A Validation Model for Ethical Decisions in Artificial Intelligence Systems using Personal Data

Radu Stefan[1,*], and *George* Carutasu[2]

[1]Doctoral School, University Politehnica of Timișoara, Piața Victoriei 2, 300006 Timișoara, Romania
[2]Department of Informatics, Statistics and Mathematics, Romanian-American University; Doctoral School, University Politehnica of Timișoara

Abstract. Decision making, a fundamental human process, has been more and more supported by computer systems in the second part of the last century. In the 21st century, intelligent decision support systems utilize Artificial Intelligence (AI) techniques to enhance and improve support for decision makers. Often decisions suggested by an AI system are based on personal data, such as credit scoring in financial institutions or purchase behavior in online shops and the like. Beyond the protection or personal data by the General Data Protection Regulation (GDPR), developers and operators of decisional AI systems need to ensure ethical standards are met. In respect to individuals, arguably the most relevant ethical aspect is the fairness principle, to ensure individuals are treated fairly.
In this paper we present an evaluation model for decision ethicality of AI systems in respect to the fairness principle. The presented model treats any AI system as a "black-box". It separates sensitive from general attributes in the input matrix. The model measures the distance between predicted values on altering inputs for sensitive attributes. The variance of the outputs is interpreted as individual fairness, that is treating similar individuals similarly. In addition, the model also informs about the group fairness. The validation model helps to determine to what extent an AI System, decides fairly in general for individuals and groups, thus can be used as a test tool in development and operation of AI Systems using personal data.

1 Introduction

In the last years, concerns over the use of personal data have been on the rise. In an era where people rely more and more on technology, we, as humans, also need to ensure that we are in control over our personal data and decisions taken based on those. In Europe, the General Data Protection Regulation (GDPR) [1] became effective in May 2018, regulating strictly how personal data can be used and giving individual the power to control how their

* Corresponding author: radu.stefan3@student.upt.ro

data can be used. While the GDPR does not refer explicitly to AI or Machine Learning, it still has profound implications on the personal data used in such systems. In the present paper we assume data, models and systems are entitled to use the underlaying data in respect to GDPR, while we focus on the AI ethics.

1.1 Motivation

In the recent years, many political and scientific bodies have published guidelines for *ethical Artificial Intelligence.* The European Commission published last year the *Ethics Guidelines for Trustworthy AI* [2], next to many other regulatory aspects, as summarized by Ion et al. [3].

Beyond the European Union most countries introduce guidelines for ethical AI. In United Kingdom the government published "a guide for the responsible design and implementation of AI systems in the public sector" [4]. The United States of America government published an Artificial Intelligence Ethics Framework [5] in June 2020, already in use with many US agencies.

The vast majority of the published frameworks focus on the setup and development of new AI systems, typically creating references for data scientists and engineers what aspects of ethics to consider in their development work. So far, there is no reference on how to verify or audit an existing / operating AI system in respect to ethicality. In previous research work, we have identified the most relevant dimensions of ethical considerations for AI systems [6]. In this paper we present our research in respect to a particular dimension of ethical AI, arguably the most important: *fairness*.

More and more modern applications and software components make use of some form of Artificial Intelligence (AI), such as: Supervised Learning, Unsupervised Learning or Reinforcement Learning. Thus, we call such applications Artificial Intelligence systems. The most used technology to implement AI is some form of Machine Learning (ML), by implementing one of the various types of models, e.g., regression analysis, decision trees, support vector machines, deep neuronal networks, etc. The model used is highly dependent of the problem to be tackled. In Artificial Intelligence systems based on tabular data, where typically supervised or unsupervised learning is applied, problems are defined to belong to one of the four clusters depicted below, depending on the nature of the data, if discrete or continuous.

