Artificial Intelligence and Law
https://doi.org/10.1007/s10506-024-09395-w

ORIGINAL RESEARCH

Exploring explainable AI in the tax domain

Łukasz Górski[1,2] · Błażej Kuźniacki[3,4,5] · Marco Almada[6] · Kamil Tyliński[7,15] · Madalena Calvo[8] · Pablo Matias Asnaghi[9] · Luciano Almada[9] · Hilario Iñiguez[10] · Fernando Rubianes[11] · Octavio Pera[11] · Juan Ignacio Nigrelli[12,13,14]

Accepted: 15 February 2024

Abstract
This paper analyses whether current explainable AI (XAI) techniques can help to address taxpayer concerns about the use of AI in taxation. As tax authorities around the world increase their use of AI-based techniques, taxpayers are increasingly at a loss about whether and how the ensuing decisions follow the procedures required by law and respect their substantive rights. The use of XAI has been proposed as a response to this issue, but it is still an open question whether current XAI techniques are enough to meet existing legal requirements. The paper approaches this question in the context of a case study: a prototype tax fraud detector trained on an anonymized dataset of real-world cases handled by the Buenos Aires (Argentina) tax authority. The decisions produced by this detector are explained through the use of various classification methods, and the outputs of these explanation models are evaluated on their explanatory power and on their compliance with the legal obligation that tax authorities provide the rationale behind their decision-making. We conclude the paper by suggesting technical and legal approaches for designing explanation mechanisms that meet the needs of legal explanation in the tax domain.

Keywords Artificial intelligence · Tax fraud · Explanation methods · Legal requirements · Duty to give reasons

1 Introduction

Over the past few years, tax authorities have started to use machine learning techniques to process the large volumes of data they have about taxpayers. In doing so, they hope to leverage that data for a multitude of purposes, such as detecting fraudulent behaviour by taxpayers, integrating new sources of information, or simply doing tasks that would otherwise be neglected due to personnel shortages (Collosa 2021).

Extended author information available on the last page of the article

Published online: 07 May 2024

The use of AI approaches can potentially allow tax authorities to increase their performance in gathering information and enforcing the law.

The development and use of tax AI in the public sector is, however, subject to legal constraints. In most countries, tax administration (just like other parts of public administration) is subject to the principle of legality: it can only act when authorized by law and in the form authorized by law (Hadwick 2022). Among these formal principles, most countries adopt some form of the duty to give reasons, which obliges administrative decision-makers to specify the facts and the laws that guide their decision-making (Fink and Finck 2022). This means that, regardless of whether an AI system is making a decision or providing inputs for a human decision-maker, the role of the system in decision-making must be explained (Bibal et al. 2021).

In this paper, we examine the suitability of current XAI techniques for providing explanations of decisions about tax. To do so, we make use of a prototype system for fraud detection, developed in collaboration with the Buenos Aires tax authority. This system, as further detailed in Sect. 3 below, is not yet at state-of-the-art performance, but it is nonetheless illustrative of the goals and approaches that power real-world applications of AI in tax. As such, it provides a realistic baseline for evaluating potential explanation models and their assessment vis-à -vis relevant legal background.

This paper contributes by:

- Giving an expert-based account of the feasibility of current XAI methods in the context of tax law, based on the dataset derived from real-life experiences.
- Analysing the interplay between legal requirements, expectations of legal experts, and technical possibilities.
- Creating a background work for future guidance on how to secure taxpayers' constitutional rights and increasing tax moral in the areas in which tax authorities rely on AI.

To evaluate said explanations, the paper proceeds as follows. Section 2 provides an overview of current scholarship on XAI and the specific requirements for explanations in the tax domain. Section 3.1 presents the dataset. Section 3.2 then discusses how tax fraud detector was implemented, i.e. what machine learning models were chosen. Section 3.3 discusses various XAI methods used and shows the result of their use. The general schema of the implemented system, as well as the how this structure connects with the structure of this manuscript is presented in Fig. 1. Section 4 moves to a qualitative assessment of the explanations produced in the previous section, contrasting them to the legal requirements that any AI-supported decision must meet. This comparison leaves somewhat to desire, so we conclude the paper by proposing technical and legal paths forward toward proper explanations in the tax domain.

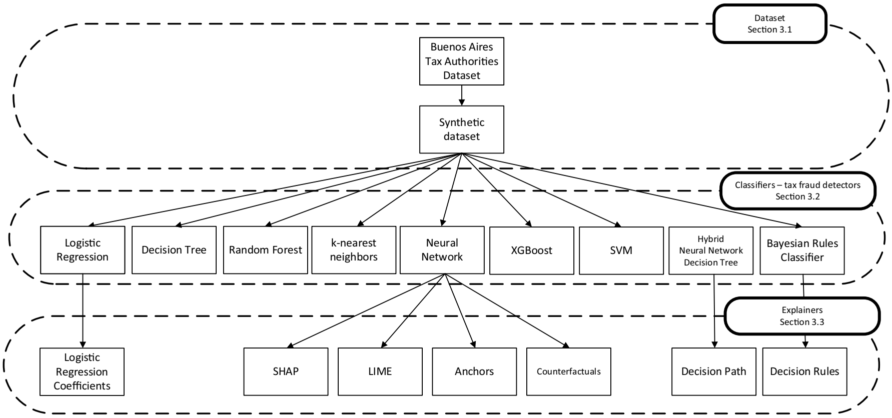


Fig. 1 The structure of implemented system and of the manuscript

2 Related work

In this paper, we will use the term explainable artificial intelligence (XAI) as the development of techniques that make the functioning of an AI system understandable for a given audience (Arrieta et al. 2020). XAI methods aim to show how the AI system's input affects its output by revealing the link between the data ingested by the system and the decision it makes. Accordingly, XAI methods may provide the decision-makers with an account of how a given AI-based system works, thereby allowing to a technically guided explanation to be transformed into justification by the public authority.

2.1 XAI curriculum

The selection of methods that have been included is based on their popularity in the XAI community, as evidenced by availability of manuals and research papers. The popularity of such explanation methods like SHAP, LIME, anchors and counterfactuals makes it a must to test them in the area of tax law. All those explanation generation methods work by examining model's input and output and do not rely on its inner working. Thus, they are in principle applicable to any category of machine learning models.

As for the usability analyses of the methods presented herein, they were the subject of scrutiny in other works. When comparing the results achieved using SHAP and LIME it was found they do not offer advantage over each other and they are assessed similarly by end users (Górski and Ramakrishna 2021). In (Slack et al. 2020) it has been proven that it is possible to prepare a malicious classifier that hides real rationale for a decision when a perturbation-based explainer (i.e. SHAP/LIME) is used. Better sampling techniques are in the works to mitigate this adversarial attack (Vreš and Robnik-Šikonja 2022), which seems a necessity if the explanation models are to evoke users' trust in the AI-based systems. In the works like

(Bench-Capon 1993) and (Schweighofer 2022) it has been shown that there can be disconnection between what the black-box classifier has internalized and the legal rationale for a decision. This can to some extent be identified using SHAP (Schweighofer 2022), but the identification is not straightforward. It does not mean that there is no trust in automation of any part of judicial decision-making process, for instance data gathering for the purpose of making legal decisions has proven to be trustworthy (Barysė and Sarel 2023).

