2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN) | 978-1-6654-4470-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/WAIN52551.2021.00021

Practices for Engineering Trustworthy Machine Learning Applications

Alex Serban*†
*ICIS, Radboud University,
†Software Improvement Group
a.serban@cs.ru.nl

Koen van der Blom‡, Holger Hoos‡ and Joost Visser‡
‡LIACS, Leiden University,
The Netherlands

***Abstract*—Following the recent surge in adoption of machine learning (ML), the negative impact that improper use of ML can have on users and society is now also widely recognised. To address this issue, policy makers and other stakeholders, such as the European Commission or NIST, have proposed high-level guidelines aiming to promote trustworthy ML (i.e., lawful, ethical and robust). However, these guidelines do not specify actions to be taken by those involved in building ML systems. In this paper, we argue that guidelines related to the development of trustworthy ML can be translated to operational practices, and should become part of the ML development life cycle. Towards this goal, we ran a multi-vocal literature review, and mined operational practices from white and grey literature. Moreover, we launched a global survey to measure practice adoption and the effects of these practices. In total, we identified 14 new practices, and used them to complement an existing catalogue of ML engineering practices. Initial analysis of the survey results reveals that so far, practice adoption for trustworthy ML is relatively low. In particular, practices related to assuring security of ML components have very low adoption. Other practices enjoy slightly larger adoption, such as providing explanations to users. Our extended practice catalogue can be used by ML development teams to bridge the gap between high-level guidelines and actual development of trustworthy ML systems; it is open for review and contributions.**

***Index Terms*—software engineering, machine learning, trustworthiness, robustness**

I. Introduction

The recent increase in use of machine learning (ML) to process personal data has resulted in more attention from policy makers and advisory bodies aiming to protect consumer interests. A leading example is the independent high-level expert group on artificial intelligence (AI) set up by the European Commission [1]. Well-intentioned but improper development of ML components can cause unintentional harm [2]. Published guidelines from policy makers (e.g., [1], [3], [4]) clearly reflect a desire for ML components to be lawful, ethical and robust [1]. However, these guidelines do not specify actions, and usually come in the form of high-level checklists [5] or broad recommendations [3].

In this paper, we aim to bridge the gap between guidelines from policy makers and operational practices for developers and their immediate collaborators. The operational practices should become part of the ML development life-cycle, together with established software engineering (SE) practices. We build on previous work by Serban et al. [6], which compiled a catalogue of SE best practices for ML applications. While the catalogue covers a variety of SE practices – e.g., testing of ML components – it lacks practices in the areas of ethics and robustness. These areas of interest are commonly grouped under the term *trustworthy ML* [1], [7].

We complement the catalogue of practices from [6] with 14 new practices for trustworthy ML. These practices are mined from literature through a multi-vocal review and tackle various topics, such as testing for bias, assuring security, and having the application audited by third parties. For all practices, we summarise related work into an actionable body of knowledge including the practice intent, motivation, applicability, description and references. We also launched a global survey to measure adoption of the practices, and assess their effects. In this paper, we describe the survey and present early results. Moreover, we invite the community to contribute to the catalogue of practices, and more generally to a body of knowledge spanning trustworthy ML and SE.

We begin by discussing background information and related work (Section II). Following that, we present the process of mining practices from literature, the resulting practices, and early results from the survey (Section III). We close with a discussion and concluding remarks (Section IV).

II. Background and Related Work

The literature on SE for ML consists of challenges faced by practitioners in the adoption of ML [8], practices [6], guidelines [9], and design patterns [10]. These lines of work tackle issues related to the ML development process, which includes rapid iterations by multi-disciplinary teams [8], an experimental approach to software development [9], the need to tackle uncertainty of ML components [11], and a strong emphasis on automation of operational aspects [6].