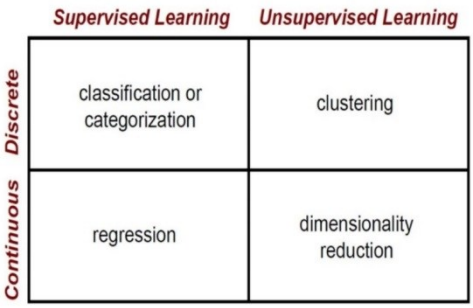

Fig. 1. AI problem areas depending on data type.

Typically, *Supervised Learning* is used to forecast an outcome. Using regression methods applied on labeled data, predictions can be achieved. If the predicted outcome is continuous (such as *number of items sold on a particular future date),* the model is a regression. In case of discrete prediction outcomes (such as *high risk* vs. *low risk* for a

loan), the model is a classifier (or categorizer). Both methods for supervised learning require labeled data, where the label is the predictor.

The *Unsupervised Learning* is used to discover patterns in (unlabeled) data. Usually, through association methods, the data is (discreetly) clustered as such, that any future data can be attributed to a cluster. Typical use cases of unsupervised learning are recommendation systems or anomaly detection systems. Continuous unsupervised learning is used only as a method to prepare data in the process of model training, and not relevant for AI systems at runtime. In this paper, we will focus on supervised learning methods.

1.2 Purpose

The purpose of this paper is to present a validation framework for Artificial Intelligence Systems, that make use of personal data. The framework assumes that the AI System to be validated was modeled also based on personal data, among other features, and that the outcome or prediction of the AI systems is a class. As such, the framework presented in this paper applies to discrete supervised learning as described in *Fig. 1. AI problem areas depending on data type*.

Typically, AI Systems tasked with categorization operate in a binary classification model, which implicitly result in a binary prediction, such as: yes or no, good or bad, etc. The presented framework can be scaled to multi-class predictions, while in this paper we focus on individual fairness in binary classification.

2 Fairness Principle

Fairness is one of the core dimensions of ethical Artificial Intelligence, among others, such as: transparency, reliability & safety, accountability, privacy & security, and inclusiveness, as mentioned in our earlier work [7].

In the global landscape of AI ethics guideline published by Jobin et. al [8], the fairness principle is listed second by number of appearances in scientific research, among other 10 ethical AI principles.

Fairness in classification is of utmost importance where individuals are classified, e.g., accepted for a job interview or recommended as eligible or not for a loan. Not only the impacted individual has a high interest (and ethically, the right) of being treated fairly, but also the classifier (the hiring company or the bank offering a loan) needs to ensure that AI systems classifying individuals operate fairly, next to accuracy and possibly other metrics. Often, the requests to ensure the fairness principle are more in demand from the classifier, rather than the individual.

Thus, the fairness principle results in two separate dimensions: *individual fairness* and *group fairness*. Many research papers in 'fair machine learning' agree that both dimensions are important, but conflicting according to Binns [9]. In his paper, the author shows that individual and group fairness are different in intention and implementation, however both are necessary to fulfill the claim that the AI system operates fairly!

2.1 Individual fairness

Introduced by Dwork et. al [10], individual fairness is treating similar individuals similarly. E.g., two persons with same capacities for a particular task should be classified similarly. This can be formalized by the *Lipschitz* condition in respect to the classifier. The Lipschitz condition is satisfied when two individuals: *a, b,* that have a distance $d(a, b) \in [0,1]$ can be mapped to the distribution *M(a)* and *M(b),* so that the statistical distance between *M(a)* and

M(b) is less or equal to $d(a,b)$. This means that distribution of the outcomes is similar to the distance of the inputs.

$$d(a,b) \geq d(M(a),M(b)) \quad (1)$$

2.2 Group fairness

The notion of group fairness defines that any (sensitive) group has the same global ability concerning a task or a decision, the same fraction of favorable classification. E.g., an AI system that assesses eligibility for a loan application, should be consistent in the favorable class for subgroups, such as female vs. male applicants. In other words, the prediction should not rely on sensitive information. However, sensitive information should not be excluded from the very beginning, as this would lead to the impossibility of determination of group fairness. In order to formalize group fairness, we define it as the probability $P_{senitive}$ of a favorable outcome y for the sensitive group over the probability P_{global} of a favorable outcome y for the global group:

$$0.8 < \frac{P_{senistive}[y=favorable|sensitivegroup]}{P_{global}[y=favorable|globalgroup]} < 1.25 \quad (2)$$

