A tool supported framework for the assessment of algorithmic accountability

1st Eleni Tagiou
University of Patras
P. O. Box 26441
Patras, Greece
tagiou@ceid.upatras.gr

2nd Yiannis Kanellopoulos
Code4Thought
P.O. Box 26441
Patras, Greece
yiannis@code4thought.eu

3rd Christos Aridas
Code4Thought
P.O. Box 26441
Patras, Greece
chris@code4thought.eu

4th Christos Makris
University of Patras
P. O. Box 26441
Patras, Greece
makri@ceid.upatras.gr

***Abstract*—Algorithmic decision making is now being used by many organizations and businesses, and in crucial areas that directly affect peoples' lives. Thus the importance for us to be able to control their decisions and to avoid irreversible errors is rapidly increasing. Evaluating an algorithmic system and the organization that utilizes it in terms of accountability and transparency bears certain challenges. Merely these are the lack of a widely accepted evaluation standard and the tendency of organizations that employ such systems to avoid disclosing any relevant information about them. Our thesis is that the mandate for transparency and accountability should be applicable to both systems and organizations. In this paper we present an evaluation framework regarding the transparency of algorithmic systems by focusing on the way these have been implemented. This framework also evaluates the maturity of the organizations that utilize these systems and their ability to hold them accountable. In order to validate our framework we applied it on a classification algorithm created and utilized by a large financial institution. The main insight for us was that when organizations create their algorithmic systems, accountability and transparency might be indeed recognized as values. However, they are either taken into account at a later stage and from the perspective of control or they are simply neglected. The value of frameworks like the one presented in this paper is that they act as check-lists providing a set of best-practices to organizations in order to cater for accountable algorithmic systems at an early stage of their creation.**

***Index Terms*—fairness, accountability, transparency, evaluation**

I. Problem Description

In many organizations, the number of key or crucial decisions that are being made by algorithms and not by people, is increasing. Algorithms, no matter how accurate they may be, will always be prone to mistakes, essentially because they are programmed by us, people. This means that as the adoption of algorithmic decision making is increasing, so does the rate of mistakes made by these employed algorithms [11]. As also [25] notes, *authority is increasingly expressed algorithmically* and decisions that used to be based on human intuition and reflection are now automated [7]. So, transparency over how these systems work matters not as an end in itself but merely as means towards accountability. Already in the last few years, individual efforts have been made to improve both human relationship with algorithms and the transparency of their decisions. Efforts have also been made to understand and integrate justice into machine learning algorithms (e.g. [3], [14], [20] and [27]).

In this paper we present a framework whose goal is to assess the transparency and accountability of algorithmic systems. This term (*algorithmic systems*) denotes systems that include not only algorithms but also human presence. In other words, the way an organization creates and utilizes an algorithm is crucial for its objective and unbiased operation. Our thesis is that the mandate for transparency and for accountability should be imposed on both of them. An organization that caters for accountability and an algorithm designed as such and is transparent can gain and have respectively the following benefits:

- There can be trust between the organization utilizing the algorithm and those affected by its output (be it clients, citizens or simple users), since the results can be explained,
- Improving the algorithm's output, since identified weighting factors and thresholds can be calibrated/fine- tuned if needed,
- Rendering the algorithm more persuasive, since its reasoning will be easier to explain.

This is not an easy task and the main challenges are the following:

- In several cases precision is preferred in expense to transparency. For instance, in the case of deep learning we need to typically comprehend the relationships among thousands of variables computed by multiple runs through vast neural networks. A task that is impossible for human brain.
- Organizations tend to keep their algorithms secret claiming they want to preserve valuable intellectual property or avoiding the risk of them getting gamed (e.g. in the case of credit scoring algorithms),
- There is no widely accepted industry-standard that defines how an algorithmic system should be evaluated in terms of transparency and accountability and what are the suggested criteria.

The framework presented in this paper extends the work of [21] which is utilizing terminology and resources from the work of [9] that mainly focuses on the news and media

domain, and [16] which is geared towards how organizations provide for accountability.

The main characteristics of the presented framework are:

- It is business-domain and technology agnostic, so it can be operationalized at any type of organization or business and algorithm,
- It is not intrusive as it doesn't require any data or input risking to disclose the specifics of an algorithmic system. It merely consists of a set of questions that require experts' input.

