ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

A Method to Test the Ethics of Some AI Classifiers - The Example of School Dropouts Problem

Antonio Ballarin[1]* and Giovanni Fruscio[2]

[1]Apex, Rome, Italy and University Canada West, Vancouver, Canada

[2]ENERGENT Rome, Italy

ABSTRACT

If an AI artifact is an emulation of human behavior in relation to the performance of some activity and if the human being, in carrying out that activity, is required to respect a behavioral framework defined by laws, rules, regulations, procedures, best practices, etc., then the AI that emulates that human behavior is also required to respect the same behavioral framework. The idea of ethics tests is developed on this principle and, precisely on the basis of this principle, apragmatic methodology can be developed that can test the correspondence in the observance of the artefact to the behavioral framework within which it will necessarily be placed in its operation. The approach proposed in this work allows us to offer an extremely pragmatic solution to the search for an "ethical behavior" for AI artifacts, bypassing the difficult applicability of the complex and abstract legislation currently in force on this topic. In order to explain the applicability of this methodology to a concrete problem, this work considers theproblem of school dropout as an example and describes the construction of two classifiers, one based on a neural network and one on a decision tree, able to predict the phenomenon. The application of the methodology clearly shows how the explainability offered by a symbolic system, such as a decision tree, is not applicable as an element of explainability in the behavior of a neural classifier.

***Corresponding author**

Antonio Ballarin, Apex, Rome, Italy and University Canada West, Vancouver, Canada.

Received: October 01, 2025; **Accepted:** October 06, 2025; **Published:** October 15, 2025

Keywords: Ethics Tests, Explainability, Classifications System, bias, Categorical Polarization

Introduction

The pervasive adoption of Artificial Intelligence (AI) invarious everyday devices, such as smartphones, TVs, food processors, cars, etc., raises increasingly relevant ethical questions. This diffusion not only improves the efficiency and functionality of the devices, but also poses significant challenges in terms of liability, privacy, security, etc.

In particular

- The collection and processing of personal data by smart devices can compromise user privacy. Sensitive information can be used without consent or exposed to breaches.
- In the event of malfunctions or accidents caused by automated systems, the question of who is responsible (the manufacturer, the software or the user) is complex and requires clear regulation.
- Algorithms can perpetuate existing biases if not carefully designed. This can lead to unfair decisions in areas such as credit, hiring, etc.

Addressing the ethical problem solely with a prescriptive approach, that is, with regulatory systems that indicate a series of elements and/or rules to be respected in order to obtain "ethical" artefacts, is completely insufficient, because it does not guarantee in any way the actual "ethical nature" of the artefact produced.

These kinds of approaches seem more oriented to protect from possible liability those who actually implement the AI solution. In fact, by slavishly following "a rule" they could always claim non-involvement in the detection of behaviors not in line with what was expected by committees or authorities responsible for AI surveillance, exactly as it happens today, e.g., with the EU General Data Protection Regulation (GDPR) [1].

The solution for an "ethical" AI cannot be guaranteed by some laws. For this reason, the paths to follow for the purposes of protection for stakeholders who use an AI and that, if nothing else, complies with the same rules, laws, norms, best practices that a human being is required to respect, are necessarily two:

- Ethics tests,
- Algorithmic explainability.

This work, starting from a typical problem of predictive classification concerning the student dropout problem (here considered as example of a classification problem), summarizes a methodology, applicable to each problem of predictive classification for the realization of ethics tests and, at the same time, as a support for the achievement of the global explanations of a model.

The Importance of Ethics in Artificial Intelligence: The Approaches of The European Union and The United States

The need to build AI artifacts that are also ethic, is well perceived in the countries that primarily develop and use AI systems. The method suggested by the various laws on how to implement

these principles, is mainly focused on the creation of a checklist, that is, a series of actions to be carried out during the creation of the artefact, which, if carried out, should lead to the emergence of an AI system in line with what is expected and, therefore, characterized by ethics.

Integrating ethics into AI is an idea that is rapidly spreading within society and that goes hand in hand with the awareness of how technological development influences every aspect of daily life. For example, both the European Union (EU) and the United States (US) have developed distinct frameworks and approaches to address ethical considerations in AI, reflecting their respective legal, cultural, and political contexts.

The European Union's Approach

The EU has taken a proactive stance in regulating AI through the AI Act, which was approved in May 2024. This landmark legislation establishes a comprehensive legal framework aimed at ensuring that AI systems respect fundamental rights, safety, and ethical principles. The Act adopts a risk-based approach, categorizing AI applications into four tiers based on their potential for harm, with stricter regulations for higher-risk categories [2,3].