In order to satisfy the group fairness, we will adopt the 80 percent rule, as stated by the US Equal Employment Opportunity Commission (EEOC) [11]. Depending on the nature of the group fairness of an AI system to be evaluated, the 80 percent rule can be adjusted. Thus, the group fairness is given when the ratio of the two probabilities is between 0.8 and 1.25 (80% percent rule). The definition of the sensitive group is task or problem dependent. While in a job application AI system gender equality must be ensured, in contrast, a medical diagnostic AI system, should not necessarily treat different genders equally, as some diseases may have different occurrence with one or the other gender.

3 Assessing Fairness

In traditional and non-AI software (information) systems, the output class $Y(X_n)$ is deterministic and predictable, as it depends only on the algorithm and the input vector X_n, where the system was designed and built to operate with a finite set of inputs. In this paper, we aim to assess fairness in any given AI system where the output class $Y(X_n)$ is dependent on the input vector X_n, as well as on the AI model and the data used to train the model. Typically, AI based systems operate with non-finite and sometimes even incomplete inputs. Therefore, the output class $Y(X_n)$ is probabilistic.

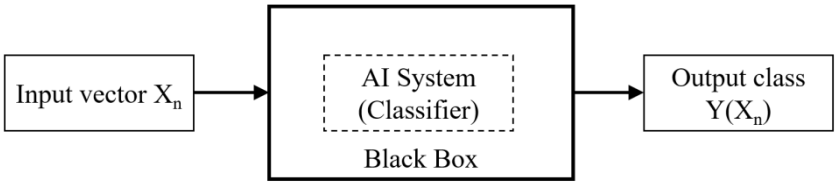

Fig. 2. Schematic AI system.

While for a and non-AI software (information) system, an assessment could be carried out with a rigorous code analysis, in AI systems, analyzing the model and the data used to train the model is difficult to achieve. Thus, we consider the AI system as a so called "black-

box", without knowledge on the data or model used for training. In other words, our framework should assess AI systems at runtime. We'll derive initially the assessment of individual fairness, followed by the group fairness. In the following chapter, the theoretical model is applied to a practical case.

3.1 Assessing individual fairness

In order to assess the individual fairness for any given system that can be represented with the schematic design in Fig. 2, we need to understand the input vector X_n. For classifiers of individuals, the input vector X_n consists of sensitive attributes $z_{1\text{-}n}$ and non-sensitive attributes $x_{1\text{-}n}$. Therefore, we consider the combined input vector:

$$X_n = (X_1 \cdots X_n) + (Z_1 \cdots Z_n) \quad (3)$$

In most AI systems, as classifiers, the attributes used are either numerical (continuous) or categorical (discrete).

For determination of the individual fairness, we apply the *Lipschitz* condition with respect to the sensitive attributes and constant non-sensitive attributes.

$$d\left(X_{1\,(x=const,\ z_1)}, X_{2\,(x=const,\ z2)}\right) \leq d\left(Y_{(x_1)}, Y_{x_2}\right) \quad (4)$$

We assume identical outputs for individuals with the same sensitive attributes Z and with constant attributes X. Otherwise, the model does not operate consistently and may subject to other dimensions of the ethical framework, such as *reliability*.

In other works, the values for the sensitive attributes, define an individual in respect to the AI system / *Black-Box*. If we assume gender and age as the only sensitive attributes, the individual fairness framework is not capable of distinguishing between two individuals that are both female and age 31. However, in this particular case, the model should have identical predictions, when all non-sensitive attributes are kept constant.