We did apply this evaluation framework at a large financial institution that developed a classification algorithm for their web and mobile banking platform. In Section 2 we elaborate on the related work in this area, while in Section 3 we present our framework and in Section 4 we discuss the results and the conclusions from our case study. Finally, at Section 5 we discuss the next steps of our research.

II. Background and Related Work

A. Decision making algorithms

As mentioned in the previous section, automated decision making based on algorithmic systems is being widely applied in a variety of areas, such as justice, journalism, healthcare, finance, education, government and others. A typical algorithmic system receives a data set as input in order to process it properly using some parameters and delivers a result by deciding and solving the problem that has been set by the organization (be it either government or private) that employs it.

Currently the public discourse is focusing on how these systems can be controlled and held accountable. The framework presented in this paper and is an extension of the work of [21] is following the *Step In* approach as described in the [7] article and its goal is to empower experts to control making algorithmic systems at technical as well as organizational level by rendering them transparent and accountable respectively. According to [7] we need people who know how to control and modify the work of computers. These people then can monitor and judge the correctness of an automated decision and avoid any error. As the article suggests as an example *"Ad buying in digital marketing is almost exclusively automated these days, but only people can say when some "programmatic" buy would actually hurt the brand and how the logic behind it might be tuned."*

Similarly also to the view of [4] which indicates, *"data is socially constructed"* our work reflects the viewpoint that algorithmic systems are being designed and implemented by humans whereas at the same time consume data constructed in one way or another from humans. That is why we want our framework to cover both social/organisational, human-related and technical aspects.

B. Algorithms' Evaluation Models

As mentioned in the previous subsection, algorithms are being created by people, and although they have a kind of logic, they cannot understand their possible mistakes or correct them. In order to maintain the accountability and transparency of algorithmic systems, we can design evaluation models that control them. Our research indicates that algorithmic evaluation models can be divided into two categories.

1) Generalized evaluation models: The first category concerns the model that is presented in this paper and includes models designed to evaluate the algorithms in general, that is, regardless of their use and the domain of their application. The effort to improve the algorithmic accountability can generally be done in two ways:

i. Either by improving the algorithm itself, namely by penetrating its mode of operation and changing the methods by which it decides [24], [28], [5], [17] and [6],
ii. Or by evaluating both the algorithm itself and the people who create and use it, but without any interference in the way that the algorithm makes a decision [11]. In the same category we can find the works [9], [10] and [12] on which our framework is based.

More specifically we utilise the basic concepts and the terminology terminology presented in this these papers present and one may use for evaluating both algorithms and organizations. These concepts are:

- Algorithmic Part: Algorithmic Presence, Data, The model, Inferencing.
- Organizational Part: Responsibility, Explainability, Accuracy, Audibility, Fairness, Human Involvement,

2) Specialized evaluation models: The second category includes assessment models designed for a specific type of algorithm. Such an example is the work of [13], which focuses on text-sorting algorithms that detect malicious comments. Also, in the field of justice, research has been conducted on algorithmic accountability, whether this concerns pre-trial decisions [8] or classification of crimes [19]. Another area where algorithmic accountability matters is social workers [26].

Finally, what we can conclude is that most efforts for maintaining the accountability of algorithms are aimed at improving the classifiers used by the algorithms and as well as their input data. This is observed at both generalised and specialized models. This can be done as we observed, merely by using statistics, probabilities or by being based in general at mathematical relationships. But is this the right direction? Our opinion is "no". People's responsibility on rendering their systems transparent ant accountable does not stop by just improving the classifiers of the algorithms they use. Algorithms are created and used by people for people, so so they should be held accountable as well. As we will see in the next section, the evaluation model we present, prescribes for both algorithms and people a series of requirements to satisfy towards accountability and transparency.

III. Design of the Framework

In this section we describe in detail the evaluation framework that we have developed , which extends the work of [21] that was based on [9], [10] and [12]. The main characteristics

of our framework and main differences of the works previously mentioned are:

- It is business-domain and technology agnostic, so it can be operationalized at any type of organization or business and algorithm,
- It is not intrusive as it doesn't require any data or input risking to disclose the specifics of an algorithmic system. It merely consists of a set of questions that require experts' input.
- It is geared towards providing both qualitative as well as quantitative results to its users.