In addition to the AI Act, the EU has published Ethics Guidelines for Trustworthy AI [4]. These guidelines emphasize seven key requirements for AI systems:

- Human Agency and Oversight,
- Technical Robustness and Safety,
- Privacy and Data Governance,
- Transparency,
- Diversity, Non-discrimination, and Fairness,
- Societal and Environmental Well-being,
- Accountability.

The United States' Approach

In contrast to the EU's comprehensive regulatory framework, the US approach to AI ethics is more decentralized and less prescriptive. While there are various initiatives at federal and state levels, including guidelines from organizations like the National Institute of Standards and Technology (NIST), there is no singular federal law equivalent to the EU's AI Act. The US emphasizes voluntary guidelines that encourage companies to adopt ethical practices rather than enforce mandatory compliance. This includes principles such as fairness, accountability, transparency, and privacy protection in AI systems [5].

The US also relies heavily on industry-led initiatives to promote ethical AI development. Organizations such as the “Partnership on AI” bring together stakeholders from academia, civil society, and industry to discuss best practices and develop ethical frameworks collaboratively [6].

Ethics in Artificial Intelligence: The Main Problem

Regardless of the type of regulatory approach, the core of the ethical problem for artificial intelligence is summarized in the diagram in figure 1 [2,3]. As noticeable the central block: "evaluation and justification" constitutes the center of gravity of the cycle: Development - Use - Analysis - Re-design.

However, the main problem is precisely what are the criteria for evaluating and justifying the behavior of an AI artifact.

In the debate in search of an ethics for AI, common sense and the actual applicability of the concepts that arise from academic discussion are often lost, forgetting the implicit complexity of development, which is often linked to the actual availability of reliable data or the application of learning models that are still very far from the actual cognitive capacities of the human mind.

However, leaving aside the highly philosophical, theological and jurisprudential debates on what ethics is or is not, in a very pragmatic way we expect the behavior of the AI artifact to be "similar" to that of a human being. Going into detail, starting from the fact that an AI is a replica of a human cognitive behavior and that an AI’s cognitive behavior, usually, acts within a construct made up of: norms, rules, laws, best practices and acquired habits, it is reasonable to place a lower limit on the ethical approach for an AI by requiring that the generic AI artifact and/or system (which imitates a given human behavior) acts within the same construct (made up of: norms, rules, laws, best practices and acquired habits) within which a human being is expected to act.

For this reason, it is reasonable to assume that the evaluation and justification criteria mentioned in figure 1, as far as the behavior of the AI artifact is concerned, require, at least, compliance with the construct defined by norms, rules, laws, best practices and acquired habits that a human being is called to satisfy for the same work in which the analyzed AI acts.

It is quite obvious that such analysis is performed in the “analysis” block of figure 1 and what is proposed in this paper is to develop this block according to the scheme reported in figure 2 [7].

And it is also obvious that the first question to ask to a system that emulates human behavior, is whether the mechanisms that regulate it can be codified in some intelligible form. Such an operation is usually carried out with a reverse engineering process. In fact, the purpose of this procedure, in the context that this paper wants to emphasize, concerns precisely the study of the possibilities of translating the actions undertaken by that specific AI artifact into "rules", understandable to a person who lives in that context defined by norms, rules, laws, best practices and acquired habits.

In case reverse engineering is not possible, what this work suggests is to proceed with tests to which the system is subjected to understand its degree of conformity to the context in which a human operator would operate.

If reverse engineering is possible, then the “rules” by which the system proceeds in its analysis should be extracted.

This article describes an example of how to implement ethics tests in order to investigate the logic implicitly contained in a typical classifier, and highlight critical issues, which, once detected, can be corrected. All this in order to comply with the same set of laws, rules, regulations, procedures, best practices, etc., that a human being is required to comply with in the business segment in which the AI artifact under analysis operates.

In order to better understand the proposed approach, we will describe the methodology here suggested with respect classification systems.