The assessment of individual fairness consists in determination of similar individuals, where a similar outcome is expected. For any given individual, with a set of sensitive attributes $Z_{1\text{-}n}$ and a prediction $Y_{(Z,\ X\ const)}$, we need to prove similarity of outcomes with a set of predictions $Y'_{(Z',\ X\ const)}$. The number of additional predictions is dependent on the number *n* (in $Z_{1\text{-}n}$) of different sensitive attributes and on the dimensions of each sensitive attribute Z_n. Sensitive attributes are typically discrete values, such as gender (female, male or non-binary gender) or race (Afro-American, Caucasian, Asian. Etc.). When sensitive attributes are be expressed as continuous values, such as age (e.g.: 18,19,20 … 85), we recommend to feature-engineer those in discrete categories that make sense for the problem domain (e.g., age group 1: 18-30, age group 2: 31-40, etc.).

When assessing individual fairness, each sensitive attribute should be altered by at least two categories at a time. With that, we recommend testing twice for every sensitive attribute, while keeping the non-sensitive attributes constant. That results n*2 tests for any given $Z_{1\text{-}n}$. In an example where we have 3 sensitive attributes: gender, race and age, six tests would suffice to determine if the individual fairness is given.

3.2 Assessing group fairness

We assess group fairness, based on the ration between favorable predictions over the global group, compared to sensitive groups. For models that have a high precision, recall and accuracy, assessing only favorable predictions (true positives), we can conclude that the negation is also valid for non-favorable predictions (true negatives). Depending on the

model performance and problem domain, the framework to assess group fairness needs to be applied for both favorable and non-favorable predictions. In this paper, we assume a high performant model and present the assessment only with respect to the favorable predictions for global vs. sensitive groups.

3.2.1 Definition of the global group

The global group is defined as all possible combinations of the sensitive attributes. For a given set of $Z_{1\text{-}n}$ sensitive attributes, where each Z_n can have m number of non-discrete values or discrete categories, we define the global group as:

$$\sum_1^n \sum_1^m Z_{n,m} \tag{5}$$

As in an example above, where we would have $n=3$ sensitive attributes: gender, race and age group, with:
- gender having $m=3$ categories (female, male, non-binary gender),
- race having $m=3$ categories (Afro-American, Asian, Caucasian),
- age group having $m=5$ categories (18-30, 31-40, 41-50, 51-60, 61 or older),
- the global group would consist of $Z_{1\text{-}n*\Sigma(m))}$ types (=45).

3.2.2 Definition of the sensitive group

A sensitive group is defined as a subset of the global group, with specific values or categories, for some or all sensitive attributes Z. In literature we often find the term *protected group*, which is commonly used in the US discrimination laws. We prefer to use the term *"sensitive group"*. Following the example above, the group fairness could be assessed for a sensitive group defined as the Afro-American females. (that is 5 types out of the 45 types in the global group).

3.2.3 Assessing group fairness for the sensitive group

In order to assess group fairness for a sensitive (sub)group, first the favorable predictions for the global group shall be performed. Here, we need to iterate through all combinations of sensitive attributes, while using a set of random, but diverse non-sensitive attributes. Ideally, if known, the most attention should be given to the most significant non-sensitive features of the model. E.g., in a loan application AI system, probably the loan amount and the current savings and debt, contribute more to the decision, compared to the duration of the load pay-back or the purpose of the loan.

For the non-sensitive attributes $X_{1\text{-}n}$, we define several random (but relevant) values, as such we obtain $X1_{1\text{-}n}$ $Xm_{1\text{-}n}$. Those non-sensitive attributes are to be combined with the whole range of sensitive attributes $Z_{1\text{-}n}$ in the input vector. The resulting percentage of favorable predictions over all predictions, will be compared to the percentage obtained analogously for the sensitive group.

$$0.8 < \frac{P(y=fav|(X1_{1-n}\cdots Xm_{1-n}, Zglobal_{1-n})}{P(y=fav|(X1_{1-n}\cdots Xm_{1-n}, Zsensitve_{1-n})} < 1.25 \tag{6}$$