In the figure below, we present our framework whose scope includes the algorithms themselves as well as the organization which utilizes them and needs to cater for their accountability. Also, in the following sub-sections we analyze how it works, that is how the qualitative and quantitative evaluation of the algorithmic systems will be done through a series of questions.

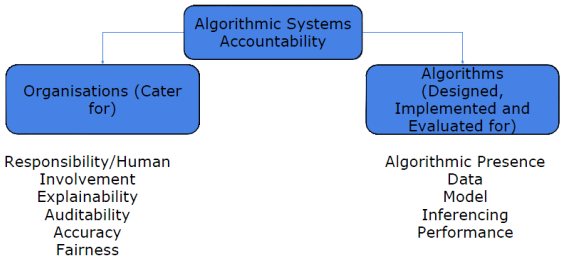

Fig. 1. Defining Algorithmic Systems Accountability from an Organizational as well as Algorithmic perspective.

A. Evaluation of Organizations

1) Responsibility: Definition:

- Algorithmic systems need to have available externally visible avenues of redress for adverse individual or societal effects,
- Organizations need to designate an internal role for the person who is responsible for the timely remedy of such issues.

Indicative Questions:

1.(i) Who is responsible if users are harmed by this product?
(ii) To what extent can a wrong decision affect users?

Absolutely	Much	Moderate	A little bit	Not at all
1	2	3	4	5

TABLE I
FIRST MEASUREMENT SCALE

2.(i) What will the reporting process and process for recourse be?
(ii) To what extent can an incorrect decision be dispensed with these procedures?

Not at all	A little bit	Moderate	Much	Absolutely
1	2	3	4	5

TABLE II
SECOND MEASUREMENT SCALE

3.(i) Who will have the power to decide on necessary changes to the algorithmic system during design stage, pre-launch, and post-launch?
(ii) How informed and objective is this person for the necessary changes? (The quantitative answer is given according to the second measurement scale)
4.(i) What are the roles of the people at your company who have direct control over the algorithm?
(ii) Are these people's roles crucial, that is, have they decision making power either from a technical or a business point of view? (The quantitative answer is given according to the second measurement scale)

$Total$ = sum of answers (1 to 5 for each answer)
Percentage of Responsibility $= \frac{100}{4*5} * Total\%$

2) Human Involvement: Definition:

- It requires explaining the goal, purpose, and intent of the algorithm, including editorial goals and the human editorial process or social- context crucible from which the algorithm was cast.

3) Explainability : Definition:

- Algorithmic decisions as well as any data driving those decisions should be explained to end-users and other stakeholders in non-technical terms.

Indicative Questions:

1.(i) Who are your end-users and stakeholders?
(ii) How much the people can understand the operation of the algorithm and how it makes a decision? (The quantitative answer is given according to the second measurement scale)
2. How much of your system / algorithm can you explain to your users and stakeholders? (The quantitative answer is given according to the second measurement scale)
3. How much of the data sources can you disclose? (The quantitative answer is given according to the second measurement scale)

$Total$ = sum of answers
Percentage of Explainability $= \frac{100}{3*5} * Total\%$

4) Auditability: Definition:

- Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.

Indicative Questions:

1.(i) Can you provide for public auditing (i.e. probing, understanding, reviewing of system behavior) or is there

sensitive information that would necessitate auditing by a designated 3rd party?

(ii) To what extent can public audit be made? (The quantitative answer is given according to the second measurement scale)

2.(i) How do you plan to facilitate public or third-party auditing without opening the system to unwarranted manipulation?

(ii) To what extent have you achieved this goal? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Auditability $= \frac{100}{2*5} * Total\%$

5) Accuracy: Definition:

- Sources of error and uncertainty throughout the algorithm and its data sources should be identified, logged and articulated so that expected and worst case implications can be understood and inform mitigation procedures.

Indicative Questions:

1.(i) What sources of error do you have and how will you mitigate their effect?

(ii) To what extent have you identified and addressed the sources of errors you have? (The quantitative answer is given according to the second measurement scale)

2. How confident are the decisions output by your algorithmic system? (The quantitative answer is given according to the second measurement scale)

3.(i) What are realistic worst case scenarios in terms of how errors might impact society, individuals, and stakeholders?