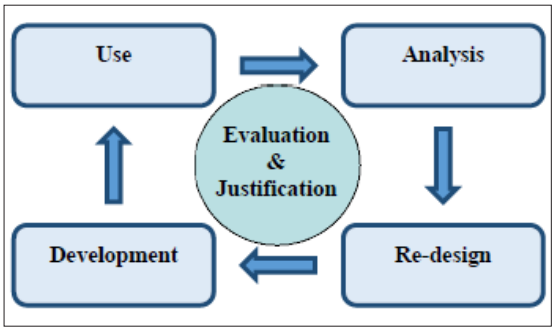

Figure 1

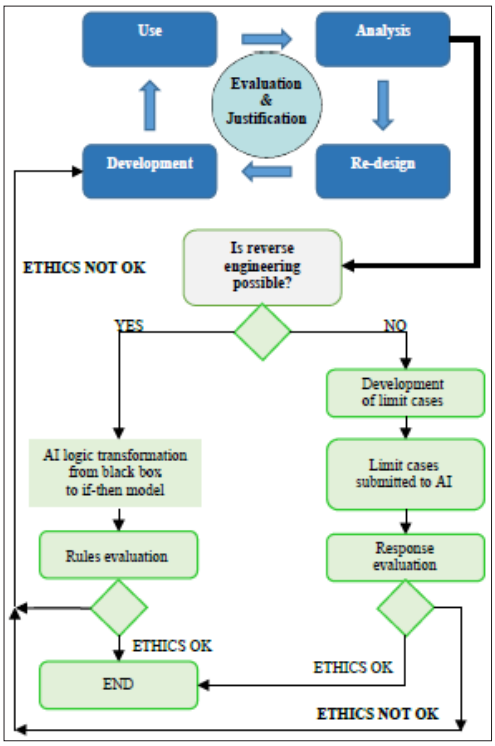

Figure 2

Experiment Description

Classification systems are undoubtedly among the most widely used models in practical applications of Artificial Intelligence. It is not surprising that such models are widely used, because "classification", as a logical paradigm, is one of the main learning methods of the biological mind. In fact, classification is indeed a fundamental paradigm in the learning processes of the human brain, reflecting how we organize and interpret our experiences. This cognitive ability is essential for making sense of the complex environments we navigate daily [8-11].

For this reason, we have considered a neural classifier and a classifier based on a decision tree model, both built to answer a specific problem, but which have the aim of being an example on which to implement the concepts expressed in figure 2.

A schematic representation of a trivial neural classifier is given in Figure 3. As well known, in this idealized classification system a series of features $P(k)_j$, with $j = 1, 2, \ldots, N$, of the generic item k belonging to some set of cases, are provided as input to the system and this, for each item, is asked to give a certain attribution to a specific class.

The approach described here is quite general, and can be applied to any classification problem. The specific classification problem considered here, as a demonstration of the proposed approach, concerns the problem of school dropout.

School dropout refers to the phenomenon where students discontinue their education before completing their degree, which can have significant implications for both individuals and society, as it reflects broader societal problems such as poverty, crime, and social inequality. High dropout rates are often indicative of deteriorating educational conditions and can perpetuate cycles of disadvantage within communities [12-14].

School dropout must be considered as a strong discriminatory phenomenon, and for this reason can be assumed as emblematic as AI solutions affected high level of risk. TABLE I reports a list of significant discriminatory factors influencing school dropout [15-17].

School dropout is a very well-studied item, in order to forecast, by the use of several techniques, including machine learning. In reviews such as [18 – 24] one can find extensive information regarding the different approaches that have been used to study this issue.

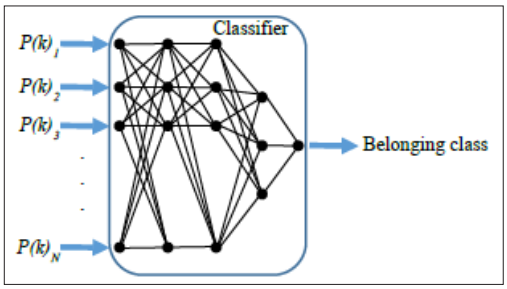

Figure 3

Table 1: Discriminatory Factors

Family Dynamics	Racial Bias
Early Marriage	Cultural Marginalization
Economic Hardship	Negative School Climate
Access to Resources	Mental Health Issues
Families with lower educational attainment	Teacher Expectations

Table 2: Features Used as Inputs for the Classifiers

Marital status	Displaced
Application mode	Educational special needs
Course	Debtor
Daytime or evening attendance	Tuition fees up to date
Previous qualification	Gender
Previous qualification (grade)	Scholarship holder
Nationality	Age at enrollment
Mother's qualification	International
Father's qualification	Unemployment rate
Mother's occupation	Inflation rate
Father's occupation	GDP
Admission grade	

Data and Models

For our work, the data was taken from the public site made available by UC Irvine Machine Learning Repository concerning higher education institution, and acquired from several disjoint databases, related to students enrolled in different undergraduate degrees [25]. The dataset consists of 4424 instances each of which is characterized by 36 features.

We used a simplified version of the original dataset, since 23 of the 36 features were considered as input values (see TABLE 2) and, as output, the target dropout consisting of only two categories: dropout or graduate, instead of the original three, which included, in addition to the two above, also the category: enrolled.

Therefore, the total instances, I_{tot}, used to fine-tune theclassifiers were 3630 instead of the original 4424.