However, none of these lines of work tackle issues related to the negative impact that improper use of ML has on society. For example, inappropriate testing for bias can have devastating effects on some social groups [2]. Public documents from policy makers and advisory bodies – such as the independent high-level expert group on AI set up by the European Commission [1] – acknowledge the potential negative impact of ML on society, and propose guidelines to address these issues. However, the recommended guidelines are not immediately operational, and need substantial interpretation and refinement to specific actions in a development context.

978-1-6654-4470-5/20/$31.00 ©2021 IEEE
DOI 10.1109/WAIN52551.2021.00021

In order to complement guidelines from policy makers, increase their operational applicability and relate them to established engineering practices, we mined practices for trustworthy ML from the literature using academic (white) and non-academic (grey) literature. The latter is known to benefit SE research by providing valuable information from experience [12], and its value in the context of SE for ML has been demonstrated before [6].

III. Practices for Trustworthy Machine Learning

Practice mining from literature. To extract the practices, we followed a similar process to [6], i.e., we ran a multi-vocal literature review on the topic of trustworthy ML. As sources of information we used Google (for grey literature search) and Google Scholar (for both white and grey literature search). We composed queries with synonyms or disambiguation of *trustworthy ML*. The queries were formed from two elements; the first suggesting the field of research, and the second suggesting sub-fields of trustworthy ML. For the first element, we used three possible variations – machine learning, deep learning, and AI. For the second element, we used multiple variations, inspired by [1] – robustness, privacy, fairness, bias, interpretable, transparent, ethical, and auditable. As an example, the first query was *machine learning robustness*.

Exclusion criteria were formulated to avoid duplicates and articles published before 2016, which are generally subsumed by later articles [6]. We included all articles which proposed challenges, requirements, practices or future directions on the topics summarised in the second element of the query. Moreover, we included all relevant documents from policy makers, as well as papers describing their content.

When interpreting the query results, we defined practices by identifying and merging common themes between challenges and solutions. The challenges provided the intent and motivation, and the solutions provided the description of the practice. This procedure is known as thematic analysis [13], and it is commonly used in qualitative SE studies. When compiling the practices, we emphasised their applicability and their match to different stages of the ML development process.

Resulting practices for trustworthy ML. In total, we identified 13 relevant articles, from which we compiled 14 new practices for trustworthy ML. Moreover, we included two additional articles, that better position the practices, or bridge ML and SE [14], [15]. To classify the practices, we used the taxonomy from [6], who provided evidence that a common taxonomy for the ML development process does not exist, and reconstructed a general taxonomy that is compatible with previous work. Since our work complements their catalogue of practices, it is natural to use the same taxonomy.

Our list of practices, together with their class and references, is shown in Table I. We further discuss the practices and their relationship to the requirements for trustworthy AI from [1], which is widely regarded as mature and authoritative. We plan to discuss the relationship of the practices with other high-level guidelines in future work.

TABLE I: Practices for trustworthy ML, along with the requirements (Req.) for trustworthy AI from [1], that they address. The requirements are (R1) human agency and oversight, (R2) technical robustness and safety, (R3) privacy and data governance, (R4) transparency, (R5) diversity, non-discrimination and fairness, (R6) societal and environmental well being, and (R7) accountability.

Nr.	Title	Class	Req.	References
T1	Test for social bias in training data	Data	R5	[16]–[19]
T2	Prevent discriminatory data attributes from being used as model features	Data	R5	[16]
T3	Use privacy preserving ML techniques	Data	R3	[7]
T4	Employ interpretable models whenever possible	Tr.	R1	[1], [7], [20]
T5	Assess and manage subgroup bias	Tr.	R5	[21], [22]
T6	Assure application security	Code	R2	[3], [7]
T7	Provide audit trails	Dep.	R7	[1], [7], [23]
T8	Decide trade-offs through an established team process	Tm.	R1	[14], [15], [24]
T9	Establish responsible AI values	Gov.	R4	[1], [7]
T10	Perform risk assessments	Gov.	R2	[1], [7], [23]
T11	Inform users on ML usage	Gov.	R4	[1], [25]
T12	Explain results and decisions to users	Gov.	R4	[1]
T13	Provide safe channels to raise concerns	Gov.	R4	[1], [7]
T14	Have your application audited	Gov.	R7	[1], [7]

The practices in the Data class describe topics such as testing for social bias, preventing discriminatory attributes from being used as inputs to ML components, and using privacy-preserving techniques. These practices address the requirements for *privacy and data governance*, and for *diversity, non-discrimination and fairness* from [1].