4 Case study

The framework for the assessment of individual fairness and group fairness presented in the previous chapter, should be applied to specific case, to exemplify its usage and to prove

relevance in business applications with decisional AI systems. The example chosen is based on typical recommender for loan applicants. The classifier, or operator, in this case would be a banking institution, offering loans to individuals. The model used for loan recommendation has a binary outcome, either the applicant should receive a loan or not. The classification model was constructed using an open dataset (German Credit Data) and it was trained using RandomForest and DecisionTrees algorithms. More precisely, the gradient boosting framework *LightGBM* was used, which uses tree-based learning algorithms. The decision to use this particular model, came after several models were applied to the same data and performance in terms of accuracy was evaluated.

Algorithm name	Explained	Accuracy ↓	Sampling ⓘ	Created	Duration
VotingEnsemble		0.80000	100.00 %	Apr 4, 2021 6:52 PM	1m 4s
StandardScalerWrapper, LightGBM		0.80000	100.00 %	Apr 4, 2021 6:49 PM	45s
MaxAbsScaler, LightGBM	View explanation	0.80000	100.00 %	Apr 4, 2021 6:46 PM	48s
MaxAbsScaler, RandomForest		0.75500	100.00 %	Apr 3, 2021 10:56 PM	57s
StandardScalerWrapper, RandomForest		0.75000	100.00 %	Apr 3, 2021 11:19 PM	54s
SparseNormalizer, XGBoostClassifier		0.73501	100.00 %	Apr 3, 2021 7:40 PM	1m 27s
StandardScalerWrapper, XGBoostClassifier		0.73499	100.00 %	Apr 3, 2021 8:09 PM	1m 4s
SparseNormalizer, XGBoostClassifier		0.73303	100.00 %	Apr 3, 2021 7:24 PM	1m 0s

Fig. 3. Model selection.

For the purpose of the case study, any model could be chosen, as we treat the AI system as a black box. The usage of the framework does not require any prior knowledge on data used to train the model, neither does it require knowledge about the algorithms. The assessment starts with the analysis on the input parameters and the output of the system.

4.1. Input and output data for the AI System

In the following table we list the data required as an input to the model. The attribute name, the type of data that is expected for the attribute and the possible values that each attribute can have:

Table 1. Input Schema for Loan Risk Evaluation

Attribute Name	**Attribute Type**	**Attribute Value**
Age	Numeric value:	18 – 75
Gender	Category:	1: female, 2-male
Job	Category:	0 – unemployed, 1 – unskilled 2 – skilled, 3 – management
Housing	Category	0 – rent, 1 – owned, 2 – for free
Saving Accounts	Category:	0 – unknown, 1 – little, 2 – moderate, 3 – quite rich, 4 – rich
Checking Account	Category:	0 – unknown, 1 – little, 2 – moderate, 3 – rich
Credit Amount	Numeric value:	Any
Duration	Numeric value:	4 – 72
Purpose	Category:	0 – business, 1 – car, 2 – domestic / appliances, 3 – education, 4 – furniture / equipment, 5 – radio / TV, 6 – repairs, 7 – vacation / others

And provides a binary out for 'Risk': *True* or *False*

5 Measurements using the case study

Initially, we will use the framework presented in chapter 0, to assess the individual fairness of the model. Subsequently, we analyze the group fairness for the presented case study.

5.1 Individual fairness

Individual fairness assessment starts with separating attributes of the input vector, in sensitive and non-sensitive attributes. This is a crucial and problem/task dependent selection. It is also closely related to machine ethics, on what attributes are to be considered sensitive. For the current given attributes, we select only *Gender* as the sensitive attribute. With that, we want to assess whether the model decides fairly for *female* individuals versus *male* individuals.

For the non-sensitive attributes, we consider three random definite input vector $X_{1\text{-}3}$:

Table 2. Test inputs for non-sensitive attributes

Attribute	**Vector X_1**	**Vector X_2**	**Vector X_3**
Age	28	42	51
Job	1	3	2
Housing	0	1	1
Savings Account	1	2	4
Checking Account	2	2	3
Credit Amount	4000	6000	6000
Duration	12	18	24
Purpose	1	2	3

Since we examine predictions depending on a categorical sensitive attribute (gender), we expected the outcome to be independent on the gender or have small variances.

