture are the following; Data Lake Ingestion, Annotation Computation, Annotation Storage, Checklists and Policy Compliance, Querying Annotations, Annotation Analytics, Visualisations and Quality Monitoring.

C. Sample Discrimination-based Selection (SDS)

Since there are a lot of Deep Learning (DL) models with the same function, those models are compared and the more suitable ones are selected [5]. The testers have to select an efficient subset of samples to rank the models as precisely as possible. A Sample Discrimination based Selection (SDS) is used to select samples that could discriminate between multiple models, like predicting the behaviours of the samples. SDS measures the sample discrimination and selects samples with higher discrimination. An empirical study with 3 widely-used image datasets and 80 real-world DL models was conducted to evaluate the SDS. The results show that SDS is an effective and efficient sample selection method to rank multiple DL models, compared to other baseline methods.

D. Co-12 Properties

The paper [11] synthesises and provides an additional number of evaluation methods from various papers. These methods are grouped into 12 sections based on their Co-12 properties namely; Correctness, Output-completeness, Consistency, Continuity, Constrastivity, Covariate Complexity, Compactness, Composition, Confidence, Context, Coherence, Controllability. The suggestion is that several evaluation methods be selected together, ensuring as many of the Co-12 properties are covered, so as to attain a multi-dimensional view of the degree of explainability. These are grouped into three property groups; Content, Presentation and User.

For correctness, the general ideology is to apply some perturbation of some form to the inputs and check that the model explanation changes. These perturbations could be in the form of modifications to the model parameters. Another method could be applying the explanation method to a white-box model and checking how closely the explanations and the model's reasoning are similar.

Completeness and consistency methods attempt to quantify the degree to which the explanations explain the model and the extent to which the said explanation is deterministic

respectively. Some of these methods include the Preservation Check and Implementation Invariance [11] respectively.

Continuity, contrastivity and covariate complexity also help evaluate the content of these explanations.

User-facing properties are ensured by evaluating the context, coherence and controllability properties of the explanations. Alignment with Domain Knowledge is a method used to evaluate whether explanations generated align with general beliefs/consensus, contained in a "ground-truth" dataset. In the case of image data, a heatmap can be used to check "location coherence" and from this, the Jaccard index, e.g. [12], outside-inside relevance ratio, or even pointing game accuracy can be used to evaluate the explanations, e.g. [13].me accuracy can be used to evaluate the explanations, e.g. [13].

E. Document Engineering for Disclosure Interfaces

Document Engineering [9] is a promising methodology for assisting in the creation of interface disclosures. DocEng is a proven methodology for defining, designing, and deploying data interfaces for facilitating document-centric applications. It comprises defining different information and data sources with new document models and reusing common or standard patterns to make documents more robust.

Following are the six primary stages involved in executing this technique:

Step 1 - Analyzing the contexts of use: The main motto of this step is to assess the contexts in which the process's outcome will be implemented. This includes determining prospective stakeholders while assessing what they might gain from the disclosure, in addition to contemplating the regulatory requirements that the disclosure must satisfy.

Step 2 - Analyzing business processes and patterns: The second stage involves examining current business processes with the objective of better understanding the types of data presented and the way it is organized, arranged and presented.

Step 3 - Document Analysis: The process of analyzing the content and structure of the obtained documents. This included examining the structure of specific sector disclosures to determine how related data is classified and whether illustrations or alternate formats might more effectively meet the recipient's needs. In addition, the disclosures are frequently segregated, necessitating technical knowledge and effort to combine and interpret them in a meaningful manner.

Step 4 - Component Assembly: The fourth step comprises modelling features (groups of related data) to investigate how disclosure material can be put back together and presented.

Step 5 - Document Assembly: Several features, like HTML-based disclosure, can be used to construct pages for similar data groups and sequential navigation links. This is done to produce a prototype for our disclosure that is integrated and consistent.

Step 6 - Implementation and Beyond: This step involves executing and enacting the new interface, developing and implementing HTML/CSS/Javascript for browser-based interactive disclosure, and providing localized archives and access to all files.

Importantly, DocEng does not end here; disclosure interfaces, like programs, require continuous improvement and modification to evolve.