(ii) How much the realistic worst case scenarios can affect society, individuals and stakeholders? (The quantitative answer is given according to the first measurement scale)

4. Have you evaluated the provenance and veracity of data and considered alternative data sources? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Accuracy $= \frac{100}{4*5} * Total\%$

6) Fairness: Definition:

- Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g. race, sex, etc).

Indicative Questions:

1. Have you defined any particular groups which may be advantaged or disadvantaged, in the context in which you are deploying, by the algorithm / system you are building? (The quantitative answer is given according to the second measurement scale)
2. Have you defined and probably quantified what is the potential damaging effect of uncertainty / errors to different groups? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Fairness $= \frac{100}{2*5} * Total\%$

B. Evaluation of Algorithms

1) Algorithmic Presence: Definition:

- It involves the disclosure if and when an algorithm is being employed at all.

Indicative Questions:

1.(i) What is the problem at hand look like?

(ii) How much of the problem do you understand? (The quantitative answer is given according to the second measurement scale)

2.(i) Given the nature of the problem, which type of algorithm are you using?

(ii) To what extent can the algorithm handle the problem your organization tries to solve? (The quantitative answer is given according to the second measurement scale)

3.(i) Which elements are being filtered away?

(ii) To what extent these elements are being filtered away? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Algorithmic Presence $= \frac{100}{3*5} * Total\%$

2) Data: Definition:

- Quality: This involves their accuracy, completeness, and uncertainty, as well as its timeliness, representativeness of a sample for a specific population, and assumptions or other limitations.
- Handling: This includes data definitions, way of collection, vetting and editing (manually or automated).

Indicative Questions:

1.(i) How are various data labels gathered?

(ii) How fair is the choice of labels? (The quantitative answer is given according to the second measurement scale)

2.(i) How do they reflect a subjective or objective process?

(ii) To what extent do they reflect a subjective or objective process? (The quantitative answer is given according to the second measurement scale)

3.(i) How are you handling missing values?

(ii) To what extent can missing values affect the result? (The quantitative answer is given according to the second measurement scale)

4. Do incorporated dimensions have personal implications if disclosed? (The quantitative answer is given according to the first measurement scale)

$Total =$ sum of answers

Percentage of Data Evaluation $= \frac{100}{4*5} * Total\%$

3) The Model: Definition:

- It involves the model itself as well as the process followed for its construction.

Indicative Questions:

1.(i) What is its input?

(ii) Is it transparent? (The quantitative answer is given according to the second measurement scale)

2.(i) What are the algorithm's parameters used as input?

(ii) Do they include sensitive features? (The quantitative answer is given according to the first measurement scale)

3.(i) What are the features or variables used?

(ii) How meritocratic are they? (The quantitative answer is given according to the second measurement scale)

4.(i) Are they weighted? If yes, then what are their weights?

(ii) If yes, how fair is the way they are weighted? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Model Evaluation $= \frac{100}{4*5} * Total\%$

4) Inferencing: Definition:

- It involves the algorithm's evaluation in terms of its accuracy and error margin and their creator's ability to benchmark them against standard datasets and standard measures of accuracy.

Indicative Questions:

1. What is the margin of error? (The quantitative answer is given according to the first measurement scale)
2. What is the accuracy rate, and how many false positives versus false negatives are there? (The quantitative answer is given according to the second measurement scale)

3.(i) What kinds of steps are taken to remediate known errors?

(ii) To what extent do these steps help? (The quantitative answer is given according to the second measurement scale)

4.(i) Are errors a result of human involvement, data inputs, or the algorithm itself?

(ii) How easy is for human intervention to cause any errors? (The quantitative answer is given according to the first measurement scale)

(iii) How easy is for data inputs to cause any errors? (The quantitative answer is given according to the first measurement scale)

(iv) How easy is for the algorithm itself to cause any errors? (The quantitative answer is given according to the first measurement scale)

$Total =$ sum of answers

Percentage of Inferencing $= \frac{100}{6*5} * Total\%$

5) Performance Evaluation: Definition:

- It involves the selection of those metrics appropriate for the algorithm's performance evaluation and comparison. Selected metrics influence how one weights the importance of different characteristics in the results and their ultimate choice of which algorithm to choose.