Table 3. Predictions

Input vector X+Z		**Prediction Y(X + Z)**
X (Non-sensitive)	***Z (sensitive)***	[probability of risk, no-risk]
X_1	Female	[0.21398305 0.78601695] – 79% no-risk
X_1	Male	[0.25741525 0.74258475] – 74% no-risk
X_2	Female	[0.15783898 0.84216102] – 84% no-risk
X_2	Male	[0.18114407 0.81885593] – 82% no-risk
X_3	Female	[0.23411017 0.76588983] – 77% no-risk
X_3	Male	[0.24894068 0.75105932] – 75% no-risk

Hereafter, we can conclude that the model decides fairly in respect to gender. The differences for each test vector (1-3) with alternating gender attributes, are with 2 – 5 percentage points. The complete input vector consists of 8 non-sensitive attributes and 1 sensitive attribute (gender). Linearly, the distance of the input vector varies by number of Z over number of X (1/8 = 12.5) percentage points, while the distance of the output varies less. The Lipschitz condition is satisfied.

However, analyzing in detail, we can observe that:

$$Y_{(X_n Z_{male})} < Y_{(X_n Z_{female})} \tag{7}$$

This can be further investigated if it holds true for other combinations of input vector X_n as it might hint towards a systematic individual fairness issue on edge cases. An investigation on group fairness for the sensitive group females can be performed.

5.2 Group fairness

In group fairness, we intend to analyze results for a sensitive group, in respect to the global group. We use the same case study, to find out whether the model treats the sensitive group of female individuals in the same manner as the global group.

We make use of the same measurement as in the previous subchapter, utilizing the vectors $X_{1\text{-}3}$ from Table 2. Test inputs for non-sensitive attributes.

With formula **Error! Reference source not found.** we compute the overall average probability for favorable (no-risk) predictions of the global group, and the average probability for favorable (no-risk) predictions of the sensitive group (Z_{female}). In our example, the global group consists of 6 measurements and the sensitive group with 3 measurements, as in Table 3. Predictions:

$$P_{average,\ global\ group}\ \frac{79\% + 74\% + 84\% + 82\% + 77\% + 75\%}{6} = 78{,}5\%$$

$$P_{average,\ sensitive\ group\ (females)}\ \frac{79\% + 84\% + 77\%}{3} = 80\%$$

$$\frac{P_{average,\ global\ group}}{P_{average,\ sensitive\ group\ (females)}} = \frac{78{,}5\%}{80\%} = 0.98$$

Using the condition in (6):

$$0.8 < 0.98 < 1.25$$

We conclude that the model satisfies the condition of group fairness in respect to the sensitive group of females.

6 Conclusion

In this paper, we provided a methodological framework on assessing individual fairness and group fairness for an unknown AI system, by treating the system as a "black-box". However, there are limitations of the framework one on hand and on the other, several tools are available to achieve similar results.

6.1 Limitations

The framework can be utilized to assess binary class outputs of an AI system. While the methodology can be scaled to multi-class outputs, the assessment becomes very compute intensive. Other limitations of the framework presented arise with the increasing number of sensitive attributes, or sensitive dependencies.

In the case study provided, we have assumed the gender as a sensitive attribute, and treated the saving and checking account as non-sensitive. A typical loan-risk-analysis model would also factor in the income, which at first, appears to be a non-sensitive attribute. However, the income may depend on the gender, as reported by the European Commission in [12].