F. Prejudice Remover Regulariser

This is a mathematics intensive method which focuses on the This is a mathematics-intensive method [4] which focuses on the quantification of the extent to which the model is biased. We shall briefly discuss it here without going into much detail. It is a method which directly tries to reduce the Prejudice Index that is denoted by Rpr. Prejudice Index is a parameter to quantify the degree of indirect prejudice. Prejudice Index (PI) can be defined as

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y,s] \ln \frac{\hat{\Pr}[y,s]}{\hat{\Pr}[y]\hat{\Pr}[s]} \qquad (1)$$

A few substitutions and simplifications lead us to the final expression for prejudice remover regularizer, RPR(D, Θ) which is

$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y \mid \mathbf{x}_i, s_i; \Theta] \ln \frac{\hat{\Pr}[y \mid s_i]}{\hat{\Pr}[y]} \qquad (2)$$

This regularizer is based on logistic regression models and is focused on classification. The random variables Y, X, and S represent a class, non-sensitive features, and sensitive features, respectively. A training data set includes D = {(y, x, s)} corresponds to the occurrences of these random variables. M[Y |X, S; Θ] models the conditional probability of a class given non-sensitive and sensitive features, where Θ is the set of model parameters. Pr denotes the conditional distribution of Y given the values for X and S. Apart from the high-end mathematics this method includes, this regularizer signifies a larger value when the class is determined based on the sensitive features thereby making them less influential in the final determination.

III. LIMITATIONS

There are a few knowledge gaps, identified by some researchers in the articles we found. In one paper [4], a knowledge gap in the definition of fairness in data analysis is identified. Indeed, the term data fairness is still being debated in the industry. The researchers mention that there are limited methods of analytic measures available at their disposal. The method can be applied to regression, but more methods need to be developed for fairness-aware clustering as well as for ranking. Due to the lack of convexity of an objective function, the method applied is occasionally trapped by local minima. To overcome this, kurtosis was used for independent component analysis. If a sensitive feature is a multivariate variable whose domain is large or is a real variable, this current prejudice remover cannot be applied directly. A sample discrimination-

based selection [5] was used to compare and predict the behaviors of various deep-learning models. The knowledge gap in this paper was to find a way to select an efficient set of models, and test and rank them in the most precise way possible. Subjective user feedback cannot be reused for all new models as the various explanations would require new reviews. This method [10] also drives a high annotation cost for multi-level human explanation masks. Additionally, if the information that the organizations disclose [9] is not useful to the people who are receiving it, then the disclosures are nothing more than highly technical "data dumps." This has a limited beneficial effect, diminishes the effectiveness of a disclosure, works to disempower, and ultimately inhibits the broader transparency goals that are being pursued. Because of this, genuine openness is hindered, as it becomes progressively difficult for any recipient who is not an expert to access and interpret the information that has been revealed. This excludes the vast majority of recipients. We suggest a broad reconceptualization of disclosures toward user interfaces as a tool for achieving more effective transparency in order to design disclosures that are more suitable, useful, and relevant for the people who are supposed to receive them. In this way, we can design disclosures that are more effective in pursuing transparency.

As the Co-12 properties are not a direct judgement of truth but rather a summary judgment of accessible explanatory virtues, evaluating a single evaluation method would be ineffective. There also seems to be much more research and methods that evaluate coherence, completeness, compactness or correctness. This makes the first limitation even more apparent. It also becomes clear that it would be unreasonable to expect current explainable AI to score highly using these metrics. In reality, tradeoffs would need to be made in deciding which of the properties needs to be prioritised for each use case.

IV. Discussion

A part of our research is focused on dealing with the causes of unfairness in data. These causes are further linked to classification, social responsibility and prejudice. We came across several regularization techniques which are applicable to a variety of prediction models. After applying this approach further to other logistic regressions, we were a step closer to our problem statement. The data is disclosed so that the person receiving it can interact with it, understand it, and act upon it efficiently. The data characteristics in the research articles encompass primary and secondary data from various sources, such as surveys, interviews, archival data, and publicly available datasets. The applications of the articles span across different domains and the analysis will involve identifying common approaches and methodologies used for evaluating data bias and data transparency. The strengths and limitations of the research methods employed were critically evaluated, and the implications of the findings for the respective application contexts were discussed. Additionally, the analysis will consider the generalizability of the findings, the reliability and validity of the data, and the robustness of the research methods.

The AUDIO [2] framework provides efficient ways of checking the completeness and correctness of results in the case of data mining activities. Its relevance to our work can be understood by the fact that it would give us some insights into a quantitative method already tested and being applied in the industry. One paper [3] deals with the issue of prejudice in face detection models to detect fraud in the usage of benefit cards in the public transport system in Salvador. The study shows a bias towards black women and men and this was due to the use of a pre-trained model. The percentage of errors in the system has a significant consequence in a real-world application. The solution to this problem is to retrain the facial detectors and update their parameters. A Fairness-Aware classifier [4] looks into a method of reducing indirect prejudice, which applies regularizers that restrict the learner’s behaviours in any prediction algorithm. The tradeoff between classification accuracy and fairness in the result is considered.