This disconnection can in principle be overcome by including domain knowledge in explainer. Other authors suggested the possibility of building a hybrid system, that includes domain knowledge for better results presentation that is better catered to needs of prospective users in a legally-oriented task (Branting et al. 2021). The authors used subsymbolic methods (embeddings, a technique that couples a text fragment with its representation in high dimensional vector space) coupled with manual annotation for supervised learning. This allowed to include domain knowledge when system prediction were explained and the authors have found this solution to be of higher value than the one that used heatmap, i.e. one that highlighted the text the neural model deemed most important, without reference to predefined legal concepts.

One of the challenges of currently available explanation methods is that they are implicitly based on the background knowledge (Combi et al. 2022). In principle, such knowledge is not in the possession of humans who are interested in the predictions of AI-based system. For example, a heatmap can show which parts of scanned brain contribute to a diagnosis, but this is of little use for a patient with limited or no medical knowledge (Robbins 2019). By analogy, the same applies to taxpayers targeted by AI systems used by tax authorities which are currently explainable only to a limited extent, as such explanations are typically understandable by AI domain experts and require more information to be used by lawyers or taxpayers. Other authors have already noticed that there are many stakeholders (groups of people interested in explanations), but most XAI scholarship seemed to cater to system developers (Langer et al. 2021). For example, if the explanations were to be presented to a judge in a court case, they would have to mediated through expert's testimony (Kuźniacki et al. 2022).

Even with support of tax experts, XAI methods in the field of tax law may not be understandable by taxpayers due to the lack of or insufficient collaboration and mutual understanding between AI experts developing and deploying AI systems used by tax authorities and tax experts. Whilst, for example, methods such as SHAP feature importance plots are able to show the variables that impact the neural network's prediction to the greatest extent, showing how they interact with other features and background knowledge to arrive at the result is whole new endeavor. In other words, currently available explanations are important steppingstones in the research that will have to provide the wider context in which the decision was made: "beliefs and motivations; hypotheses of other (human, animal or AI) agents' intentions; interpretation of external cultural expectations; or, processes used to generate its own explanation" (Dazeley et al. 2021). Current explainability methods lack in terms of causality: the presentation of model's relevant modules and input data, which does not necessarily end in user's satisfaction and understanding in the context of a given task (Holzinger et al. 2019). With the use of generative AI the new possibilities of generating explanations are presented and they yield promising results that have capacity to make XAI easy to understand by laypersons (Yu et al. 2022).

2.2 Legal requirements for XAI in the tax domain

For the purposes of this paper, it is important to distinguish between three concepts: interpretability, explainability, and justification. The first two are occasionally used as synonyms, but we follow some scholars (Arrieta et al. 2020) in ascribing different meanings to them. *Interpretability* refers to an inherent quality of a machine learning model that allows a human (usually an expert (Kolkman 2022) to make sense of it, whereas *explainability* refers to the possibility of designing an interface that allows a human to make sense of the model. In both cases, what one wants to understand is the model and the outputs it produces (Creel 2020).

Justification, in contrast, is less concerned with understanding and more with the legal value of a decision. Under the principle of legality (Craig 2020), administrative decisions are only valid to the extent that they are grounded on legal authority. From a legal perspective, it follows that a decision by a tax authority—which is a form of administrative body representing the executive state's (fiscal) power—must be justified with reference to the existing laws, regulations, and other legal instruments applicable to a decision. In fact, these authorities are obliged, to a large extent, to present these reasons to the persons affected by the decision and sometimes to the public (Schauer 1994; Bardutzky 2022). This reason-giving duty continues to apply whenever an AI system becomes part of an administrative decision-making procedure (Bibal et al. 2021).

Since explanation and justification are different things, XAI techniques, by definition, are not sufficient to produce justifications of the kind expected by the law. Nonetheless, it has been argued that explanations of AI-based decisions are necessary for evaluating any justification of such a decision. Since the information provided by AI systems is an important factor in decision-making processes (Demková 2021), any assessments of how the law is applied to a given context must engage with how the system processed data and how that data was used. Consequently, some authors (Fink and Finck 2022) have argued that explanations must be a part of the reason-giving whenever an AI system is used in administrative contexts. At the same time, others (Ferrario and Loi 2022; Mehdiyev et al. 2021; Zerilli, Bhatt, and Weller 2022) have argued that explanations can contribute to the acceptance of AI-based tax decisions by taxpayers. However, some have raised warnings about how the use of XAI may create undue constraints to legal decision-making (Esposito 2022), or launder unacceptable decisions through the manipulation of explanations (Bordt et al. 2022) or, more generally by validating institutional practices of secrecy (Busuioc et al. 2023). XAI, therefore, is not a panacea for algorithmic transparency in the government, or an automation of justification, but a necessary element of the overall governance of public sector AI.

To fulfill this role, XAI solutions in the public sector must be tailored to meet the informational needs imposed by law. In the tax domain, such tailoring means that an explanation of a tax decision must provide the information that is needed to evaluate whether that decision complies with the applicable laws and the rights of taxpayers (Kuźniacki et al. 2022). In particular, XAI can play an important role in preventing arbitrary and discriminatory decisions infringing right to privacy without a proper judicial oversight, such as those already detected in some jurisdictions (Amnesty International 2021). Notably, AI systems without oversight might: (i) biased decisions, (ii) be used for purposes beyond the legitimate scope that motivated their introduction, or (iii) be used in ways that deprive taxpayers of their right to contest potentially wrongful decisions (Kuźniacki et al. 2022). All these risks are compounded by the various forms of opacity that surround AI systems, which may preclude taxpayers from learning about the tax decision-making procedure or even about the existence of a decision based on an AI system in the first place. In order for the legal system to reach the stage at which automated decision making could be implemented there is a need to change the infrastructure and information gathering process so it can be served by the machines and provided with sufficient explanations (Reiling 2020). Additionally, it has been noted that to change the existing procedures in the way legal decision-making is organized there is a need to involved actors of the system such as judges, tax administration to fully reorganize the procedures and create ecosystem that is capable for the use of new technologies (Sourdin 2022).

The decisions of tax authorities must comply with the principle of formal motivation. To comply with this principle, all factual and legal grounds on which the decision is based should be mentioned and explained by the tax authorities, unless a tax and / or trade secrecy requires tax authorities to not reveal certain (especially factual) information concerning their decisions[1] (Kuźniacki et al. 2022). The justification for such decisions must be clear and precise and reflect the real motives behind the decision.[2] If human decision-makers have no access to the explanations of systems they rely on, they might end up simply following any recommendations from those systems (Wagner 2019), or even adopting a selective form of compliance, in which they disregard any solutions that "seem off" and follow what "seems plausible" (Alon-Barkat and Busuioc 2023). Whenever that happens, these decision-makers cannot offer any reason beyond "computer says no", a rationale that is incompatible with the principle of legality (Oswald 2018), as held by courts in various jurisdictions (Fink and Finck 2022; Kuźniacki et al. 2022; Zandstra and Brouwer 2022). Hence, the use of AI systems in the tax administration is unlikely to be

[1] Tax and trade secrecy may prevent an explanation of tax decisions based or supported by AI systems. This may be called a legislative opacity or an opacity by legal (by contrast or in addition to technological) design. However, this article focuses only on technological aspects of explainability of AI in tax related cases. More for tax and trade secrecy and explainability of AI in tax domain is available in (Kuźniacki et al. 2022) at sec. 2.2, as well as (Kuźniacki and Hadwick 2023b; 2023a).

