EGOV-CeDEM-ePart

X-RAI: A Framework for the Transparent, Responsible, and Accurate Use of Machine Learning in the Public Sector

Per Rådberg Nagbøl*, Oliver Müller**

*IT University of Copenhagen, Denmark, pena@itu.dk
**University of Paderborn, Germany, oliver.mueller@uni-paderborn.de

Abstract: This paper reports on an Action Design Research project taking place in the Danish Business Authority focusing on quality assurance and evaluation of machine learning models in production. The design artifact is a Framework (X-RAI) which stands for Transparency (X-Ray), Responsible(R), and explainable (X-AI). X-RAI consist of four sub-frameworks: the Model Impact and Clarification Framework, Evaluation Plan Framework, Evaluation Support Framework, and Retraining Execution Framework for machine learning that builds upon the theory of interpretable AI and practical experiences tested on nine different machine learning models used by the Danish Business Authority.

Keywords: Machine Learning Evaluation, Government, Interpretability

Acknowledgement: Thanks to all the involved employees at the Danish Business Authority

1. Introduction

Recent years have seen breakthroughs in the field of AI, both in terms of basic research and development as well as in applying AI to real-world tasks. The AI Index 2019 Annual Report of the Stanford Institute for Human-Centered Artificial Intelligence (Perraul et al., 2019), which summarizes the technical progress in specialized tasks across computer vision and natural language processing, attests that AI is now on par or has even exceeded human performance in tasks such as object classification, speech recognition, translation, and textual and visual question answering. However, augmenting and automating tasks previously performed by humans can also lead to serious problems. Research studies and real-world incidents have shown that AI systems–or better the machine learning models they are based on-– can err, encode societal biases, and discriminate against minorities. These issues are amplified by the fact that many modern machine learning algorithms are complex black boxes whose behavior and predictions are almost impossible to comprehend, even for experts. Hence, more and more researchers and politicians are calling for legal and ethical frameworks for designing and auditing these systems (Guszcza et al. 2018). Against this background, the government of Denmark released a national strategy for AI in 2019. The strategy

covers a broad array of initiatives related to AI in the private and public sectors, including an initiative concerning the transparent application of AI in the public sector. As part of this initiative, common guidelines and methods will be created to enforce the legislation's requirements for transparency. As one of the first steps, the government launched a pilot project to develop and test methods for ensuring a responsible and transparent use of AI for supporting decision making processes (Regeringen, 2019). The pilot project takes place at the Danish Business Authority (DBA) in collaboration with the Danish Agency of Digitization. In this paper, we report on the first results of an Action Design Research (ADR) project accompanying the pilot project. The overall ADR project is driven by the following research question: How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting? To answer this question, the project draws on literature and theory on interpretability of machine learning models and practical testing on machine learning models in the DBA.

2. Explainable AI Through Interpretable Machine Learning Models

Modern machine learning algorithms, especially deep neural networks, possess remarkable predictive power. However, they also have their limitations and drawbacks. One of the most significant challenges is their lack of transparency. Complex neural networks are opaque functions often containing tens of millions of parameters that jointly define how input data (e.g., a picture of a person) is mapped into output data (e.g., the predicted gender or age of the person in the picture). Hence, it is virtually impossible for end users, and even technical experts, to comprehend the general logic of these models and explain how they make specific predictions. As long as one is only interested in the predictions of a black box model and these predictions are correct, this lack of transparency is not necessarily a problem. Broadly speaking, there are two alternative approaches to open up the black box of modern machine learning models (in the following see Lipton, 2018, Molnar, 2019, Du et al., 2020). First, instead of using black box deep learning models, one can use less complex but transparent models, like rule-based systems or statistical learning models (e.g. linear regression, decision trees). These systems are intrinsically interpretable, but the interpretability often comes at the cost of sacrificing some predictive accuracy. The transparency of these systems works on three levels: Simulatability concerns the entirety of the model and requires models to be rather simple and ideally human computable. Decomposability addresses interpretability of the components of the model, such as, inputs, parameters, and calculations.

