Challenges and Future Directions for Accountable Machine Learning*

Agne Zainyte and Wei Pang**

Heriot-Watt University, Edinburgh, Scotland, United Kingdom, EH14 4AS
{ az66, w.pang }@hw.ac.uk

Abstract. In recent years, machine learning (ML) algorithms have been applied in many areas such as healthcare, finance and autonomous vehicles. At the same time, there is an increasing need for making ML systems accountable, which would help deal with situations when these systems made wrong decisions or predictions. Currently there exist three major frameworks for Accountable ML: Model Card Toolkit, Datasheets, and FactSheets. However, the greatest limitation of these frameworks is that they are mostly focusing on qualitative information about the machine learning models. In this research, we discuss in detail these three frameworks and future development directions of Accountable ML frameworks; we recommend the implementation of causality, decision provenance, and computational tests for achieving better ML accountability.

Keywords: Accountable, Machine, Learning, Artificial, Intelligence, Transparent, Frameworks.

1 Introduction

Machine Learning (ML) is a type of Artificial Intelligence (AI) algorithms that are capable of learning from data and improving themselves accordingly. Due to fast increasing computational power, in recent years researchers have been able to develop sophisticated ML algorithms that are capable of solving complex problems. These ML algorithms are now being applied to healthcare, finances, banking, marketing, and even autopilots.

The utilization of ML in automated decision making has brought the attention of several regulatory bodies; for example, EU's General Data Protection Regulation (GDPR) commission, which has now established a legal right to explanations of decisions made by automated means, such as ML algorithms [1]. Moreover, automated decision-making systems (with ML systems being one type of these systems) have been

**WP is supported by the RAInS project funded by EPSRC (EP/R033846/1).

increasingly used in situations that if an error occurred it would have detrimental effects on human lives. This further accelerates the need for accountability in ML models.

This paper aims at reviewing currently existing Accountable ML frameworks, pointing out challenges associated with facilitating accountability in ML, and suggesting future directions in developing accountable ML frameworks.

The rest of the paper is organised as follows: In Section 2, we discuss the accountability in machine learning. This followed by the introduction of the three main accountable machine learning frameworks in Section 3. Then in Section 4, we point out the challenges of accountable machine learning. In Section 5, we present the future directions of accountable machine learning. Finally, Section 6 concludes the paper.

2 Accountability in Machine Learning

There was an infamous incident in 1991 when Dean Pomerleau was pioneering in autonomous vehicle research by trying to teach a computer how to drive. Pomerleau was training the computer for a few minutes by driving himself and then later would allow it to drive the vehicle itself. Unfortunately, the computer made an error that would have caused a car crash if Pomerleau would have not taken control back in time [2]. The error occurred due to the training data being inefficient – it mostly contained grassy roadsides and when the computer captured the bridge it was confused as it was never trained on it before.

Although, the research came far in last 30 years, given the fact that AI is being extensively used in decision making, the concerns still arise what if AI makes mistakes. The extensive discussion of who would be held accountable for AI's mistakes was presented by Hevelke *et al.,* [3] who analysed the situation where an autonomous vehicle got in a car crash. The authors stated that the most intuitive solution would be to hold the manufacturers responsible for any crash caused by the vehicle (if it was system error that caused it), however, the question of who is accountable if the vehicle is designed in such a way that it continuously develops and improves itself is highly difficult to answer [3]. Hevelke *et al.* did not reach a solid conclusion of where accountability lies in self-improving algorithms. Furthermore, Mittelstadt *et al*. [4] continued the debate, highlighting the challenge that machine learning algorithms are capable of learning autonomously and are rather opaque in their reasoning. However, designers could provide information about how algorithms have been taught, which could provide more insights into the factors that could have affected the inference that the computer has made [4].

3 Current Techniques for Accountable Machine Learning

3.1 Model Card Toolkit

The Model Card Toolkit (MCT) [5] is a library developed by Google that generates Model Cards [6], which are documents that give insights into models' internal works. Model Cards were proposed to prevent mistakes caused by systematic failure that

became apparent only after the models have been put in use; Model Cards would define the model's performance before the model has been put to use. Since MCT has a collection of tools that assist in compiling and filling in Model Cards, it was meant to be a way to standardise reporting on machine learning model's ethics, provenance, usage and evaluation [6].