In fact, excessive information can hinder understanding of what is happening by masking useful information behind or amid other information (the identified 'transparency paradox'). More precisely, we demonstrate the prospect for disclosure interfaces, in which disclosures designed as user interfaces are modified according to the requirements, interests, and expectations of the recipient. We demonstrate one feasible method for their development and evaluation, which is document engineering. The disclosure interface is essential for, potentially contributing to the empowerment and knowledge of stakeholders such as users, regulators, and organisations.

In brief, we could say that the three main causes of unfairness are prejudice, underestimation, and a negative legacy. Sometimes, even if the sensitive features are removed from the calculations, they indirectly can affect the model thereby leading it to unfair determinations. Such a phenomenon is called Red Lining Effect or indirect Discrimination. Other such parameters used to measure the extent of bias are Cross Variation score which is defined by subtracting the conditional probability of the positive class given a sensitive value from that given a non-sensitive value.

As a countermeasure for the limitations faced by the Co-12 properties, it is suggested that the selection of properties to be prioritised, be performed in an optimised fashion so as to maximise their effectiveness. In addition, for any of the evaluation methods to work effectively, there need to be some guidelines established as part of existing data laws and regulations.

V. Conclusion

Data transparency cannot be evaluated independently, It’s dependent on data bias, insufficiency and discrimination. If data is biased, then there are more chances that the organizations will not be forthcoming about the source of the data. Or they might release the information in a way that’s hard for normal people to understand it. To avoid this, there should be a rule that asks the enterprises to release the information that is formatted according to the audience. For example, there can be a specific format for laypersons, a

format that is more technical for researchers or scientists and a particular format for government organizations.

The current nature of the data landscape requires that these methods discussed here within, be further studied as urgently as possible. As a result, studies like ours are important and should be able to help enforcers and organisations alike, to make informed decisions that help mankind.

REFERENCES

[1] Elisa Bertino, Ahish Kundu, and Zehra Sura. 2019. “Data Transparency with Blockchain and AI Ethics”. J. Data and Information Quality 11, 4, Article 16, 8 pages (2019).

[2] Liu, R., Wang, H., Monreale, A., Pedreschi, D., Giannotti, F., Guo, W. “AUDIO: An Integrity Auditing Framework of Outlier-Mining-as-a-Service Systems”. Springer, Berlin, Heidelberg, Flach, P.A., De Bie, T., Cristianini, N. (eds) Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol 7524, pp. 1 -18 (2012).

[3] Ferreira, M.V., Almeida, A., Canario, J.P., Souza, M., Nogueira, T., Rios, R. “Ethics of AI: Do the Face Detection Models Act with Prejudice?”. Springer, Cham. Britto, A., Valdivia Delgado, K. (eds) Intelligent Systems. Lecture Notes in Computer Science, vol 13074, pp. 89–103 (2021).

[4] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J. “Fairness-Aware Classifier with Prejudice Remover Regularizer”. Springer, Berlin, Heidelberg, Flach, P.A., De Bie, T., Cristianini, N. (eds) Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol 7524, pp. 35-50 (2012).

[5] Linghan Meng, Yanhui Li1, Lin Chen, Zhi Wang, Di Wu, Yuming Zhou, Baowen Xu, “Measuring Discrimination to Boost Comparative Testing for Multiple Deep Learning Models” IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Madrid, ES, pp. 385-396 (2021).

[6] N. Imtiaz, J. Middleton, J. Chakraborty, N. Robson, G. Bai and E. Murphy-Hill, “Investigating the Effects of Gender Bias on GitHub” IEEE/ACM 41st International Conference on Software Engineering (ICSE), Montreal, QC, Canada, 2019, pp. 700-711 (2019)

[7] Saravanan Thirumuruganathan, Mayuresh Kunjir, Mourad Ouzzani, and Sanjay Chawla. “Automated Annotations for AI Data and Model Transparency”. J. Data and Information Quality 14, 1, Article 2, 9 pages (2022).

[8] Rogelio P.C. Amorim and Credine Silva de Menezes. “Evaluation Methodology of Municipal Transparency Websites”. Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era - Volume 1. Brazilian Computer Society, Porto Alegre, BRA, 17–24 (2016).

[9] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. “Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability”. ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, 679–690 (2022).

[10] Sina Mohseni, Jeremy E Block, and Eric Ragan. “Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark”. 26th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, 22–31 (2021).

[11] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. “Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI”. ACM Comput. Surv. (2023).

[12] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 3319-3327 (2017).

[13] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards Explanation of DNN-based Prediction with Guided Feature Inversion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 1358–1367 (2018).