[2] Art. 41 of the Charter of Fundamental Rights of the European Union (the EU Charter), Official Journal of the European Union C 326/391, 26 October 2012. For national law, see, for example, the Belgian law of 29 July 1991 on the formal motivation of administrative decisions.

lawful without some support from XAI techniques. The specific configurations of these techniques, however, will depend on the particular requirements of the jurisdiction in which the system is used.

3 Empirical studies

3.1 Dataset

Any evaluation of XAI approaches in the tax domain must consider the context in which AI is used. However, some areas of the government—notably law enforcement (Curtin 2020) and tax authorities (Hadwick 2022)—use the law to prevent, or at least restrict, disclosure of information about the algorithms they use and the data that feeds those algorithms. Accordingly, there is little published technical work on the application of XAI techniques in the context of tax authorities (Kuźniacki et al. 2022; Mehdiyev et al. 2021; Kuźniacki 2022).

There is a number of legal datasets targeted to machine learning projects, but they differ in creation methodology. The COMPAS dataset discloses the decision that was made in practice and the criteria for that decision. The criteria are extracted from relevant authorities, using public records and dataset's authors merged it with prison and jail information (Barenstein 2019). This dataset is, however, in the domain of criminal law and we are unaware of alternatives that are publicly accessible and targeted to the tax law domain. There are other datasets (Savelka and Ashley 2021; Walker et al. 2019; Chang et al. 2020) that focus on an argumentative structure on interpretative exercises that are executed during the decision-making process and not on the decision itself. In addition to aforementioned datasets, legal sources themselves form a vast repository of data, including statutes, judgments, decision, or writs, often available for scraping (Rissland et al. 2003). Such scraped data is often not immediately useful and further processing is nevertheless often needed. The data needs to be manually extracted/annotated or the specialized tools for preprocessing have to be developed (Górski et al. 2020).

In this paper, we make use of a dataset prepared by the Buenos Aires tax authorities for the purposes of fraud detection. It stands out from aforementioned datasets by focusing on tax law, and the facts of the case that are used by tax authorities to assess the risk of fraudulent activity. This dataset is not yet available to the public (and the tax authorities have received our feedback regarding its future development), but the following lines describe its general characteristics, starting with the relevant legal background.

Argentina has three jurisdictions with taxing powers: national, provincial and municipal. At the provincial level, all the 23 Argentine provinces as well as the Autonomous City of Buenos Aires (CABA) impose a Gross Turnover Tax (GTT) on the regular activity of commerce, industry, services or any other activity carried out within their jurisdictions. This GTT is levied on gross revenues resulting from the regular and onerous exercise of commerce, industry, profession, business, services or any other onerous activity conducted on a regular basis within the respective provincial jurisdiction.

In order to detect the possible existence of fraud or tax evasion, understood as "elimination or reduction of tax produced within a country, by those who are legally bound to pay it and who achieve that result by means of fraudulent or omissive conducts that violate legal provisions" (Villegas Héctor 2001), that is, the unfulfillment of the tax obligation through illegitimate means, making a difference from legal ways of avoiding said obligation (elusion) or directly choosing not to carry out the event giving rise to the obligation in itself (economy of choice); the presumptive tool can legitimately be used.

In that regard, the legislator establishes that "may be used as indicators, among others: the capital invested in the exploitation, fluctuations in assets, volume of transactions or sales from previous tax periods, the amount of purchases, the existence of merchandise, the existence of raw materials, dividends, general expenses, wages and salaries, the rent of property used for the business, industry or exploitation and of the house-room, the taxpayer's standard of living, the normal performance of businesses, exploitations or similar enterprises from the same branch; and any other elements of judgement that are in the possession of the Administration or which must be provided to it by the taxpayer or liable person, chambers of commerce or industry, banks, trade union associations, public or private entities, collection agents or any other person who possesses useful information in this respect related to the taxpayer which is related with the verification and determination of the taxable events" (arts. 247 s paragraph, Fiscal Code). The formula laid by the legislator is broad, including, but not limited to, the indicators aforementioned. In addition, it establishes in article 248 different systems for the determination over presumptive basis.

Based on that broad legal definition, a dataset that reflects the practical side of tax authorities' work was prepared by the Buenos Aires Tax Authorities. The dataset consists of nine features denoting the existence of the facts that they use to assess the probability of the tax fraud and the status of taxpayer (cf. Table 1). This dataset consists of binary features, in a form of a table that denotes whether a given fact took place in a given case and whether a given case it was finally assessed that a fraud was committed.

There are some cases with missing data, i.e. it is not declared whether a certain fact occurred in a given case. Upon consultation with Buenos Aires tax authority we have found out that denotes unavailability of a given data for a particular case. We have found that lack of data is also a valuable information and encoded the lack of data using an arbitrarily chosen number (2).

Figure 2 presents the number of times a given value was assigned to a given feature in the dataset. 1 denotes existence of a fact in a given case, 0 – that such fact did not occur, NaN denotes missing data. The dataset consists of 6465 cases, of which 3290 rows have at least one NaN value. As this is a real-life dataset, it exhibits data imbalance. That is, there is a significant disproportionality between the number of cases that are fraudulent and not (612 instances whole). Such imbalance is inevitable due to the underlying phenomenon described, i.e. tax frauds always–by nature–constitutes a tiny fraction of behaviour of all taxpayers, here within the group of taxpayers subject to the tax law as described above (Zareapoor and Shamsolmoali 2015).

Table 1 Description of risk features, as delivered by Buenos Aires Tax Authority

Feature name	Description
Fraud GTT (target variable, i.e. the one that denotes whether the fraud was committed)	difference between Net Assessed Price and GTT's tax base is greater than 10% for the whole quarter
Pesos_Sales	The taxpayer has underdeclared sales. There is a discrepancy between purchases and sales
Labour_Cost_Sales	The taxpayer has underdeclared sales. There is a discrepancy between his purchases and the labor cost and the value of the sales
Labour_Cost_Net_Sales	The taxpayer has underdeclared sales. There is a discrepancy between labor cost and net sales
F931	If there is any F931 filing in the first quarter that is, if it has employees in its charge, and the relationship with the activity it carries out
Underreported_Work_Hours	Failure to correctly declare employees work hours
Incorrect_Rate	Declare an incorrect proportional share
Excess_Deduction	The taxpayer takes more withholdings than the ones declared by the Collection Agent
Collecting_Agent	Is a Collection Agent
Higher_Rate_Tax_Payer	Is a High-Income Taxpayer

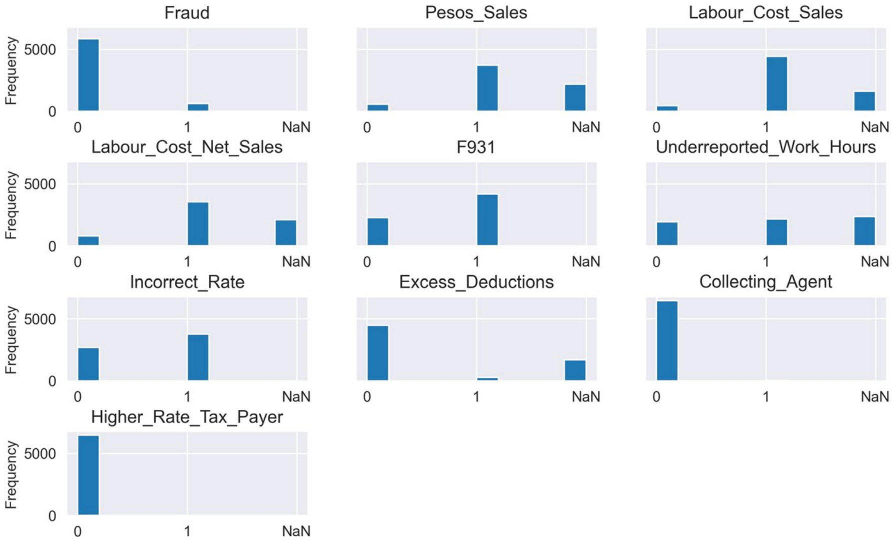


Fig. 2 Histogram of features' values distribution. This is a part of data analysis performed before the development of tax fraud detector and generation of explanations