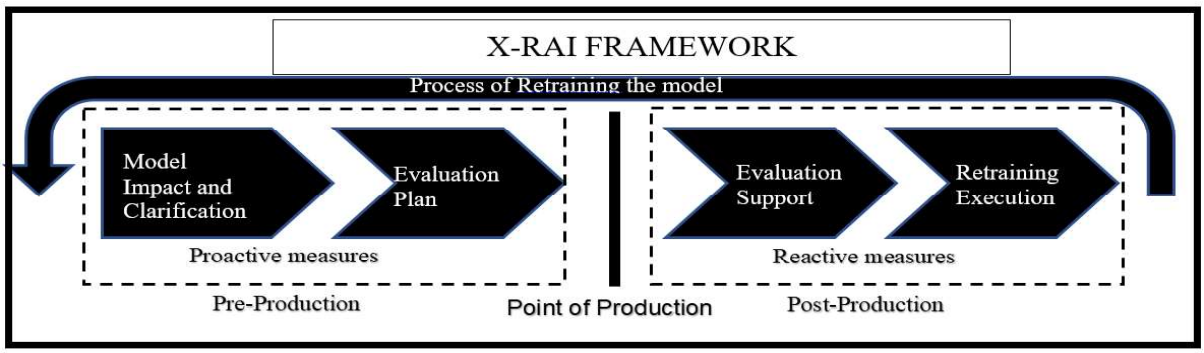
Consequently, decomposability requires interpretable model inputs and disallows highly engineered or anonymous features. Algorithmic transparency concerns the training/learning algorithm. A linear model's behavior on unseen data is provable, which is not the case with deep learning methods with unclear inner workings. Second, instead of using transparent and inherently interpretable models, one can develop a second model that tries to provide explanations for an existing black box model. This strategy tries to combine the predictive accuracy of modern machine learning algorithms with the interpretability of statistical models. These so-called post-hoc examinability techniques can be further divided into techniques for local and global explanations. Local explanations are explanations for particular predictions, while global explanations are explanations that provide a global understanding of the input-output relationships learned by the

trained model. In other words, a local explanation would explain why a concrete person on a picture has been predicted to be female, while global explanations would explain what general visual features differentiate females from other genders. Different types of post-hoc explanations exist. Text explanations use an approach similar to how humans explain choices by having a model generating explanations as a supplement to a model delivering predictions. Visualizations generate explanations from a learned model through a qualitative assessment of the visualization. Explanations by example let the model provide examples showing the decisions the model predicts to be most similar (Lipton, 2016). Local Explanations for particular predictions (Doshi-Velez & Kim, 2017) such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP for explaining feature importance (Lundberg, S & Lee, S, 2017). Focusing on the local dependence of a model helpful when working with neural networks being too incomprehensible to explain the full mapping learned satisfactorily (Lipton, 2016). When choosing which approach and technique to use in order to create an explainable AI system, it is worth to consider *why* there is a need for explanation (e.g., to justify decisions, enhance trust, show correctness, ensure fairness, and comply with ethical or legal standards), *who* the target audience is (e.g., a regular user, an expert user, or an external entity), *what* interpretations are derivable to satisfy the need, *when* is the need for information (before, during, or after the task), and *how* can objective and subjective measures evaluate the system (Rosenfeld, A & Richardson, A, 2019).

3. The X-RAI Framework as a Design Artifact

The X-RAI framework is an ensemble consisting of four artifacts (Fig. 1). First, the Model Impact and Clarification (MIC) Framework, which ensures that a ML model fulfills requirements regarding transparency and responsibility. Second, the Evaluation Plan (EP) Framework, which plans resource requirements and the evaluation of ML models. Third, the Evaluation Support (ES) Framework that facilitates the actual empirical evaluation of ML models and supports the decision whether a ML model shall continue in production, be retrained or shut down. Fourth, the Retraining Execution (RE) Framework, which initiates the process of sending an ML model back to the Machine Learning Lab (ML Lab) for retraining.

Figure 12: The X-RAI Framework



The first two artifacts are part of the decisive foundation for a steering committee regarding launching the ML model into production (pre-production). The last two artifacts support the

continuous evaluation and improvement of the ML model after it goes live (post-production). The design artifacts in ADR are solutions to problems experienced in practice and with theory ingrained. The problems must be generalizable outside the context of the project (Sein et al., 2011). X-RAI is a solution to problems experienced in the context of the Danish Business Authority where government officials are the intended end users. The government officials are, in our case, educated within the sciences of law, business, and politics as well as data scientists with plural backgrounds. Their expertise varies according to the governmental institution. X-RAI must be capable of involving and utilizing stakeholders with variating expertise without excluding some by setting an unachievable technological barrier of entry.