The neural network used to train the model is a fully connected feedforward model: it takes input feature vectors with a dynamically determined size based on the dataset's dimensionality. The architecture consists of five fully connected layers with gradually decreasing numbers of neurons. The first layer has 256 neurons and is followed by a rectified linear unit (ReLU) activation function to introduce non-linearity [26-28]. The second layer reduces the neuron count to 128 and includes a ReLU activation function, followed by a dropout layer with a probability of 0.01 to mitigate overfitting by randomly deactivating some neurons during training. The third layer further reduces the size to 64 neurons with a ReLU activation, while the fourth layer reduces it to 32 neurons and incorporates both a ReLU activation and another dropout layer with the same dropout rate. The final fully connected layer outputs logits corresponding to the number of target classes, in this case just two, which are later processed by the loss function for classification. The ReLU activation in the hidden layers ensures the network can model non-linear relationships in the data, while the dropout layers add regularization to improve generalization. The network's weights are optimized using the Adam optimizer, which adapts the learning rate during training, and the Cross Entropy Loss is used as the loss function to calculate errors, combining the softmax operation and negative log-likelihood [29,30]. The model is designed to be trained in an iterative process over 10,000 epochs, updating the weights based on mini-batches of data.

The second model is built by using a classification pipeline with the Decision Tree Classifier to predict two classes. The dataset is first preprocessed to handle categorical and numerical variables. The target column, which represents the classification labels, is encoded using a Label Encoder to convert categorical values into numerical ones [31]. Categorical features are then one-hot encoded, creating binary indicators for each category, while avoiding multicollinearity by dropping the first category of each feature [32]. The resulting dataset is split into training and validation sets (split 60-40), ensuring a separate subset for model evaluation. Numerical features are standardized using a Standard Scaler to normalize the feature values, ensuring they are scaled to have zero mean and unit variance [33]. The scaled training data is used to train a Decision Tree Classifier with the Gini index criterion to measure node impurity.

The results obtained with these two different AI models are summarized in Table 3.

As it is possible to observe, the two classifiers produce very similar performance results. When tackling simple binary classification problems, such as distinguishing between "good or bad" or "black or white," both neural classifiers and decision tree-based classifiers can be effective [34]. However, when a neural classifier produces performance results comparable to those obtained with a decision tree classifier, it can have significant cognitive implications, particularly in the context of understanding model behavior and the nature of the data being analyzed.

The substantial equivalence of the two classification systems leads to the hypothesis that the explainability, allowed by the use of the classifier based on a decision tree, is transferable to the classification produced by the neural model. However, as reported later in this paper, this hypothesis is surprisingly denied when one tries to extend the classification properties of the two models to synthetically produced cases.

Table 3: Performances of the Classification Models

	Decision Tree	NN
Accuracy	0.8223	0.8588
Precision	0.8552	0.8666
Recall	0.8552	0.8588
F1 Score	0.8552	0.8541

Application of the Proposed Methodology

The first question that needs to be asked for the application of the introduced methodology, concerns how the classifiers, built on real data, behaves towards potentially existing cases, but not included in the training set, derivable from the possible combinations of the categories of the various features.

In fact, the cases contained in the public database, as numerous as they may be, are never fully descriptive of the possible observable reality. More precisely, the 3630 cases used to build the classifier constitute only a small part of the possible combinations deriving from the combination of the number of characteristics of the 23 input parameters. If Ntot represents this number, then N_{tot}

$$N_{tot} = \prod_j nc^j \cong 1.37 \cdot 10^{18} >> I_{tot} \qquad (1)$$

where ncj is the number of characteristics of the component j of $P(k)_j$, which represents the N-dimensional vector of the generic case k. Table 4 reports numbers of characteristics of each input parameter.

Table 4: Numbers of Characteristics for Each Feature

Feature	Nub. of cat.	Feature	Nub. of cat.
Marital status	6	Displaced	2
Application mode	18	Educational special needs	2
Course	17	Debtor	2
Daytime or evening attendance	2	Tuition fees up to date	2
Previous qualification	17	Gender	2
Previous qualification (grade)	5	Scholarship holder	2

Nationality	21	Age at enrollment	3
Mother's qualification	29	International	2
Father's qualification	34	Unemployment rate	5
Mother's occupation	32	Inflation rate	5
Father's occupation	46	GDP	5
Admission grade	3		

Thus, the basic problem is to understand whether the h cases, where $h \in [1,3630]$, used for the implementation of the classifiers and deriving from the public database, are sufficiently descriptive of the observed phenomenon.

There is no answer to this question because, due to problems of this nature, it will be difficult to find a database that contains the entire universe under observation. But if this observation is true, then it means that it is impossible *a priori* to establish whether the cases in the database, from which the classifier is built, are subject to bias or not.