Testing for bias is also present in the Training (Tr.) class, where it refers to assessing whether a trained model exhibits subgroup bias, which may arise from balancing groups in the training data. Moreover, some techniques for managing subgroup bias involve changing the loss function of ML components, or performing calibrations [21]. These techniques are applied when training a model, and not during data preparation. Another practice for training relates to using interpretable models whenever possible. The Training practices address the requirements for *human agency and oversight*, and for *diversity, non-discrimination and fairness* from [1].

In the Coding (Code) class, there is a single practice related to ensuring security. This practice tackles ML security topics, such as robustness, and more traditional security concerns, such as penetration testing. The practice addresses the requirement for *technical robustness and safety* from [1].

In the Deployment (Dep.) class, there is a single practice related to logging and storing audit trails. This practice addresses the requirement for *accountability* from [1].

We also identified a single practice in the Team (Tm.) class, related to defining team processes for deciding on ML-specific trade-offs. For example, processes for deciding on prioritisation of interpretable ML methods over non-interpretable methods, on prioritisation of non-biased models with lower accuracy, or on whether to use ML in the first place. This

Test for social bias in training data *Data*

Intent
Identify instances of social bias in training data, [...].
Motivation
Bias in data is one of the main sources of unfairness [...].
Applicability
Testing for social bias in training data should be done [...].
Description
In order to avoid social bias in ML algorithms, it is imperative to continuously check that the training data [...].
Adoption
Adoption rates, grouped by distinct demographic factors.
Related
Practices 2, 5, and 10 from Table I.
References
[16]–[19]

Fig. 1: Example of practice in online catalogue.

practice addresses the requirement for *human agency and oversight* from [1].

The largest class of practices – Governance (Gov.) – covers organisational practices which guide the development of ML components. Examples of such practices include the adoption of responsible AI values, subscription to a code of conduct, risk assessments, and auditing of ML components by third parties. Moreover, the Governance class includes a series of practices based on enforceable laws, such as explaining results and decisions to users (i.e., satisfying the right to an explanation). These practices address the *accountability* and *transparency* requirements from [1].

To make the practices available to practitioners, we added them to an existing online catalogue[1] (previously built by Serban et al. [6]), consisting of detailed descriptions and concise statements of intent, motivation, applicability, related practices and references. A brief (and abbreviated) example is illustrated in Figure 1. A curated reading list with these references, further relevant literature, as well as a selection of supporting tools is maintained online[2]. Both catalogue and reading list are open for community contributions.

Practice adoption. In order to measure the adoption of the practices, we extended the survey from [6] with new questions for the practices in Table I. The extended survey allows joint assessment of practices for trustworthy ML and more established ML engineering practices from [6]. The survey's questionnaire was designed following the recommendations of Kitchenham et al. [26]. It is a cross-sectional observational (i.e., participants were asked at the moment of taking the questionnaire to what extent they adopted the practices) concurrent control study (i.e., participants are assigned to specific groups, enabled by several preliminary questions).

[1]https://se-ml.github.io/practices/
[2]https://github.com/SE-ML/awesome-seml

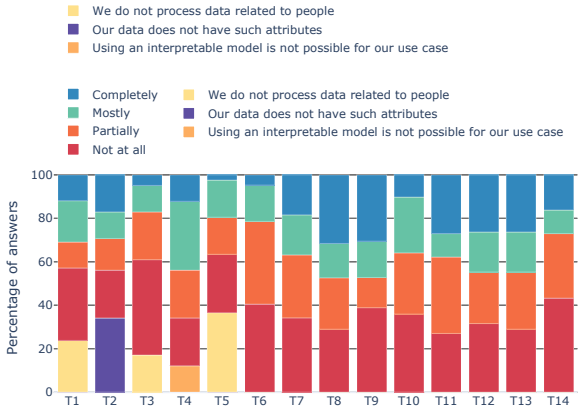

Fig. 2: Adoption of SE practices for trustworthy ML.