3.1. Model Impact and Clarification Framework

The MIC Framework has been applied and tested on four ML models--three times in its initial version and one time in its current version. The MIC is a questionnaire that enables the questionee to describe and elaborate on issues related to different aspects of ML related to transparency, explainability, responsible conduct, business objectives, data, and technical issues. The primary purpose of the MIC Framework is to improve, clarify, and guide communication between various stakeholders, such as developers with technical expertise, caseworkers with expertise in the ML models decision space and management. The idea of the MIC Framework derives from an analysis of the Canadian Algorithmic Impact Assessment (AIA)[1] tool that was found to have a strong link to the Canadian directive on automated decision-making[2]. MIC differs from AIA since it is grounded in theory and business needs instead of legislation. The algorithmic information in Box 1 contains information about the ML model. Box 2 is filled out by the future owner of the system enabling them to state their needs concerning the use, explainability, transparency, users, and accountable actors. Box 3 builds directly on Lipton's descriptions of transparency with the following three sub-levels: simulatability, decomposability, algorithmic transparency. In addition, it builds on types of post-hoc interpretability with the following approaches: text explanations, visualization, local Explanations, and explanation by example (Lipton, 2016). These are supplemented with three concrete explainability methods, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP (Lundberg, S & Lee, S, 2017). The output verification is bound to the fact that ML models in the DBA are decision-supportive, not decision-making, which reduces the need for an explanation if the end-user can validate the truthfulness of the model output instantly. Box 4 focuses on the data dimensions of the ML model including the relation to data sources and other ML models. Box 5 explains every feature to avoid opaque ML models due to highly engineered or anonymous features (Lipton, 2016) and supplements methods such as SHAP (Lundberg, S & Lee, S, 2017). Box 6 draws on the special categories from the 2016 European Union's General Data Protection Regulation[3] and the 2018 Danish Data Protection Act[4], repeating the questions on other data

1 See https://canada-ca.github.io/aia-eia-js/ and https://github.com/canada-ca/digital-playbook-guide-numerique/tree/master/en

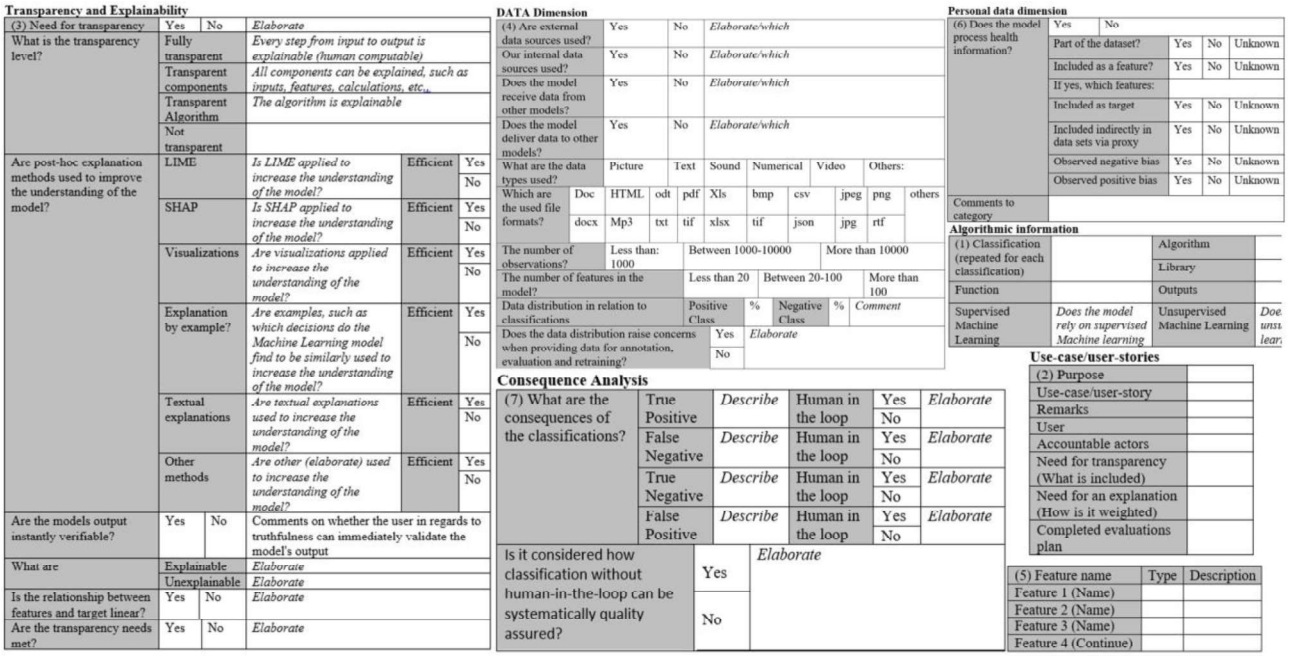
2 See https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

3 See https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN

4 See https://www.retsinformation.dk/Forms/r0710.aspx?id=201319 (all links last checked 01/06/20)

categories to avoid discrimination. Box 7 focuses on the consequences of the output, mitigation of consequences, and ensuring the responsible application of ML models. It takes inspiration from the confusion matrix enabling an easy estimate of the frequency of each outfall.