Moreover, it has been discovered by us that the majority of rows are either repeated completely or differ only whether a fraud was committed in a given case or not. In other words, there are dataset rows that exhibit the same feature values and differ only in the end result of fraud being committed. In machine learning terms, the dataset is noisy. This proves that more features that differentiate various cases should be introduced in the future, and this need has been relayed to Buenos Aires tax authority. This lack of predictive potential is a known phenomenon in the case of datasets that consist of categorical features. Also, the features Collecting_Agent and Higher_Rate_Tax_Payer are constant in the dataset (they are always 0), thus they offer no predictive power and they are not used in modelling.

We have generated a synthetic dataset based on the one described in the preceding paragraphs. This was done using the normalizing flow algorithm (Durkan et al. 2019). Synthetic dataset generation aims to create completely new datasets that mimic the distribution of samples in the original one. This algorithm-based generation allowed to create a dataset that contained of 1300 samples, 999 non-fraudulent and 301 fraudulent. This synthetic dataset allows us to prepare machine learning models without the need of performing a human subject research study, whilst still allowing to assess the algorithms' performance. The usage of such dataset further mitigates any worries regarding the privacy and data safety of the taxpayers mentioned in the dataset.

3.2 Classifiers

The dataset described above was used as the starting point for the implementation of several well-known classification models. While these implementations were meant as a proof of concept for a potential automated detector, they are not the focal point of this study. However, it is still important to present the classifiers implemented. Such system could in principle be used by tax authorities to either identify the potentially fraudulent behavior for further scrutiny, or to provide assessment as for the fraudulent nature of conduct. The nature of dataset and features identified by tax authorities effects the overall system to be more suited for the former role, for the support of tax administration as it exercises its lawful discretion, which is also suggested by the perspective of efficiency and compatibility with fundamental taxpayer rights. The discretion, itself, raises various questions in the context of AI, but the answers to them lie beyond the scope of our paper (cf. (De Cooman 2023)).

The classifiers implemented and presented in this chapter are as follows: decision tree, random forest, logistic regression, simple neural network (with three fully connected hidden layers, sized 20, 15, 10 respectively), hybrid neural network-decision tree model (with tree created from the model using the ANN-DT algorithm (Schmitz et al. 1999)), k-nearest neighbors (KNN), Bayesian rule lists, and XGBoost (XGB). One-hot encoding was used. Hyperparameter space was also explored and the best model was chosen.

This exploration used the traditional method of testing hyperparameters of varying order of magnitude using the grid search method. This implementation used fivefold cross-validation to perform the search (Agrawal and Agrawal 2021). This has led to setting the following parameters: for logistic regression, C = 10, solver = newton_cg, tol = 1e-5, penalty = l1; for the decision tree, max_depth = 7, criterion = gini, max_features = 10, min_samples_leaf = 1, min_samples_split = 5,

Table 2 Performance of various implementations of tax fraud detectors. Those detectors were created before explanations of their decisions were generated

Model	Accuracy/95% confidence intervals	Confusion matrix	ROC AUC/95% confidence intervals
Logistic Regression	0.59 0.54–0.64	$\begin{bmatrix}80 & 70\\ 11 & 34\end{bmatrix}$	0.6 0.57–0.65
Decision Tree	0.6 0.53–0.66	$\begin{bmatrix}85 & 65\\ 12 & 33\end{bmatrix}$	0.63 0.56–0.69
Hybrid Neural Network + Decision Tree	0.65 0.63–0.67	$\begin{bmatrix}103 & 47\\ 21 & 24\end{bmatrix}$	0.61 0.6–0.62
KNN	0.6 0.54–0.66	$\begin{bmatrix}123 & 27\\ 31 & 14\end{bmatrix}$	0.6 0.54–0.66
Random Forest	0.64 0.6–0.69	$\begin{bmatrix}104 & 46\\ 18 & 27\end{bmatrix}$	0.64 0.59–0.69
XGB	0.66 0.62–0.71	$\begin{bmatrix}99 & 51\\ 15 & 30\end{bmatrix}$	0.66 0.61–0.72
SVC	0.63 0.57–0.68	$\begin{bmatrix}95 & 55\\ 17 & 28\end{bmatrix}$	0.63 0.57–0.69
Neural Network	0.66 0.58–0.74	$\begin{bmatrix}105 & 45\\ 21 & 24\end{bmatrix}$	0.62 0.5–0.69
Bayesian Rule List	0.31 0.24–0.4	$\begin{bmatrix}18 & 132\\ 2 & 43\end{bmatrix}$	0.54 0.49–0.58

min_weight_fraction_leaf = 0, splitter = random; for KNN, algorithm = ball_tree, leaf_size = 100, n_neighbors = 10, p = 1; for Random Forest, max_depth = 7, max_features = 1, min_impurity_decrease = 0, min_samples_leaf = 1, min_samples_split = 5, min_weight_fraction_leaf = 0, n_estimators = 10; for XGB, eta = 1, gamma = 0.1, max_depth = 6; for SVC, C = 10, kernel = poly. Neural network was trained for 250 epochs (using early stopping), with batch size = 16. This was implemented using the following Python libraries: TensorFlow 2.10.0, scikit-learn 1.1.2, XGBoost 1.6.2, imodels 1.3.18.

The classifiers' performance has been evaluated using a number of metrics. Herein (Table 2), we show the results in terms of accuracy and ROC AUC. Those values are presented as means obtained using bootstrapping (with n = 1000), alongside the 95% confidence intervals (Adibi 2004). For hybrid neural network + decision tree solution, the network that achieved 0.66 accuracy was chosen as the base to generate a tree from during the bootstrapping. Additionaly, confusion matrices are presented for those bootstrapped models which accuracy score was closest to the mean one.