At the same time, a limited number of well-selected cases could also be "sufficiently descriptive" of the phenomenon. But in this hypothesis, it is impossible to know *a priori* what it means, for a case under analysis, to "be sufficiently descriptive of the phenomenon", nor to know if that limited number of cases in the dataset actually are. Thus, also in this case, and again, it is impossible a priori to establish whether the cases in the database, from which the classifier is built, are subject to bias or not.

The ethics test for a classifier of this nature, therefore, can only be:
(*i*) the artificial production of missing cases in the database,
(*ii*) their evaluation by the classification procedure built with real cases.

Given the large number of theoretically possible cases, we produced several sets of randomly generated artificial cases, each containing several tens of thousands of synthetic cases. Then, we submitted the synthetic cases to the classifier (built with the real cases) to check for possible biases of the system.

Our focus, in this study, has mainly addressed on three variables considered particularly emblematic for the description of the methodology: *Nationality, Mother's qualification and Father's qualification.*

The reasons for choosing these variables are simple and very understandable. The phenomenon of school dropout is considered, at a social level, a sort of failure, both for the person who suffers (or chooses) the dropout, but also for the educational institution.

- In the context of school dropout, students from certain *nationalities* may experience discrimination or bias within the educational system, affecting their engagement and performance. This can further exacerbate dropout risks among these groups [15,35].

- The qualification variables of mothers and fathers can significantly introduce bias in the analysis of dropout rates among students, because parental education levels are closely linked to children's educational outcomes, including their likelihood of dropping out of school [36-38].

The way to evaluate a possible polarization of classifiers towards certain categories of the variables under analysis is the following:

1. 10 sets of randomly extracted synthetic data SD_i, with $i = 1, 2, \ldots, 10$, are created from a total number N_{tot}. The synthetic cases contained in each set, n_i, is equal to 20,000
2. The synthetic cases are subjected to the evaluation of the classifiers built with the I_{tot} instances actually contained in the starting database.
3. For the only cases d_i of the i-set that result with the output variable equal to *dropout*, the mean and the standard deviation calculated on the categories of the input features considered are performed. More precisely, we need to consider how d_i is distributed among the various categories of the feature $P(k)_j$. If each category c^f_j (where $f = 1, 2, \ldots, nc^j$) of feature j, has a number of dropout cases $D(c^f_j)$, it is clear that the sum over f of the cases c^f_j will be equal to d_i.

Therefore, the average value of the *dropouts* divided by the different categories of the feature j is,

$$\mu\left(D\left(c_j^f\right)\right)=\frac{1}{d_i}\sum_{f=1}^{nc^j}D\left(c_j^f\right) \qquad (2)$$

while the standard deviation is

$$\sigma\left(D\left(c_j^f\right)\right)=\sqrt{\frac{1}{d_i}\sum_{f=1}^{nc^j}\left[\mu\left(D\left(c_j^f\right)\right)-D\left(c_j^f\right)\right]^2} \qquad (3)$$

We define a characteristic f of feature j, c^f_j, as *polarized*, if the *dropout* number $D(c^f_j)$ for that feature holds the relation:

$$D\left(c_j^f\right)>\mu\left(D\left(c_j^f\right)\right)+\sigma\left(D\left(c_j^f\right)\right) \qquad (4)$$

Nationality Variable

All sets of synthetically generated cases, submitted to the evaluation of the classifier based on neural networks built with the real cases, showed a polarization of the Nationality variable towards some specific countries. *Nationality* variable includes the following categories, i.e. the following countries form which student are coming from: Portugal, Germany, Spain, Italy, Netherlands, UK, Lithuania, Angola, Cape Verde, Guinea, Mozambique, Sao Tome and Principe, Turkey, Brazil, Romania, Moldova, Mexico, Ukraine, Russia, Cuba, Colombia.

The neural classifier, in analysing the synthetic cases, reported a higher probability of school dropout, thus demonstrating to have a constant bias, for the following countries: Moldova, Mexico, Ukraine, Russia, Cuba, Colombia. In other words, the number of dropouts for these countries satisfy eq. (4). This evidence systematically appears for all sets of synthetic cases subjected to the evaluation of the neural classifier.

This result, reported in this way, would be extremely alarming. A forecast of higher school dropout rates for certain countries (in this case some from Eastern Europe and Latin America) can stigmatize entire populations, negatively influencing educational policies and opportunities for students. Assigning a higher probability of dropping out of school based on the country of origin can be considered an act of discrimination, especially when taking into account the socio- economic and cultural disparities that influence access to education [39-42].

Contrary to the result obtained with the neural classifier, the decision tree-based predictive classification system does not report any bias.