Figure 13: Model Impact and Clarification Framework



3.2. Evaluation Plan

The Evaluation Plan (EP) was applied and tested on eight ML models in three incrementally different versions. The EP structures the ongoing evaluation of a ML model throughout its lifetime and thereby illuminates the necessary resources for maintenance. The Evaluation Plan clarifies uncertainties such as time and frequency for the evaluation meetings, involved actors including roles and obligations, data foundation, and meeting preparation. The goal is to ensure that all ML models fulfill the defined quality requirements from the cradle to the grave. The theory is ingrained indirectly in the EP through the MIC framework. The choices made when using the MIC framework influences how the ML model can be evaluated. The ML model's degrees of transparency and explainability influences the possibilities of the evaluations. The evaluation detects data drift in a procedure similar to the application-grounded evaluation where the ML model is evaluated accordingly to domain experts performance on the task (Doshi-Velez & Kim, 2017). The EP encourages the first evaluation to be as early as possible due to the difficulties in predicting complex methods such as neural network on unseen data (Lipton, 2016).

Figure 14: Evaluation Plan Framework

(1) The name of the model and version number
(2) Participants for an example the application manager, caseworkers, ML lab etc.
(3) When is the first evaluation meeting?
(4) Expected evaluation meeting frequency: (How often are we expected to meet? And are there peak periods which we need to take into consideration?)
(5) Foundation for evaluation: For an example logging data or annotated data (Annotated data is here data where the domain experts classification is compared to the machine)
(6) Resources: (who can create the evaluation/training data, internal vs. external creation of training data, what is the quantity needed for evaluation, time/money)
(7) Estimated resource requirement for training, training frequency, and complications degree (procedure regarding regular bad performance)
(8) The Role of the Model: Is it visible or invisible for external users.
(9) Is the models output input for another/is the models input an output from another model.
(10) What are the criteria of success and failure (When does a model perform good/bad. How many percent?)
(11) Is there future legislation that will impact the model performance? (Including: bias, introduction of new requirements/legal claims, abolition of requirements/legal claims, bias, etc..
(12) When does the model need to be retrained?
(13) When should the model be mutet?

3.3. Evaluation Support

The Evaluation Support (ES) framework was applied five times on three different ML models in three incrementally changed editions.

Figure 15: Evaluation Support Framework

(1) The name of the model and version number
(2) Date of evaluation
(3) When was the last evaluation of the model?
(4) What was the result of the last evaluation?
(5) Participants in the evaluation meeting
(6) Who is doing the current evaluations?
(7) How many cases/documents has been processed in the evaluation (find minimum)
(8) Was the data used for the evaluation satisfying?
(9) Was is the result of the evaluation
(10) Has the performance of the model decreased?
(11) Has the performance of the model increased?
(12) What is the threshold set at?
(13) What is the history of the threshold setting?
(14) Should the threshold level be changed?
(15) Why is the threshold setting changed?
(16) Does the model still satisfy a business need? If not should the model then be shut down?
(17) Is there future legislation that will impact the model performance? (Including: bias, introduction of new requirements/legal claims, abolition of requirements/legal claims, bias, etc..
(18) Should the model be retrained based on the evaluation?

A fourth edition is ready for testing. The ES facilitates the evaluation of the ML model at the evaluation meetings. The domain specialist responsible for the ML model answers relevant fields in the framework before the meeting. The stakeholders complete the remaining framework collaboratively at the meeting and decide if the ML model shall continue in production, be retrained, or shut down. The ES strives to evaluate the ML model accordingly to the task as described in the

applications-grounded evaluation (Doshi-Velez & Kim, 2017). In our case, we let the caseworker that normally would do the task of the ML model evaluate the classifications and report it in the ES framework. The ES primarily focuses on fulfillments of performance requirements while it lets transparency and explainability be subcomponents of interpreting the reason for ML model performance. The reason is important if the model needs retraining.