The results are presented in Table 2, in terms of test set accounting of 20% of all the instances. In general, the results are on par, with the exception of Bayesian

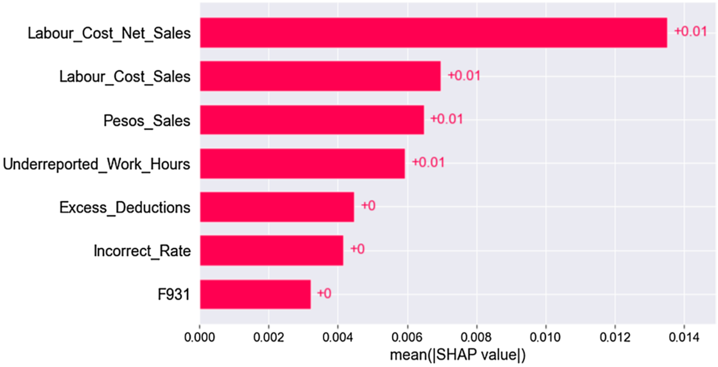
Rule List. In this research, we did not strive to maximize the performance metrics for the models. Rather, we have treated them as a starting point for explainability analysis for tax law applications. In this respect, it can already be noted that the neural network, and the hybrid neural network+decision tree solution built from the same network achieved very similar performance metrics. The latter is also a readily interpretable solution (cf. the next section). Any neural network is susceptible to be represented by a decision tree (Aytekin 2022) and the end user, a non-technical person, might be in better position to assess the results stemming from decision tree-based solution. Moreover, if an interpretable model can be trained at a similar level of performance than what was achieved by the black box model, one might have little reason to use a black box model in the first place. Some legal scholars have, in fact, argued that the use of interpretable models might be legally required in some circumstances (Babic and Cohen 2023). Such a requirement, however, is not an absolute demand, as it hinges on the specific role that an AI system plays in a given context and on the rules that apply within a given jurisdiction. Given that, as mentioned above, we focus on an application that carries out an auxiliary task that lies within the administrator's range of discretion, we see no *prima facie* duty to adopt an interpretable model instead of adding an explanation to a black box model. The choice between interpretable models and an XAI approach hinges, therefore, on which approach leads to more suitable trade-offs between transparency and performance (Kuźniacki et al. 2022). If, on the one hand, there is no reason to believe black box models always perform better (Semenova et al. 2022), some empirical research has pointed out that interpretable models are not always more legible to humans in the decision-making loop, either (Bell et al. 2022).

3.3 Explaining the classifiers

We aim to explore explanations of a model that was trained in the previous section, for the purpose of using it in the reality of tax administration work. There are two categories of explications we present here, in line with Fig. 1.

The first one applies to tax fraud detectors that in principle are understandable out-of-the-box. In other words, their inner workings (how they reach a prediction) are accessible to any sufficiently-trained person versed in machine learning. Such models, that we implemented, include decision trees, Bayesian rule lists, or logistic regression. In contrast, black-box models, like neural networks, need to have additional methods employed to give insight into how they reach a decision. Herein, we use the following model-agnostic methods to explain the black-box classifier based on neural network: LIME, Anchors, SHAP (for local as well as global explanations), counterfactuals. Out of implemented black-box models, the neural network was chosen as a basis for explanation generation, as it is typical example of a black-box model that suits well needs and nature of work of tax authorities, i.e. they rely on having access to huge volumes of tax-sensitive data that are very favourable to the use of AI technologies based on neural networks.

Table 3 SHAP global feature importance

Explanation	
Natural language based explication	The most important feature for the predictions generated by the neural network is the discrepancy between labor cost and net sales, with the feature pertaining to purchases, labor cost and the value of the sales following. The other important features include pesos sales and underreported work hours

In the following section, the output of the methods is presented as images and tables. Those have been supplemented with additional descriptions in natural language created by technical authors of this paper. The need for additional explication arose, as, in general, authors with legal backgrounds have found the outputs of XAI methods as well as inner-workings of white-box models to be hard to approach and incomprehensible. Thus, further natural language descriptions were needed to make sense out of raw outputs of XAI methods. Their content was thus created in dialogue with this paper's legal experts. This gives a complete account on the extent of explication that was needed for the explanations as well as interpretable models to be approachable for non-technical experts. It is possible to write a postprocessing layer that would generate similar descriptions automatically, though it is a separate research and development endeavor, out of the scope of this paper. Various studies have also already scrutinized the possibilities of generating natural-language based explanations (Cambria et al. 2023). For simplicity, where applicable, ordinal encoding is used, even though underlying classifier uses one-hot encoding.

Global feature importance has been calculated using SHAP (Table 3). SHAP itself is a method that uses game-theoretical concept of Shapley values, as well as the local explanation, to calculate features' impact. Global feature importance presents the average attribution of a given feature across all the samples in the dataset. The domain experts were supplied with a number of local explanations as well. Those for presentation purposes were based on a single arbitrarily chosen true positive sample from the dataset (the sample is presented in Table 4.).

Local explanations generated for tax fraud detector include: SHAP force plots (Table 5), LIME (Table 6), anchors (Table 7), counterfactuals (Table 8). For interpretable models, the following were employed: Bayesian rule lists (Table 9), interpretation of linear regression's coefficients (Table 10), as well as the decision path from the decision tree created out of a neural network (Table 11). Following Python

Table 4 Exemplary true positive sample used for presentation of explanations and interpretable models

	Pesos sales	Labour Cost Sales	Labour Cost Net Sales	F931	Underreported Work Hours	Incorrect Rate	Excess Deductions
True Positive – encoded	2	0	1	1	0	1	0
Natural language meaning	Unknown	False	True	True	False	True	False

Table 5 SHAP force plots-based explanation

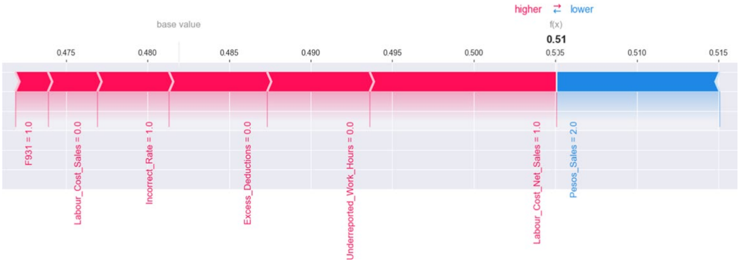
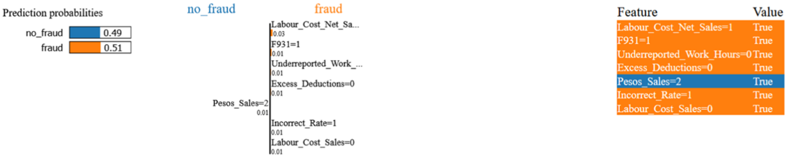
Explanation	
Natural language based explication	Force plots show how the resultant prediction was "pushed" towards the prediction by feature values. In the example below we can see that the values of F351, Labour_Cost_Sales, Incorrect_Rate, Excess_Deductions_Underreported_Work_Hours and Labour_Cost_Net_Sales contributed towards increasing the probability of this case being fraudulent. Pesos_Sales being 2 (i.e., no data about it is available) decreased the fraud proability by a small margin. This type of explanation can support the decision making process by showing which features are the most important for the given prediction

Table 6 LIME-based explanation

Explanation	
Natural language based explication	In this case, fraud was declared with 51% probability. Dually, there is a 49% probability of this case not being fraudulent, thus this case lies on the decision boundary. The most important features for the purpose of the "fraud" prediction are the values of Labour_Cost__Net_Sales and F931 (orange color in the diagram). A probability of this being not a fraudulent case can be explained with reference to missing data about Pesos_Sales. This type of explanation shows which features contribute to what extent towards fraud and no fraud predictions. This can be helpful when analyzing the logic behind making certain predictions. In terms of potential bias in the data it can be seen which specific feature is more dominant for the prediction and allows the users and the developer of the AI system to question if the data should be used for the intended purpose

libraries were employed: SHAP 0.41.0, LIME 0.2.0.1, anchor 0.0.2.0, DiCE 0.7.2. Out of the models used, LIME implements a local surrogate model-based explanation that uses an approximation of explained model to generate explanation based on the perturbations of original instance data. Anchors aim to explain individual predictions by finding a set of rules involving sample's features that cause the classifier to make a given prediction, irrespective of changes to its other features. In other words, "The anchors method explains individual predictions of any black box classification model by finding a decision rule that "anchors" the prediction sufficiently. A rule anchors a prediction if changes in other feature values do not affect the prediction" (Molnar 2020). However, generation of anchors has been problematic for this