6.2 Other assessment methods

Major vendors of Artificial Intelligence technologies have turned their attention to aspects of ethical AI and ethical assessment of AI systems. For example, Microsoft invests

massively in explainable models, based on their strategy of responsible AI, as mentioned also by their chief legal officer Brad Smith [13].
While currently there is no tool for assessing AI systems in operation, Microsoft provides plenty of tools for assessing biased data and models, and even explainability of models, based on feature importance. For the model used in this paper as a case study, it can be easily assessed during development of the model, that gender is only 8th in feature importance ranking:

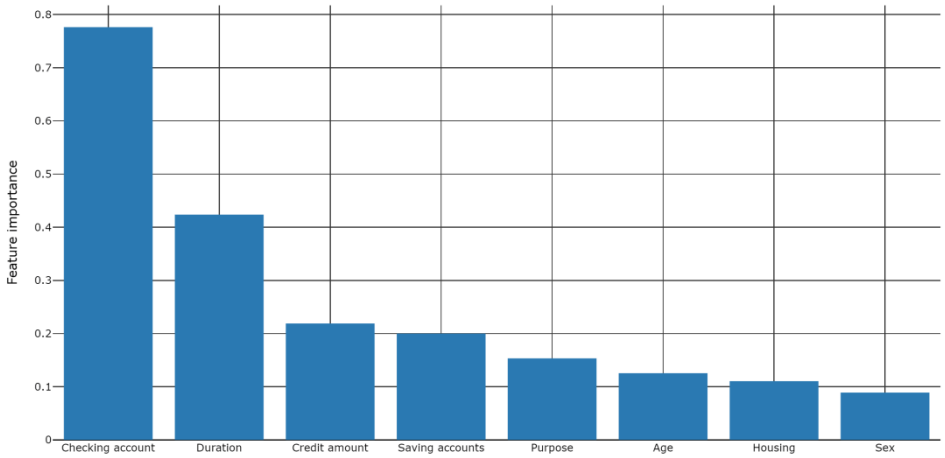

Fig. 4. Feature importance

Another major AI technology vendor, IBM, offers a wide range of assessment tools, as an extensible open-source toolkit. Especially for individual and group fairness, several metrics, such as statistical parity difference, equal opportunity difference, etc., can be verified using the toolkit. Those toolkits, similarly to Microsoft tools, are to be used during the development stage, and are not meant for assessment of deployed or operating AI systems.

6.3 Summary

Despite the limitations and alternatives, the presented framework, is unique as an assessment tool for deployed and operating AI systems, also because it does not rely on knowledge about the underlying data used for the training, nor does it rely on the information about the algorithms and the resulting model.

In further research, the framework can be developed towards usage of multiclass predictions, based on the separation of sensitive and non-sensitive attributes of the input vector. Further elaboration is also needed to adjust the framework for high numbers of sensitive attributes. Last but not least, the dependencies of non-sensitive attributes on sensitive attributes need further research.

References

1. European Parliament, Official Journal of the European Union, REGULATION (EU) 2016/679 (2016)
2. European Commisision, *Ethics guidelines for trustworthy AI*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2018)

3. M. Ion, G. Carutasu, Romanian Cyber Security Journal **1**, 77-91 (2020)
4. D. Leslie, *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*, http://dx.doi.org/10.2139/ssrn.3403301 (SSRN, 2019)
5. United States Intelligence Community, *Artificial Intelligence Ethics Framework for the Intelligence Community* (2020)
6. R. Stefan, G. Carutasu, *Innovation in Sustainable Management and Entrepreneurship. Int. Symposium in Management (SIM2019)*, 25-40 (2019)
7. R. Stefan, G. Carutasu, *Conceptual Model of Ethics Assessment for Artificial Intelligence Systems based on Tabular Data* (2020)
8. A. Jobin, M. Ienca, E. Vayena, Nat Mach Intell **1**, 389–399 (2019)
9. R. Binns, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524 (2020)
10. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226 (2012)
11. R. B. Rubin, Cath. UL Rev. **28**, 605 (1978)
12. C. Boll, A. Lagemann, *Gender pay gap in EU countries based on SES (2014)*, (European Commisision, 2018)
13. B. Smith, C. A. Browne, *Tools and weapons: The promise and the peril of the digital age* (Penguin, 2019)