3.4. Retraining Execution Framework

The Retraining Execution (RE) Framework was applied and tested two times on two different ML models in two incrementally changed versions. The RE initiates the process of sending a ML model back to the machine-learning lab for retraining. The retraining occurs when the ML model needs to improve performance and will continue to provide value. The RE framework focuses on the reusability of evaluation data and old training data for retraining, the occurrence of new technological possibilities, the detection and elimination of bias, changes in data types and legislation, the urgency for retraining, and if the input and output are related to other models. Transparency and explainability of the ML model become relevant when explaining a root cause for the need for retraining.

Figure 16: Retraining Execution Framework

(1) The name of the model and version number
(2) What is the reason for having the model retrained?
(3) What is the result of the last evaluation?
(4) Own suggestion of root cause, why does the model need retraining? (changes in document type, legislation, tenders etc..)
(5) Is new training data available for retraining (including estimation of required resources)
(6) How important is it to have the model retrained?
(7) Is the model dependent on other models? Yes/no – what is the status on them?
(8) What is the status of training data in the current situation? (Changes in document form, legislation, tenders, etc..)
(9) Can new data be added to the existing data or is there a need for a whole new training dataset? (What old training data is reusable?)
(10) Observed suspicion (bias against industry, gender, business type, etc.) Is it a problem? Yes/No
(11) Is the models output input for other models? Yes/no – status on them
(12) Is there developed algorithms that can solve the problem better since the model was put in production?
(13) "concluding text felt" Is there taken a decision regarding the model need to be retrained? (Has all stakeholder agreed on that the model has to be retrained?)

Data distribution becomes relevant if the data are skewed and slows down and thereby increases the cost in a data annotation process with the focus on providing training examples for the minority class. The use of the retraining execution framework restarts the X-RAI process by leading to the use of the MIC framework.

4. Conclusion and Outlook

The X-RAI framework was successfully developed, applied, and tested on nine different ML models used in the Danish Business Authority accordingly to the ADR principle of authentic and concurrent

evaluation (Sein et al.. 2011). The iterations have let to incremental changes in the frameworks. The frameworks are currently standard procedures and mandatory for all ML models developed by the ML Lab in the Danish Business Authority, which we conclude to be successful in the aspect of organizational adoption of artifacts and procedures. Artifacts must have theory ingrained accordingly to ADR (Sein et al.. 2011). Interpretability theory, including the subcategories of transparency and explanation, is ingrained into the frameworks. The lens provides a strong foundation for informing how the ML models work. Future work will focus on analyzing the evaluation data and using it to design IT artifacts and integrate them into the Danish Business Authority's IT-ecosystem. An additional theoretical lens will be ingrained in the artifacts to create a theoretical foundation for responsible conduct in the design.

References

Perraul, R. & Shoham, Y. & Brynjolfsson, E. & Clark, J. & Etchemendy, J. & Grosz, B. & Lyons, T. & Manyika, J. & Mishra, S. & Niebles, J. C. (2019). The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Guszcza, J. & Rahwan, I. & Bible, W. & Cebrian, C. & Katyal, V. (2018) Why We Need to Audit Algorithms. https://hbr.org/2018/11/why-we-need-to-audit-algorithms.

Regeringen (2019) Finansministeriet og Erhvervsministeriet: National strategi for kunstig intelligens

Molnar. C. (2020): Interpretable Machine Learning A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Doshi-Velez, F. & Kim, B. (2017) Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608v2

Du, M. & Liu, N. & Hu, X. (2020). Techniques for Interpretable Machine Learning. Communications of the ACM. Volume 63. Issue 1. https://dl.acm.org/doi/10.1145/3359786

Rosenfeld, A. & Richardson, A (2019). Explainability in Human-Agent Systems. arXiv:1904.08123v1

Sein, M.K. & Henfridsson, O. & Purao, S. & Rossi, M. & Lindgren, R. (2011) ACTION DESIGN RESEARCH. MIS Quarterly, Volume 35, Issue 1, page 37-56

Lipton, Z (2016) The Mythos of interpretability. Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. last revised 6 Mar 2017. arXiv:1606.03490v3

Lipton, Z (2018). The Mythos of Model Interpretability. ACM QUEUE. Volume 16, issue 3 https://queue.acm.org/detail.cfm?id=3241340

Lundberg, S & Lee, S (2017). A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

About the Authors

Per Rådberg Nagbøl

Per Rådberg Nagbøl is employed as a Ph.D. fellow at The IT University of Copenhagen and does a collaborative Ph.D. in collaboration with the Danish Business Authority.

Oliver Müller

Oliver Müller is Professor of Management Information Systems and Data Analytics at Paderborn University.