Table 7 Anchor-based explanation

Explanation	Excess_Deductions = False AND
	Labour_Cost_Sales = False AND
	Incorrect_Rate = True AND
	F931 = True AND
	Labour_Cost_Net_Sales = True

Natural language based explication	Anchors present a rule which indicates which values of the features lead to the true positive prediction. Thus, according to this explainer, provided that the taxpayer did not take excess deductions and did not have undeclared sales (Excess_Deductions, Labour_Cost_Sales = no), but did declare incorrect proportional share, made F931 filling and had underdeclared sales, the prediction would have been positive, generally irrespective of other features, provided the non-noisy nature of the dataset

dataset, due to its low generalizability and noisy nature, causing the samples to lay at the edge of the classifier's decision boundary. Counterfactual explanations generate samples with altered features that, if fed to the classifier, would cause it to give an opposite prediction. Counterfactuals are thought to be human-friendly, because of their contrastive nature (they show how a given sample can be changed) and their selectivity (usually the number of feature changes is low) (Molnar 2020). Social sciences maintain that they offer a similar conceptual framework to that employed when humans explain their decision to each other; empirical evidence suggest that for such explanation to be useful, ultimately a domain expertise is nevertheless needed (Wang and Yin 2021).

4 Evaluation of the produced explanations and white-box AI systems

4.1 Methodology for and scope of the evaluation

The previous section shows an overview of various explanation techniques deployed at the prototype system under analysis as well as white-box AI models for tax fraud detection. These techniques and models were selected due to their widespread use in XAI practices, which means that their suitability as sources of explanation is largely accepted in the literature. Such acceptance, however, is insufficient for the legal purposes discussed in this paper. After all, most XAI techniques and white-box AI systems are developed for use in cooperative contexts, such as scientific discovery, in which all actors are aligned in their pursuit of knowledge (Creel 2020).This assumption of alignment does not hold true in legal contexts, as they usually are germane to contradicting interests of different parties, especially those of taxpayers (minimizing tax payments) and tax authorities (maximizing tax collection). Hence, both taxpayers and tax authorities will favour explanations that present their behaviour typically aiming to achieve different purposes in the best possible light. Moreover, white-box AI systems used by tax authorities to tax fraud detection will be subject to tax secrecy either entirely

Table 8 Counterfactual-based explanation

Explanation		Pesos sales	Labour Cost Sales	Labour Cost Net Sales	F931	Under-reported Work Hours	Incorrect Rate	Excess Deductions	Fraud
	True Positive	Unknown	False	True	True	False	True	False	True
	Counterfactual 1	Unknown	***True***	True	True	False	***False***	False	***False***
	Counterfactual 2	Unknown	False	True	***False***	***Unknown***	True	False	***False***
	Counterfactual 3	Unknown	False	True	True	False	True	***Unknown***	***False***
Natural language based explication	In the instance presented herein for the tax fraud dataset, the first row shows the original sample, the three rows below show examples that would cause the classifier to issue alternate decision; in this example – no fraud. By inspecting the table, we can see that in the first example (Counterfactual 1) should the value Labour Cost Sales be True (i.e. did occur in this case) and there would be no incorrect rate, the classifier would generate a "no fraud" prediction. The second counterfactual example shows that if no F931 filling was made, the classifier would once again yield a "no fraud" decision. The third counterfactual example shows that in case of lack of missing data regarding excess deductions, the classifier would also issue a "no fraud" prediction								

Table 9 Bayesian Rule List-based explanation

Explanation	IF Labour_Cost_Sales is true and Underreported_Work_Hours is true THEN probability of Fraud: 22.4% (16.3%-29.3%)
	ELSE IF Pesos_Sales is true THEN probability of Fraud: 28.9% (21.6%-36.8%)
	ELSE IF F931_0 > 0.5 and Underreported_Work_Hours is true THEN probability of Fraud: 56.8% (49.3%-64.2%)
	ELSE IF F931_0 > 0.5 THEN probability of Fraud: 6.4% (1.4%-14.8%)
	ELSE IF Incorrect_Rate is true and Labour_Cost_Sales is true THEN probability of Fraud: 51.0% (44.1%-57.9%)
	ELSE IF Labour_Cost_Sales is true THEN probability of Fraud: 27.2% (19.8%-35.3%)
	ELSE IF Incorrect_Rate is true and Labour_Cost_Sales is true THEN probability of Fraud: 66.7% (61.4%-71.7%)
	ELSE IF Labour_Cost_Sales is true THEN probability of Fraud: 63.4% (56.8%-69.7%)
	ELSE probability of Fraud: 98.1% (93.0%-100.0%)

Natural language based explication	In case of the example above, we can see that it is sufficient to check if it is true that there is a discrepancy regarding the labour cost sales accompanying underreported work hours to decide the activity is fraudulent, with 22.4% probability. If it is not so, then it is checked if pesos sales feature is problematic. If this is the case, then the sample is decided fraudulent (with 28.9% probability). If this condition is not fulfilled, the classifier moves on to check other values, according to the list above. The final else clause applies to samples that do not fulfill any of the rules. In that case, the sample is fraudulent with 98.1% probability

or at least to the extent to which it will not be possible (not without a sophisticated reverse-engineering) to understand their inner workings by taxpayers. However, for the purposes of this research we will evaluate not only XAI methods but also white-box AI systems in order to comprehensively cover the issue of XAI in tax law. Prospectively, the tax policy makers should consider reducing or uplifting tax secrecy whenever AI systems are used by tax authorities to the detriment of taxpayers without a sufficient degree of protection of their fundamental rights (Kuźniacki and Hadwick 2023b; 2023a). Hence, the evaluation of white-box AI system is prospectively of relevance. It is also important to keep in mind that the degrees of freedom involved in building an XAI models and white-box AI systems give tax authorities ample leeway to shape the explanation as seem fit to justify their diverging interests and purposes(Bordt et al. 2022). An assessment of XAI techniques and white-box AI models must consider that phenomenon: different interests and purposes of taxpayers and tax authorities shape different roles of the explanations play for each of them.