The greatest advantage of using MCT is that it has semi-automated process when creating Model Cards, and this ease of use could encourage developers to implement accountability into their products. However, there are several limitations for MCT, one of them being that it is only applicable to TensorFlow library [7]. Secondly, Model Cards are generated only after the model has been developed, and more specifically, it is lacking stage-to-stage analysis before its model is compiled.

3.2 Datasheets

In 2018, the so-called "Datasheets for Datasets" was proposed by Microsoft [8]. The authors raised the issue that there was no standardised way to document how and why datasets were created, what information it contains, in which scenarios it is applicable and whether it might raise any ethical or legal concerns. This type of documentations would highly benefit datasets that were not sharing their details to public, such as those containing sensitive information. A proposed solution was a Datasheet that contains a set of proposed questions that had to be filled in by dataset creators. These questions are divided into seven categories: motivation, data pre-processing, data distribution, data maintenance and legal issues and ethical considerations [8]. Datasheets primarily aim at qualitative information about datasets; however, it is lacking any visualisation, which could give greater insights about data and would help with interpretability.

The greatest advantage of Datasheets is that it acknowledges the fact that datasets could be held accountable; as Hutchinson *et al.* [9] argued, datasets directly influence how self-improving algorithms will reassemble themselves and therefore datasets play a major part in Algorithmic Decision Making (ADM) [9]. Notably, there are no follow-up research along this line as far as we are aware.

3.3 FactSheets

FactSheets is a framework that was developed by IBM research team [10]. It was probably inspired by and built upon both Model Cards and Datasheets. The idea is that each FactSheet contains sections of information about relevant attributes of a machine learning model, such as intended use, performance, safety and security. Additional detailed information about how the model was created, trained and deployed could be supplied, alongside with scenarios of use cases and legal and ethical considerations. It is clear that FactSheets are the more detailed version of Model Cards; however, it lacks the information that was captured automatically by MCT during pipeline development. The aim of creating FactSheets was to help prevent overgeneralisation of models, which would in turn help reduce the errors that are the result of unintended use of these models [10].

Ever since their release in 2019, The development of FactSheets has been active and research papers are openly available on IBM's website [11, 12, 13]. However, we

feel that information required to be filled in FactSheets is quite excessive and some of it may not be necessary the key to effective traceability and hence accountability.

4 Challenges of Accountable Machine Learning

We believe that the greatest challenge in accountable ML development is the identification of what accountability in ML comprises of. According to Mohseni *et al.* [14], accountability is the desired outcome of transparency in algorithms; the. authors argue that accountability would be inevitable if transparency and explainability were implemented correctly [14]. However, a counterargument is presented by De Laat [15]; De Laat states that more transparency does not equate to more accountability, and in fact models should be able to implement accountability even when transparency is not excessive [15]. This type of approaches would also be applicable to models that are opaque due to not being open source about their development and to those models that are opaque due to the concerns of potential loss of competitive edge [15].

Interestingly, Ananny *et al.* [16] argues that ML algorithms are more than just code, and they are also artifacts that implement practices and norms captured in data and employed by their developers – therefore, in order to facilitate accountability in ML models, it requires more than transparency of components comprising the model, but also information about how the algorithm works as a system. This highlights the second challenge in accountable ML development: the problem of what information needs to be captured in order to understand systems as a whole.

According to Cambridge Dictionary definition of accountability is “the fact of being responsible for what you do and being able to give a satisfactory reason for it, or the degree to which this happens” [17]. Similar definition but adapted to ML systems is given by Naja et al. [18] as authors state that “by accountability, we mean the ability to inspect, review or otherwise interrogate an AI system”. Additionally, we believe that accountability in ML should include sufficient level of transparency (although not necessarily full transparency), explainability, robustness, fairness and bias. In the section below we further discuss what additional information should be captured to the facilitate implementation of necessary components of accountable ML.

5 Recommendations for Accountable Machine Learning Frameworks

Current accountable machine learning frameworks focus on descriptive approaches. It is expected that detailed factual information would allow greater transparency, which in turn would help facilitate accountability. However, below are recommended future directions when developing a framework for Accountability in Machine Learning models.

5.1 Causality

It is crucial to understand what influence each stage of the model development has. By understanding the effects that each component has, the design of ML models can be improved by making sure that the effects and causes are clearly outlined and understood [19]. This type of causality would compensate for opaqueness found in algorithms and would allow traceability and improved transparency without having to reveal 'sensitive' information about the model's development.