This result is surprising because it is different from what was obtained by the neural classifier, and does not allow us to understand the possible underlying logic that leads to the distortion. In any case, this is an unexpected result, given the substantial equivalence of the two classification models, as reported in Table 3.

However, the result of the decision tree classifier is not as surprising as the result of the neural classifier. In fact, by carefully observing the 3630 cases of the starting dataset, we observe that the vast majority of cases (about 97.6%) belong to a single country (Portugal), while all the others are distributed across the other countries.

All this leads to the consideration that the unequal distribution of the categories of a feature in the dataset that collects the real cases, inevitably leads to the creation of a distorted system.

Mother's and Father's Qualification

With the experience gained through the study of the Nationality category, we have implemented the same methodology described in steps 1 - 5 of the previous paragraph for the two variables: Mother's qualification and Father's qualification. In this case, however, only students from Portugal were considered, who constitute the majority of the instances I_{tot}.

We divided the characteristics of both the Mother's qualification and Father's qualification variables into three groups:

- Basic Education,
- Medium Education,
- Higher Education.

Tables 5, 6 and 7 show the details of this partition, mapped on the original classifications.

Table 5: Basic Education for Mother's and Father's Qualification

Father's Qualification	Mother's Qualification
7th Year of schooling	7th Year of schooling
7th Year (Old)	9th Year of Schooling - Not Completed
8th Year of schooling	9th Year of Schooling - Not Completed
9th Year of Schooling - Not Completed	10th Year of Schooling
10th Year of Schooling	11th Year of Schooling - Not Completed
11th Year of Schooling - Not Completed	Other - 11th Year of Schooling
Other - 11th Year of Schooling	General commerce course
12th Year of Schooling - Not Completed	Basic Education 1st cycle (4th/5th year) or Equiv.
2nd Year complementary high school course	Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.
General commerce course	Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.
Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.	Can't read or write
Complementary high school course	Can read without having a 4th year of schooling
Can't read or write	
Can read without having a 4th year of schooling	
Basic Education 1st cycle (4th/5th year) or equiv.	
Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.	
Unknown	

Table 6: Medium Education for Mother's and Father's Qualification

Father's Qualification	Mother's Qualification
Secondary Education - 12th Year of Schooling or Equivalent	Secondary Education - 12th Year of Schooling or Eq.
Frequency of Higher Education	Frequency of Higher Education
Technical-professional course	Technical-professional course
Complementary high school course - not concluded	2nd cycle of the general high school course
2nd cycle of the general high school course	
General course of administration and commerce	
Supplementary accounting and administration	

Table 7: Higher Education for Mother's and Father's Qualification

Father's Qualification	Mother's Qualification
Higher Education - Bachelor's Degree	Higher Education - Bachelor's Degree
Higher Education – Degree	Higher Education – Degree
Higher Education - Master's	Higher Education - Master's
Higher Education – Doctorate	Higher Education – Doctorate
Technological specialization course	12th Year of Schooling - Not Completed
Higher Education - degree (1st cycle)	7th Year (Old)
Specialized higher studies course	Unknown
Professional higher technical course	Technological specialization course
Higher Education - Master (2nd cycle)	Higher Education - degree (1st cycle)
Higher Education - Doctorate (3rd cycle)	Specialized higher studies course
	Professional higher technical course
	Higher Education - Master (2nd cycle)
	Higher Education - Doctorate (3rd cycle)

Also, for these characteristics, all sets of synthetically generated cases were submitted to the evaluation of the classifier based on

neural networks built with the real cases. What we have observed is a polarization, in the sense indicated by eq. (4), of both *Mother's qualification* and *Father's qualification* features, on categories related to a *Basic Education*.

Although there is no method to trace the implicit reasoning developed by a neural network, the result obtained is in line with social theories on the subject [43-45]. That is, it has been widely studied how low parental education is one of the causes of dropout [36-39].

In this sense, therefore, it is possible to observe a significant agreement between the field experience and the outputs obtained by the classifier.

However, the results obtained with the decision tree-based classifier proved, also in this case, to be surprisingly in disagreement with those of the neural classifier.

In fact, the *polarization* of the decision tree classifier, shown on the synthetically produced cases, was systematically found on the features related to *Higher Education*.

From literature, we can see that research indicates that parental involvement (such as supervising homework, attending school meetings, and discussing academic progress), is positively correlated with lower dropout rates. For instance, a study found that children whose parents actively engaged in their education had a significantly reduced risk of dropping out [46]. This involvement provides emotional support and academic guidance, which are crucial during challenging educational phases. Not only that. Parents with higher levels of education tend to have higher expectations for their children's academic success, that can motivate children to persist in their education [47]. Again, higher-educated parents have better access to resources families with greater financial resources can provide a more conducive learning environment and reduce stressors that might lead to school dropout, etc [48].