Indicative Questions:

1.(i) Given the algorithm at hand, which are the evaluation metrics used?

(ii) How to ensure the validity of your evaluation metrics? (The quantitative answer is given according to the second measurement scale)

2.(i) What is the reasoning behind the selection of these metrics?

(ii) How fair is the reasoning for the selection of these metrics? (The quantitative answer is given according to the second measurement scale)

3.(i) How are these metrics being utilised and interpreted?

(ii) How meritocratic is this process? (The quantitative answer is given according to the second measurement scale)

$Total =$ sum of answers

Percentage of Performance Evaluation $= \frac{100}{3*5} * Total\%$

Note: The final score is based on the average of all the percentages found above.

C. Database design

Since we have designed the evaluation framework for the transparency and accountability of algorithmic systems, we need to create a database so that we can store our results for easier processing, as well as for future use as we will see in the next chapter.

1) Entity-Relationship Model: The following E-R model shows our database whose goal is to store all useful information for the future creation of an automated algorithm transparency and accountability model.

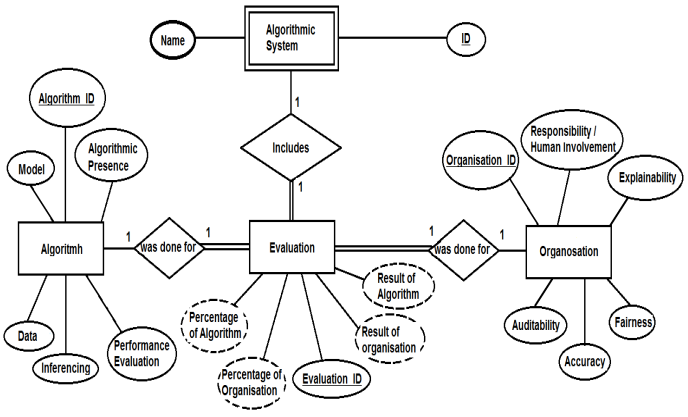

Fig. 2. Model of Entity-Relationships (E-R) of Algorithmic System Assessment.

First, we see that all relationships are one-to-one. This is because for each algorithmic system we make an evaluation and each evaluation concerns a specific algorithmic system. Then we see that the Algorithmic System entity is in a double rectangle. This is because it is a weak entity that gets its characteristics from the powerful entity "Assessment". In addition, this entity includes a feature that is encompassed by a severe shortage, because the Name attribute is complex, that is, it consists of a set of values. Also, the "Rating" entity includes four attributes (Organization Percentage, Algorithm Percentage, Organization Result, and Algorithm Effect) that

are included in an intermittent shortage. These attributes are called outputs, which mean that they are the result of a calculation. In particular, it is the average of the characteristics of the organization and the algorithm respectively and for their final evaluation. Finally, the ”Evaluation” entity cannot exist without the entities ”Algorithmic System”, ”Organization” and ”Algorithm”, so its participation in the corresponding correlations is total (the total participation is indicated by a double line).

2) Relational Data Model: Once we have created our E-R model, we will rely on this to design the relational data model based on which the database tables will be created. The steps we followed are shown below.

Step 1o:

- For each type of entity, we create a relationship that includes all simple attributes
- We represent the complex features with their elementary features.
- We ignore the traits produced.
- We select a candidate key as a primary key.

Algorithmic System	
ID	Name

Evaluation
Evaluation_ID

Organisation		
Organisation_ID	Responsibility/Human Involvement	Explainability
Auditability	Accuracy	Fairness

Algorithm		
Algorithm_ID	Algorithmic Presence	Data
Model	Inferencing	Performance Evaluation

Step 2o:

- For each 1: 1 association, we select the relation with the full participation and enter into it the key of the other relationship as a foreign key
- we also introduce all the attributes of the association into the relationship

Evaluation			
Evaluation_ID	Algorithmic System	Organisation	Algorithm

a) : Relationships as derived from the implementation of the transition steps from the entity - association model to the relational data model are shown in the figure below.