In the taxation domain, explanations are directly relevant for three categories of stakeholders (Kuźniacki et al. 2022). For tax authorities, explanations help with compliance with their reason-giving and transparency duties (Fink and Finck 2022), which are both meant to provide accountability toward society and allow the authorities to control and improve their internal processes. For taxpayers, explanations

Table 10 Interpretation of linear regression's coefficients

Explanation	Feature	Coefficient	Probability change
	Pesos_Sales is false	-0.23	– 21%
	Pesos_Sales is true	0.09	9%
	Pesos_Sales is unknown	0.15	15%
	Labour_Cost_Sales is false	0.15	15%
	Labour_Cost_Sales is true	0.19	21%
	Labour_Cost_Sales is unknown	-0.34	– 28%
	Labour_Cost_Net_sales is false	-0.4	– 32%
	Labour_Cost_Net_Sales is true	0.63	88%
	Labour_Cost_Net_sales is unknown	-0.23	– 20%
	F931 is false	0.24	27%
	F931 is true	-0.24	– 21%
	Underreported_Work_Hours is false	0.16	17%
	Underreported_Work_Hours is true	0.12	13%
	Underreported_Work_Hours is unknown	-0.29	– 24%
	Incorrect_Rate is false	-0.1	– 9%
	Incorrect_Rate is true	0.1	10%
	Excess_Deductions is false	0.47	60%
	Excess_Deductions is true	-0.84	– 56%
	Excess_Deductions is unknown	0.36	43%
Natural language based explication	In the case of logistic regression, the probability of the fraud prediction increases when the value of the coefficient is positive, by the value expressed by the formula $e^{coefficient}$. When the coefficient is negative, fraud probability decreases by the value of 1—$e^{coefficient}$. In the table below, the results are presented in the Probability change column In our example when Pesos_Sales is false (i.e. there is no discrepancy regarding sales), the coefficient equals to -0.23. Thus, the value of $e^{-0.23} \approx 0.79$, so the odds of the model predicting fraud goes down by 21%/ By taking the same approach, when Pesos_Sales is true, coefficient is 0.09, so the odds increase by 9% time when there is a discrepancy in this regard. The same goes for the rest of coefficients		

provide the information they need to evaluate whether their rights are observed in practice and to contest any decisions that fail to do so. For judges and other adjudicating actors, explanations must allow an unbiased view of the decision-making process, to allow effective oversight of the decisions involving AI. Each of these tasks requires different kinds of information, which in turn require different approaches to the production of explanations.

For this initial study, our paper focuses on the taxpayer's perspective. Previous work on user-centric evaluation of explanations (Speith 2022) has suggested that explanations should be evaluated in light of three criteria. The first one is the *comprehensibility* of the explanation produced by the XAI technique to its recipient. The second criterion is the *fidelity* of the XAI approach to the underlying logic of the original decision-making system. Finally, explanations must also be evaluated in light of their *assessability,* that is, by how much they help their recipient in evaluating whether the system meets its intended goals. These three criteria are broadly construed, and so it becomes necessary to specify their content in particular situations.

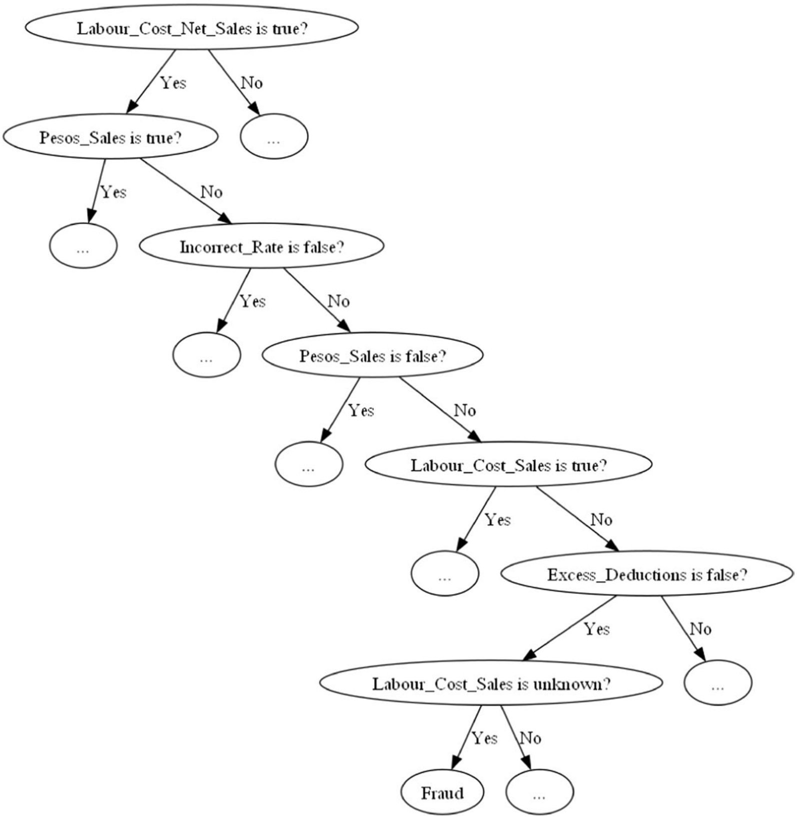
This paper focuses on the comprehensibility and assessability criteria.[3] Our assessment is aimed at sketching tax-specific dimensions of evaluation. We want to point out factors that need to be accounted for when applying a well-defined methodology (e.g. IEEE 7001–2021 standard for transparency of autonomous systems), but without focusing on the particulars of any audit methodology (which can be a matter for a follow-up study). In the evaluation below, comprehensibility refers to whether a taxpayer can make sense of a XAI system's outputs or the outputs of the white-box AI system: an output that can be understood with no AI expert support is very comprehensible, whereas an output that require more handholding is not as immediately accessible and might only be useful for taxpayers with more resources at their disposal or more knowledge. Assessability, in turn, is understood in terms of how much a system enables contestation of algorithmic outputs (Almada 2019; Kaminski and Urban 2021), for example by pointing out features or decision rules that warrant further scrutiny. In this paper, the evaluation of these features was conducted by two of the authors with legal background (BK & MA), as a baseline for future studies.

4.2 The evaluation

Following the order used in Sect. 3, this subsection provides an evaluation of XAI methods in terms of comprehensibility and assessability, followed by an evaluation

[3] We leave fidelity aside for two reasons. The first one, pointed out in (Bordt et al. 2022), is the fact that most explanations in tax are likely to be used in adversarial contexts, and so any evaluations of their fidelity must consider the particular forms of manipulation that might happen in such a context. Such an assessment is still at its early stages in scholarship. It is further complicated by the fact that fidelity may be affected by the organizational context in which an AI system is deployed, as factors such as trade and state secrecy may increase the margin for manipulating explanations: see (Busuioc et al. 2023).

Table 11 Decision path of a decision tree created out of a neural network

Explanation	 Labour_Cost_Net_Sales is true? Yes No Pesos_Sales is true? ... Yes No ... Incorrect_Rate is false? Yes No ... Pesos_Sales is false? Yes No ... Labour_Cost_Sales is true? Yes No ... Excess_Deductions is false? Yes No Labour_Cost_Sales is unknown? ... Yes No Fraud ...
Natural language based explication	The first node checks whether, in this case, there is a discrepancy regarding the labour cost net sales. In the case of current sample, it is so. Thus, the classifier moves on to check whether there is a issue regarding pesos sales. It is not so in this case, so the classifier moves on to check whether there is incorrect proportional share declared. It did not happen in this case, thus the classifier takes the "no" path and checks whether there is a discrepancy regardin labour cost sales. The "no" path is taken (in our case there was no such discrepancy), and subsequently we check the absence of excess deductions. It is present in our case, so finally we check whether there is missing data regarding labour cost sales. It is so in our case, so the case was classified as fraudulent

of interpretable AI methods along the same dimensions. The results of the evaluation are synthesized in Table 12, and further detailed in the following paragraphs of text.

The plus signs in Table 12 are a summary of a qualitative evaluation. Methods evaluated with "+" are deemed to have considerable limitations from a legal perspective, while a score of "+ +" indicates some virtues, and "+ + +" points out to a very positive evaluation. These scores are not meant as a comprehensive evaluation, but as a rule of thumb to suggest the usefulness of deploying specific methods in tax contexts.