Such a result requires a deeper analysis of the reasons that lead the decision tree classifier to polarize on the categories identified as *Higher Education*. The possibility of navigating the reasoning implicitly contained in the decision tree, simplifies the understanding of the logic used in this particular case under analysis.

Discussion

As shown in Figure 2, in many cases it is not possible to understand, in a simple and linear way, the motivations that guide an AI artifact in its choices. It is well known how complex explainability is for a neural classifier and how this can be achieved, e.g., with various techniques that, however, can only achieve partial explainability [49-51]. Furthermore, the explainability offered by a symbolic-based AI artifacts, apparently identical in their classification performance to neural-based artifacts, is not necessarily sufficient and/or adequate and/or repeatable by extending the use of the classifier to cases not directly deriving from the learning dataset.

In this context, explainability offered by a typical classifier based on decision trees can lead to the discovery of unknown phenomena which necessarily require experiments on observed reality to be confirmed.

Starting from this scenario, the use of tests such as those described in this work, and called *ethics tests* for simplicity, prove to be a valid tool for investigating the internal logic of an artefact and making corrections to the artefact itself. In order to make it compliant with that set of laws, rules, regulations, procedures, best practices, etc., that a human being is required to respect in the same segment on which the analyzed AI acts.

The use of multiple technologies compared on the same problem, using well-constructed and balanced test cases, is an extremely useful technique, in order to obtain confirmations for phenomena whose internal dynamics are known, but also to highlight aspects not known *a priori*, which are worth investigating for a better and deeper understanding of reality.

References

1. Foulsham M, Hitchen B, Denley A (2019) GDPR - How To Achieve and Maintain Compliance. London: Routledge.
2. Council of the European Union, Responsible Artificial Intelligence: Ethics and Regulation. https://consilium-europa.libguides.com/c.php?g=690732&p=4948483.
3. Law HK (2024) The European Union's AI Act: What You Need to Know. https://www.hklaw.com/en/insights/publications/2024/03/the-european- unions-ai-act-what-you-need-to-know.
4. European Commission, Ethics Guidelines for Trustworthy AI. https://digital- strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
5. (2023) USAID AI Ethics Guide. https://www.usaid.gov/sites/default/files/2023- 12/_USAID%20AI%20Ethics%20Guide_1.pdf.
6. Partnership on AI, https://partnershiponai.org/.
7. Ballarin A, Vincenti M, Lo Sapio G, Fruscio G, Ballarin C (2024) A reasonable methodology for the realization of ethical artificial intelligence artifacts: From Turing test to ethics tests. AIP Conf Proc 3220: 1.
8. Friston K (2003) Learning and inference in the brain. Neural Networks16: 1325-1352.
9. Grossberg S (1981) Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control. Boston, MA D Reidel Publishing Company.
10. Anderson JR (1991) The adaptive nature of human categorization. Psychological Review 98: 409-429.
11. Shepard RN, Hovland CI, Jenkins HM (1961) Learning and memorization of classifications. Psychological Monographs: General and Applied 75: 13.
12. Wong Z (2023) The alarming epidemic of school dropouts: Causes and consequences. Journal of Educational Sciences Research 13: 25-40.
13. Winding TN, Andersen JH (2015) Socioeconomic differences in school dropout among young adults: The role of social relations. BMC Public Health 15: 23-91.
14. Fall AM, Roberts G (2012) High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. Journal of Adolescence 35: 1017-1028.
15. Kumar P, Patel SK, Debbarma S, Saggurti N (2023) Determinants of school dropouts among adolescents: Evidence from a longitudinal study in India. PLoS One 18: 3.
16. Ressa T, Andrews A (2022) High school dropout dilemma in America and the importance of reformation of education systems to empower all students. International Journal of Modern Education Studies 6: 15-30.
17. Sarette N (2022) The relationship between race and high school and college dropout rates. Perspectives 14: 45-60.