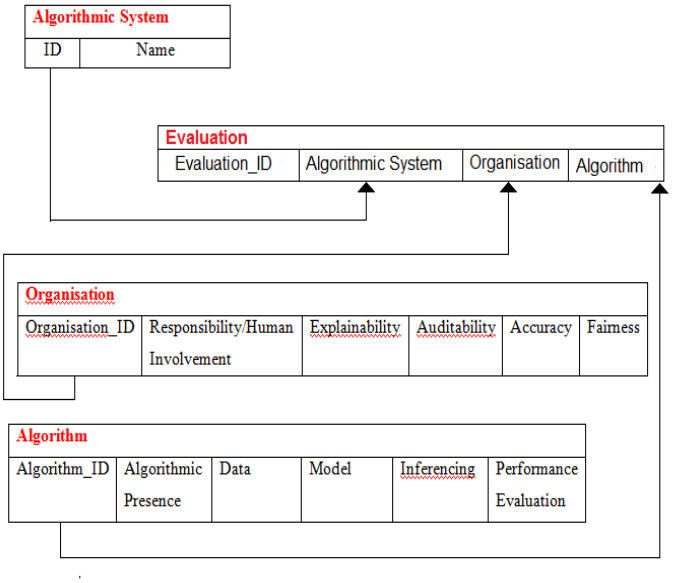

Fig. 3. Relational Data Model.

IV. Case Study

In the previous sections we described the problem and related challenges when trying to evaluate the accountability and transparency of a given algorithmic system. We propose a framework that helps overcoming the described challenges and represents our thesis that the mandate for transparency and for accountability should be imposed to both the algorithmic systems and the organizations that use them.

Our proposed framework is geared towards this direction. As described in Section 2, it extends the work of [21] and combines theoretical aspects from the work of [9], [10] and [12] together with empirical elements reflecting the current state of practice within corporations. We wanted to validate our model based on the following criteria:

- Its practicality and relevance merely to machine learning practitioners or in general those responsible for the technical implementation of an algorithmic decision making model,
- Its ability to be domain agnostic,
- Its capability to provide useful insights when operationalised.

For those reasons we created a questionnaire out of our framework and applied it at a large financial organization. We chose to apply our work there, as financial sector is a highly regulated one. Especially now and under the GDPR [18] directive, there are more demands on transparency and the way algorithmic systems form their decisions (e.g. Article 22 of GDPR that concerns automated individual decision-making, including profiling and Article 24 that concerns the responsibility of the controller). Thus our framework potentially can help them improving towards this direction.

The scope of our analysis was a classification [23] algorithm whose goal was to classify the users of a mobile and

web platform based on how digitally literate they are. The organization was trying to improve its users' experience by providing advanced functionality to the literate ones and basic to those who are classified as non-literate.

For this case we followed a process that consists of a series of sessions as described below:

- Framework presentation session: In this session we presented our framework to all participants/stakeholders in order to familiarise them with the questions and the goals of the exercise. An interesting observation at this point was the difficulty to identify the right persons to participate. As our framework's goal is to assess the technical as well as the organizational aspects of an algorithmic system, it does not come as a surprise that we had to align with multiple stakeholders, from both the IT as well as the Business departments of the organization.
- Data collection sessions: In these sessions our team worked in parallel with the teams responsible for implementing and managing the classification algorithm respectively. By interviewing those teams we managed to fill in the questions for the algorithmic as well as the organizational parts of our questionnaire.
- Data validation sessions: As soon as we collected all necessary data and upon expert analysis we presented our findings to both teams in order to validate our views and ensure we haven't omitted valuable information,
- Final Presentation session: As soon as our findings were validated we presented them to the higher management of the organization along with our recommendations on how they can improve from a technical as well as an organizational perspective.
- Evaluation of our framework - Feedback Session: Some weeks upon the final presentation we organised a session in order for our team to receive feedback from the financial organisation's practitioners regarding the validity and value of our framework to their way of working. A structured questionnaire was used for the purposes of this session.

V. Results

By applying our framework and following the evaluation process described in the previous section we managed to provide valuable insights to the client organization as well as to gather valuable feedback for the validation of the proposed framework. These are presented below.