5.2 Decision Provenance

Decision provenance involves provenance methods to provide information about decision pipelines, and it shows how data flow throughout the model and what connections were being made that lead to the final decision. The information gained from decision provenance could allow developers to be proactive during the model development process and mitigate the risks associated with ADM [20]. It would symbiotically work with causality – understating how algorithms change and what component of the system powers the change would provide useful insights about models behaviour that would help with accountability of the models [15].

5.3 Computational Tests

Arnold *et al*. [23] discussed the concept of the "big red button" that enables developers or users to divert or interrupt the system, however, the downfall of such approach is that it concentrates on the point where the system has already 'gone rogue'. A more prominent solution would be a preventative measure, implemented in the form of tests that would check whether a model behaves the way it was anticipated to behave. Such tests would also allow registering the errors that occurred if model failed the tests. Moreover, these tests could be applied to "hidden information", which is not possible to visualise and is hidden from the developers [21, 22, 23].

6 Conclusion

In conclusion, it is important to address the challenges of Accountable ML as ML is being utilised in more and more areas that affect human lives and the society. This results in more regulations being applied to automated decision making, which increases the pressure of algorithms being developed with accountability implemented within them. Our suggestions are that in the future, the developers would enrich Accountable ML frameworks with information about causality, decision provenance and would implement computational test frameworks that would check the behaviour of models.

References

1. General Data Protection Regulation. Online: https://gdprinfo.eu (2021)
2. Pomerleau, D.A.: Efficient Training of Artificial Neural Networks for Autonomous Navigation. Neural Computation 3(1), pp.88–97 (1991)
3. Hevelke, A. and Nida-Rümelin, J.: Responsibility for crashes of autonomous vehicles: an ethical analysis. Science and Engineering Ethics 21(3), pp. 619-630 (2015)
4. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data Society, 3(2), p.2053951716679679 (2016)
5. Model. Card Toolkit. Online: https://www.tensorflow.org/responsible_ai/model_card_toolkit/guide (2021)
6. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T.: Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency. pp. 220- 229 (2019)
7. ML Metadata. Online: https://www.tensorflow.org/tfx/guide/mind (2021)
8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daum´e III, H. and Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
9. Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P. and Mitchell, M.: Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 560-575 (2021)
10. Arnold, M., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K.N., Olteanu, A., Piorkowski, D. and Reimer, D.: FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), pp.6-1 (2019)
11. Hind, M., Houde, S., Martino, J., Mojsilovic, A., Piorkowski, D., Richards, J. and Varshney, K.R.: Experiences with improving the transparency of ai models and services. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-8 (2020)
12. Richards, J., Piorkowski, D., Hind, M., Houde, S. and Mojsilovic, A.: A Methodology for Creating AI FactSheets. arXiv preprint arXiv:2006.13796 (2020)
13. Piorkowski, D., Gonz´alez, D., Richards, J. and Houde, S.: Towards evaluating and eliciting high-quality documentation for intelligent systems. arXiv preprint arXiv:2011.08774 (2020)
14. Mohseni, S., Zarei, N. and Ragan, E.D.: A multidisciplinary survey and frame- work for design and evaluation of explainable AI systems. arXiv preprint arXiv:1811.11839 (2018)
15. De Laat, P.B.: Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability?. Philosophy technology, 31(4), pp.525-541 (2018)
16. Ananny, M. and Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media society, 20(3), pp.973-989 (2018)
17. Cambridge Dictionary. Online: https://dictionary.cambridge.org/dictionary/english/accountability (2021)
18. Naja, I., Markovi, M., Edward, P., Cottril, C.: A semantic framework to suppost AI system accountability and audit. In: ESWC 2021. p. in press. Greece (2021)
19. Kacianka, S. and Pretschner, A.: Designing Accountable Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 424-437 (2021)

20. Singh, J., Cobbe, J. and Norval, C.: Decision provenance: Harnessing data flow for accountable systems. IEEE Access, 7, pp.6562-6574 (2018)
21. Nushi, B., Kamar, E. and Horvitz, E.: Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 6(1) (2018)
22. Lepri, B., Oliver, N., Letouz´e, E., Pentland, A. and Vinck, P.: Fair, transparent, and accountable algorithmic decision-making processes. Philosophy Technology, 31(4), pp.611-627 (2018)
23. Arnold, T. and Scheutz, M.: The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics and Information Technology, 20(1), pp.59-69 (2018)