Besides 14 new questions with the practices in Table I, we added one question regarding the effect of adopting practices for trustworthy ML. The answers were standardised on a Likert scale, with four possible answers, designed to express degrees of adoption – e.g., "not at all" or "completely" – rather than degrees of agreement such as "agree" or "strongly agree". When the answer scale did not match the full range of possible answers, we added specific answers that helped to avoid noisy results (i.e., variations of "not applicable").

Practice adoption results. In the time span of three months, we received 103 answers, from which we filtered out participants that were not part of a team using ML. Moreover, we filtered on the percentage of questions that were answered in the prerequisites (at least 50 %) and in the practice adoption questions (at least 50 %), resulting in *42 complete responses*.

Figure 2 presents our preliminary results on practice adoption. The first 5 practices had specific answers to avoid noisy results, such as "using interpretable models is not possible for our use case" – which accounts for use cases where black box models from deep learning are heavily used (e.g., computer vision). Moreover, for questions related to social bias, we added an answer specifying that the data set does not contain personal data, and thus the practices are not applicable ("we do not process data related to people"). Similarly, for using discriminatory attributes, the participants were allowed to specify that the data sets used do not contain such attributes, through the "our data does not have such attributes" answer.

More than 20 % of the participants reported that they do not use personal data in their projects, or data with discriminatory attributes (T1, T2, T5). For practice T3 – i.e., use privacy-preserving ML techniques – a smaller percentage of participants chose the extra answer. We hypothesise that participants found the "Not at all" answer a better fit for this question.

Overall, the practice adoption for trustworthy ML is rather low, based on the large percentage of "Not at all" and "Partially" answers. The least adopted practice is T6, on assuring application security. This result is cause for concern, given the interest in both software and ML security from academia and industry. We hypothesise that although a large body of academic literature regarding security and robustness of ML exists, it is still limited in its applicability. For example, all

defences against adversarial examples – a known threat for ML components – have been breached [27]. This practice is also linked to practice T10, on performing risk assessments, which also has low adoption. Following that is practice T14, which involves having the application audited by third parties.

At the other end of the spectrum, the practices related to establishing team processes for deciding trade-offs (T8) and establishing responsible AI values (T9) have higher adoption. Similarly, practice T12 on explaining results to users, practice T13 on providing safe channels to raise concerns, and practice T4 on employing interpretable models have also have slightly higher adoption. We hypothesise that larger adoption of some practices (T4, T12, T13) is motivated by more mature legislation in the area (e.g., the right to an explanation). Other practices are more established in the engineering community (T8), or have higher practical feasibility (T9).

IV. Discussion and concluding remarks

We introduced a set of 14 new practices for trustworthy ML, which complements the existing catalogue of ML engineering practices from [6]. We argue that requirements related to the development of trustworthy ML (outlined by policy makers and regulatory bodies) can be translated into operational practices and should become part of the ML development life cycle. We also believe that the adoption of trustworthiness-specific and general ML engineering practices is interconnected; for instance, the practice of continuous integration [6] can make the practices for bias testing more effective. Using results from our extended survey, we plan to study the joint adoption of practices and their effects in more detail.

We emphasise that the practices introduced in this paper are meant to complement, not replace, the guidelines from policy makers and advisory bodies. We also note that no practices that directly address requirement R6 from [1] could be identified through our literature review. While practices (T1-T3, T9) address the societal aspects of R6, the environmental well-being is only addressed indirectly by T9. In the future, we plan to test the completeness of our catalogue through validation interviews with practitioners. Moreover, we welcome contributions from the community to further enrich the catalogue.