18. Romero C, Ventura S (2010) Educational data mining: A review of the state of the art. IEEE Trans Syst, Man Cybern, Part C Appl Rev 40: 601-618.
19. Mduma N, Kalegele K, Machuve D, A survey of machine learning approaches and techniques for student dropout prediction. Data Sci J 18: 1-10.
20. Shahiri AM, Husain W, Rashid NA (2015) A review on predicting student's performance using data mining techniques. Procedia Computer Science 72: 414-422.
21. Rastrollo-Guerrero JL, Gómez-Pulido J, Durán-Domínguez A (2020) Analyzing and predicting students' performance by means of machine learning: A review. Applied Sciences 10: 10-42.
22. Beaulac C, Rosenthal JS (2019) Predicting university students' academic success and major using random forests. Research in Higher Education 60: 1048-1064.
23. Hoffait AS, Schyns M (2017) Early detection of university students with potential difficulties. Decision Support Systems 101: 1-11.
24. Martins MV, Tolledo D, Machado J, Baptista LMT, Realinho V (2021) Early prediction of student's performance in higher education: A case study. in Advances in Intelligent Systems and Computing 1365: 16-25.
25. UCI Machine Learning Repository Student Performance Dataset. https://archive.ics.uci.edu/dataset/697.
26. Fukushima K (1975) Cognition: A self-organizing multilayered neural network. Biological Cybernetics 20: 121-136.
27. Banerjee C, Mukherjee T, Pasiliao E (2020) The Multi-phase ReLU activation function. in Proceedings of the 2020 ACM Southeast Conference.
28. Kulathunga N (2021) Effects of nonlinearity and network architecture on the performance of supervised neural networks. Algorithms 14: 51.
29. Shao Y (2024) An improved BGE-Adam optimization algorithm based on entropy weighting and adaptive gradient strategy. Symmetry 16: 623.
30. Li L, Doroslovački M, Loew MH (2020) Approximating the gradient of cross-entropy loss function. IEEE Access 8: 111626-111635.
31. Breskuvienė D, Dzemyda G (2023) Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. International Journal of Computers Communications & Control 18: 3.
32. DE Farrar, Glauber RR (1967) Multicollinearity in regression analysis: The problem revisited. The Review of Economics and Statistics 49: 92-107.
33. Zheng A, Casari A (2018) Feature Engineering for Machine Learning - Principles and Techniques for Data Scientist. USA O'Reilly.
34. Sya RF, Rahmadani R (2023) Performance comparison analysis of classifiers on binary classification dataset. Indonesian Journal of Data and Science 4: 77-84.
35. Pan C, Zhang Z (2024) Examining the algorithmic fairness in predicting high school dropouts. in Proceedings of the 17th International Conference on Educational Data Mining, Atlanta USA14-17.
36. Aina C (2013) Parental background and university dropout in Italy. Higher Education 65: 1-22.
37. Rumberger RW Lim SA (2008) Why students drop out of school: A review of 25 years of research. California Dropout Research Project 15: 2008.
38. Jeynes W H (2015) A meta-analysis - The relationship between father involvement and student academic achievement. Urban Education 50: 387-423.
39. Smith MH, Beaulieu LJ, Israel GD (1992) Effects of human capital and social capital on dropping out of high school in the South. J Res Rural Educ 8: 75-87.
40. Mezzanotte C (2022) The social and economic rationale of inclusive education: An overview of the outcomes in education for diverse groups of students. OECD Education Working Papers, no. 263, OECD Publishing, Paris.
41. Hadjar A, Scharf J (2018) The value of education among immigrants and non-immigrants and how this translates into educational aspirations: A comparison of four European countries. Journal of Ethnic and Migration Studies 45: 811-831.
42. Martin M, Stulgaitis M (2022) Refugees' access to higher education in their host countries: Overcoming the 'super-disadvantage. UN High Commissioner for Refugees (UNHCR) & UNESCO.
43. Aravantinos V, Diehl F (2019) Traceability of deep neural networks. arXiv preprint arXiv:1812: 06744.
44. Tambon F, Laberge G, An A (2022) How to certify machine learning based safety-critical systems? A systematic literature review. Automated Software Engineering 29: 38.
45. Traylor A, Feiman R, Pavlick E (2023) Can neural networks learn implicit logic from physical reasoning?. in Proceedings of the ICLR Kigali Rwanda.
46. Paul R, Rashmi R, Srivastava S (2021) Does lack of parental involvement affect school dropout among Indian adolescents?. Evidence from a panel study. PLoS ONE 16: e0251520.
47. Ross T (2016) The differential effects of parental involvement on high school completion and postsecondary attendance. Education Policy Analysis Archives 24.
48. Karhina K, Bøe T, Hysing M, Askeland KG, Nilsen S A (2024) Parental separation and school dropout in adolescence. Scandinavian Journal of Public Health 52: 632-639.
49. Ortigossa ES, Gonçalves ES, Nonato LG (2024) EXplainable Artificial Intelligence (XAI) – From theory to methods and applications. IEEE Access 12: 12345-12360.
50. Seo B, Li J (2024) Explainable machine learning by SEE-Net: Closing the gap between interpretable models and DNNs. Scientific Reports 14: 77507.
51. Renftle M, Trittenbach H, Poznic M, Heil R (2024) What do algorithms explain? The issue of the goals and capabilities of Explainable Artificial Intelligence (XAI). Humanities and Social Sciences Communications 11: 32-77.