A. Insights based on our framework's application

1) Qualitative evaluation: The conducted analysis indicated that the organization was partially in control of its algorithm. However, when it comes to the implementation of the algorithm itself, the main issue was the inability of reliable inferencing. The reason for this was the lack of a benchmark that the organization could employ to interpret the deducted results and assess their validity. The main findings from our analysis were:

- The organization has full control over the quality and the selection of the data used for feeding the algorithm. On the other hand, there seems to be no formal processes for handling any issues caused by the algorithm (e.g. harmed or disappointed/frustrated users).
- There is partial control over the algorithm design as the team decided to employ a set of classical data mining techniques such as clustering [15] and association rules [22] that are easier to explain compared to a neural network or a deep learning model.
- However, for the task at hand, that is to classify users based on their digital literacy, the selected approach was not solving it adequately as the created algorithm cannot provide reliable inferencing. The main reason for this is the lack of labeled data and of the ability to interpret the derived clusters and rules using human judgement. Nonetheless, state-of-the-art semi-supervised techniques like [1], [2] could be employed if an internal domain expert had annotated a few initial training examples.

2) Quantitative evaluation: Quantitative evaluation, in fact, is done by the organization itself. As we can see in Chapter 3, there are questions whose answer corresponds to a certain grade. The respondents of the organization who answered the questionnaire gave us an answer / grade on each such question. So our own task was to calculate the rates of answers in the way we refer in Chapter 3.

Below we present and analyze the results of the quantitative evaluation of the algorithm.
Percentage of Responsibility $= 50\%$
Percentage of Explainability $= 33.3\%$
Percentage of Auditability $= 50\%$
Percentage of Acciracy $= 60\%$
Percentage of Fairness $= 40\%$

Organizational total $= 46.7\%$

The quantified results show that the organization does not have the absolute control of the algorithm. From the percentage of explanation and fairness, we may conclude that the client organization is problematic in the following areas:

1. Lack of the ability to explain the algorithm and its decisions to end users.
2. The exemption of algorithmic decisions from unbiased, objective views.

From the deducted insights from both the qualitative and the quantitative parts of the assessment we may observe that they coincide regarding the ability of the client organization to cater for the accountability and transparency of their classification algorithm.
Percentage of Algorithmic Presence $= 66.7\%$
Percentage of Data Evaluation $= 70\%$
Percentage of Model Evaluation $= 75\%$
Percentage of Inferencing $= 53.3\%$
Percentage of Performance Evaluation $= 80\%$

Algorithm total $= 69\%$

The level of transparency of the classification algorithm seems to be higher. More analytically, the rate of data, model, and performance is over 70%, while the inferencing is moderate. This means that:

1. The quality of the input data is moderately good, as is their handling.
2. The model has the ability to make a fair decision, since it does not use sensitive features to a large extent.
3. The team responsible to implement the algorithm tries to use the most appropriate metrics for performance evaluation.
4. It is not possible to properly evaluate the classification algorithm in terms of accuracy, margin of error and the ability of their creator to compare it with standard datasets and measures of accuracy.

Similarly to the evaluation of the organizational part of our framework, here we can also observe the consistency of the conclusions between the quantitative and qualitative assessments.

B. Framework validation

As described already, based on the feedback session we organized with the client organization we had the ability to identify the value of our framework as well as the points for its improvement.

More specifically, and in a more general context, the practitioners from the client organization felt that it was extremely interesting and could reveal aspects that had escaped them and they had overlooked. They stated that using such a model can help them both to avoid mistakes when implementing an algorithm and to improve the maturity of the organization. Moreover, they consider that it is not time-consuming and should be gradually adopted by organizations using decision-making algorithms.

Regarding the questions used in our framework those were deemed as practical and adequate. However, what can be improved is their clarity and accuracy, so a suggestion was to add examples where it is necessary in order to avoid any misunderstandings.

As for the deducted results, the practitioners from the client organization did appreciate mostly the qualitative feedback whereas the stakeholders from management did find the quantitative results useful. In any case, establishing such an evaluation as a step of the implementation and deployment of an algorithmic system helps indeed building trust among the several departments of an organization as well as towards the algorithmic system itself.

Finally, an important point indicated by all teams of client organization is that complementing the framework with some techniques for explaining the decisions of their classification algorithm, would be of value as it will give them the ability to control the algorithm and the possibility to provide insights to interested parties and stakeholders.

VI. Conclusions and future work

As we have seen, the need for algorithmic transparency and accountability is growing increasingly along with the fast adoption of automated of decision-making algorithms. We believe that in order to be able to talk about transparency, it is not enough to control and evaluate only the algorithms, but also the people who create and use them. Our objective in this paper was to develop an evaluation framework (qualitative and quantitative) regarding the accountability and transparency of algorithmic systems.