The analysis of the initial answers from our survey revealed low adoption of practices for trustworthy ML. In the future, we plan to further increase the number of respondents. New responses will also enable fine-grained analyses of the data – e.g., conditional analysis by demographics, correlation among practices, and prediction of effects from practice adoption. By repeating the survey at regular intervals, we will be able to monitor future adoption trends.

While in this paper we focused on the high-level requirements for trustworthiness from [1], similar guidance documents exist, or are being developed. We believe it is important to perform a broad and deep study of such documents, to identify where additional operational practices need to be provided. We also plan to conduct interviews with practitioners to identify new practices, that might not be described in literature, to further reduce the gap between policy and practice.

In this work we measured practice adoption through self-assessment by practitioners. A challenge for future work is to complement this with more objective, evidence based methods.

References

[1] High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI." https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. [Online; accessed 14-01-2021].

[2] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency*, vol. 81 of *PMLR*, PMLR, 23–24 Feb 2018.

[3] National Science and Technology Council (US). Select Committee on Artificial Intelligence, "The national artificial intelligence research and development strategic plan: 2019 update." https://www.nitrd.gov/news/National-AI-RD-Strategy-2019.aspx. [Online; accessed 14-01-2021].

[4] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, 2019.

[5] High-Level Expert Group on AI, "Assessment list for trustworthy artificial intelligence for self-assessment." https://ec.europa.eu/digital-single-market/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. [Online; accessed 14-01-2021].

[6] A. Serban, K. van der Blom, H. Hoos, and J. Visser, "Adoption and effects of software engineering best practices in machine learning," in *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, ACM, 2020.

[7] M. Brundage *et al.*, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," *arXiv:2004.07213*, 2020.

[8] A. Arpteg, B. Brinne, L. Crnkovic-Friis, and J. Bosch, "Software engineering challenges of deep learning," in *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, IEEE, 2018.

[9] D. Sculley, G. Holt, D. Golovin, *et al.*, "Hidden technical debt in machine learning systems," in *NeurIPS*, 2015.

[10] H. Washizaki *et al.*, "Studying software engineering patterns for designing machine learning systems," in *IWESEP*, IEEE, 2019.

[11] A. Serban *et al.*, "Towards using probabilistic models to design software systems with inherent uncertainty," in *ECSA*, Springer, 2020.

[12] V. Garousi, M. Felderer, and M. V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," *Information and Software Technology*, vol. 106, 2019.

[13] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, 2006.

[14] G. Ruhe, "Software engineering decision support–a new paradigm for learning software organizations," in *International Workshop on Learning Software Organizations*, Springer, 2002.

[15] J. Branke, K. Deb, K. Miettinen, and R. Slowiński, *Multiobjective optimization: Interactive and evolutionary approaches*, vol. 5252 of *Lecture Notes in Computer Science*. Springer, 2008.

[16] M. Hardt, "Fairness." https://www.youtube.com/watch?v=Igq_S_7IfOU, 2020. [Online; accessed 14-01-2021].

[17] L. T. Liu, M. Simchowitz, and M. Hardt, "The implicit fairness criterion of unconstrained learning," in *ICML*, PMLR, 2019.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv:1609.05807*, 2016.

[19] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, PMLR, 2013.

[20] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.

[21] Ú. Hébert-Johnson *et al.*, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *ICML*, PMLR, 2018.

[22] M. Kearns *et al.*, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *ICML*, PMLR, 2018.

[23] I. D. Raji *et al.*, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Conference on Fairness, Accountability, and Transparency*, PMLR, 2020.

[24] G. Fandel, "Group decision making: Methodology and applications," in *Readings in Multiple Criteria Decision Aid*, Springer, 1990.

[25] M. Mitchell *et al.*, "Model cards for model reporting," in *Conference on Fairness, Accountability, and Transparency*, PMLR, 2019.

[26] B. A. Kitchenham and S. L. Pfleeger, "Principles of survey research part 2: designing a survey," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 1, 2002.

[27] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: a comprehensive survey," *ACM CSUR*, vol. 53, no. 3, 2020.