We have applied our framework to a large financial institution and we have seen how it can actually influence and improve the accountability of both organizations and their algorithmic models. It is important to mention that the quantification of the results has helped our evaluation, since the results are easier perceived when they are presented in a measurable way.

In general we may deem the application of our framework in a real-life case as encouraging.

As future steps for our research work we have identified the following:

- We need to apply the model to more industrial cases in order to create a benchmark that can evaluate in an automated manner the accountability of a learning engineering model and the organization that uses it. Currently, the model is based on human judgement and expertise,
- We may explore the possibility to complement our evaluation framework with techniques explaining the results of algorithms.

Below we will elaborate on a high level on our ideas for future work:

- Having applied our framework many times, we will have a database which we can use as a pool of training data in order to able to predict whether an algorithmic system is accountable or not in a more automated manner by utilising machine learning techniques.
- A similar idea is to use all previous evaluations as a corpus in which by looking specific keywords (or the most relevant words) we may be able to deduct whether a system is accountable or not without having to undergo a more intensive manual process.
- Finally, by exploiting all the quantitative model evaluations in our database we may be able to create a benchmark which we can utilise in order to rate models in terms of their specific characteristics as defined by our framework.

References

[1] Aridas, C. and Kotsiantis, S. (2015). Combining random forest and support vector machines for semi-supervised learning. *In Proceedings of the 19th Panhellenic Conference on Informatics (PCI '15), ACM*, pages 123–128.

[2] Aridas, C. K., Kotsiantis, S. B., and Vrahatis, M. N. (2017). Hybrid local boosting utilizing unlabeled data in classification tasks. *Evolving Systems*.

[3] Barocas, S. and Selbst, A. (2015). Big data's disparate impact. *SSRN eLibrary*.

[4] Brussard, M. (2018). Artificial unintelligence: How computers misunderstand the world. *MIT Press*.

[5] Celis, E. L., Deshpande, A., Kathuria, T., and Vishnoi, N. K. (2016). How to be fair and diverse? *FATML*.

[6] Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. *Rome: FATML Workshop version*.

[7] Davenport, T. H. and Kirby, J. (2015). Beyond automation. *Harvard Business Review*.

[8] Davies, S. C., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *In Proceedings of KDD '17*, page 10 pages.

[9] Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Commun. ACM*, 59(2):56–62.

[10] Diakopoulos, N. (2017). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, page 3(3):398–415.

[11] Diakopoulos, N. and Friedler, S. (2016). How to hold algorithms accountable. *Commun. ACM*.

[12] Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., and Zevenbergen, B. (2017). Principles for accountable algorithms and a social impact statement for algorithms. *Digital Journalism*.

[13] Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification.

[14] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. *In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12*, pages 214–226.

[15] et al., A. S. (2017). A review of clustering techniques and developments. *Neurocomputing*.

[16] FATML (2017). Principles for accountable algorithms and a social impact statement for algorithms.

[17] Ferreira, M., Zafar, M. B., and Gummadi, K. P. (2016). The case for temporal transparency: Detecting policy change events in black-box decision making systems. *FATML*.

[18] GDPR.

[19] Hooker, J. N. and Kim, T. W. (2018). Toward non-intuition-based machine ethics.

[20] Islam, A. C., Bryson, J. J., and Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR, abs/1608.07187*.

[21] Kanellopoulos, Y. (2018). A model for evaluating algorithmic systems accountability. *Cornell University Library*, page 4 pages.

[22] Kotsiantis, S. and Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, pages 32(1): 71–82.

[23] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*.

[24] Naman, G., Mohammad, Y., and Boi, F. (2018). Non-discriminatory machine learning through convex fairness criteria. *Artificial Intelligence Laboratory*.

[25] Shirky, C. (2009). A speculative post on the idea of algorithmic authorityn.

[26] Vosloo, S. (2017). How to have algorithmic accountability in ict4d – your weekend long reads.

[27] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. (2015). Fairness constraints: A mechanism for fair classification. *In 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[28] Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., and Weller, A. (2017). From parity to preference-based notions of fairness in classification. *FATML*.