4.2.1 XAI Methods

4.2.2 LIME

Since most popular LIME implementation provides graphical and numerical indication of the probability of tax fraud, and it highlights the features with bigger contribution to the outcome, a non-technical reader can extract information at a glance. LIME fares better when it comes to assessability, as the indication of features in the explanation provides a starting point for evaluation of the system's outputs. Since the method shows which features contribute to an outcome, and to which extent they do so, an explanation produced by LIME would allow interested parties to contextualize the algorithm's outputs, for example by identifying potential sources of bias.

4.2.3 SHAP

The SHAP approach can be used for two roles. On the one hand, SHAP can provide global explanations by presenting the global importance of a feature for the model. On the other hand, SHAP force plots allow us to identify the relevance of a feature for a particular outcome. While both approaches draw from the same techniques, they have different implications from the perspective of those who are consuming the information produced by the explanation technique.

Considering the non-technical audience intended for these explanations, a SHAP force plot is somewhat unintuitive. The idea of presenting features as vectors that pull in one direction or another is very useful if one is thinking in terms of gradients. However, it might mislead observers when it comes to identifying the tipping points that would make the output go one way or another, and it does not allow for the easy comparison between features in a particular case.

Contrastingly, the plot of the global importance of features allows untrained users to contrast a feature with others and have a notion of their relative importance. On the other hand, the general character of these explanations reduce their value in assessing particular cases, as it might be the case that a feature that is not particularly relevant in general becomes crucial for a taxpayer's specific situation, or vice-versa. Hence, global explanations provide suggestions of what taxpayers might look at, and authorities need to make clear that such general lines should not be mistaken for case-specific guidance.

4.2.3.1 Anchors Anchors provide explanations in natural language, framed in conditional terms: if this, then that. While the actual contents of such a conditional might not be straightforward to parse, this conditional structure allows taxpayers

Table 12 Comprehensibility and Assessability for the methods used in Sect. 3

Method	Comprehensibility	Assessability
XAI techniques		
SHAP global feature importance	+ +	+
LIME	+ +	+ + +
SHAP	+ +	+ +
Anchors	+	+ +
Counterfactuals	+	+ +
White-box AI methods		
Bayesian Rules List	+ + +	+ + +
Coefficient interpretation	+	+
Hybrid Decision Tree	+ +	+ + +

to identify decision rules that should be discussed in court because of their importance in determining tax consequences in a concrete tax case.

4.2.3.2 Counterfactuals Counterfactual explanations have the potential to be useful for lawyers, as their selective and contrastive nature makes them closer to the kind of explanation produced by human beings. Current implementations of counterfactuals, however, fall short of that potential. Whenever the change in the outcome is due to a change in present values, the reader of the table might be able to turn that counterfactual into actionable information. The features are to be read the same as in the case of the dataset itself (cf. Table 1 and its description). For example, the change in the feature Excess_Deductions, as evidenced in the Table 8 with counterfactuals, i.e. from the presence of the fact denoted by that feature (1) to the lack of such presence (0), leads to a change in the outcome, and so taxpayers might adjust their behaviour in the future. However, the meaning of the counterfactual is less clear when the absence of data leads to a change in outcome: for example, when the second counterfactual leads to no tax fraud detection because of the change in the data for Labour Cost Sales – from the absence of the fact (0) to missing data about it (2).

From the perspective of taxpayers, the low comprehensibility of counterfactuals makes this method less assessable (weak + +), as the interpretation of the XAI outputs will require expert interpretation before it can be used. While some taxpayers, particularly those with more resources, might be able to extract more insights from the outputs of the counterfactual model tested above, this is unlikely to be the case of the ordinary taxpayer, who is not supported by resources of major tax advisory firms. Contrastingly, a more comprehensive formulation of the counterfactuals might make the outputs more useful for users, even if it comes at the expense of some of their potential for assessment in the hands of sophisticated users.

4.2.4 Interpretable methods

4.2.4.1 Bayesian rule list A non-technical reader might get easily acquainted to the outputs of a Bayesian Rule List. Such a list relies on the kind of if–then-else rules described for anchors, which are also presented in natural language. However, rule lists have the advantage of allowing users to use these rules to evaluate both the general behaviour of the system and its application to particular cases. As such, the technique is very suitable for situations in which taxpayers might need to make adversarial inquiries to contest discriminatory and arbitrary outputs of tax AI systems.

4.2.4.2 Coefficient interpretation While numerical coefficients might be understandable for software developers and domain experts, a taxpayer might lack the context needed to understand the relative magnitude of a coefficient. As such, the interpretation of coefficients for a logistic regression might be useful for taxpayers if they are mathematically savvy or supported by technical experts. Otherwise, it might not provide much in terms of actionable insights for evaluating or contesting the decisions made by the algorithm.

4.2.4.3 Hybrid neural network with decision tree solution In abstract terms, the logic of such a hybrid solution is very accessible to the lay taxpayer. Depending on the positive or negative answer regarding each feature in each node, this decision tree goes down through all nodes and branches until a definitive classification of tax fraud or the lack of it is revealed. In practice, however, the sheer number of factors involved in decision-making might result in trees that exceed human cognitive abilities (Miller 1956).

In the example presented in Sect. 3, we have used some techniques to reduce the complexity of the tree, such as presenting only a bit of it. But, even though the figure presents only a tiny branch of that tree, it is still difficult to understand. In addition, looking at only a part of the tree may mislead observers: not only the same decision might be achieved by a different combination of the factors in that tree, but a particular branch of the tree might not include all elements considered in the decision. As such, much of the value of the decision tree approach will depend on what approaches are used to select the part of the tree that is presented as an explanation.

5 Conclusions

In this paper, we have introduced a new dataset, synthetically generated by Buenos Aires tax authorities based on real-life, tax-related data. This presentation was supplemented by a study aimed at uncovering the feature requirements for explainable system that such authorities may be interested in using, as well as the technical study in which we compared several explainers.

LIME scores the best in the evaluation of XAI methods but in the current legal system setup not enough to meet the minimum standard for direct use in tax decision making. For that to happen, it would need to be slightly more comprehensible. Counterfactuals, if more comprehensible, will also be a good candidate to contribute to explainable tax AI. Perhaps a good approach would be to design a XAI method which would merge LIME and counterfactuals with high (at least strong++) comprehensibility in mind. This would also translate to very high (+++) assessability, thereby creating an XAI method which meets a minimum standard for explainable tax AI.

When it comes to interpretable models, the Bayesian Rules List scored the best overall, and its outputs are also clearer than those produced by all XAI methods. This approach appears to ensure the explainability of tax AI for taxpayers in the best way due to its strong comprehensibility and assessability. The hybrid decision model also showed some potential, but it ultimately is not comprehensible enough for use by most taxpayers. To address these shortcomings, a potential direction for development would combine a Bayesian Rules List with a deep neural network, using the former to reflect the latter's output production logic. By doing so, it might be possible to provide explanations that are even more informative than the Bayesian Rules List for the purposes of tax AI.

In the nearest timeframe, we aim to mitigate the limitations of the technical study, strengthening the contributions of this paper. Firstly, we base the crux our findings on qualitative analyses and base them on the various experiences of the authors, overcoming the difficulties that tend to arise from such co-operations (Ratcheva 2009). Nevertheless, the dataset presented herein calls for a more quantitative study to be performed. Secondly, the dataset is based on features identified by the tax authorities themselves. Although they were screened on the needs of ML-based system, there is a distinct possibility that a better performance could be attained if additional features could be brought into the dataset. This work was able to highlight the challenges encountered by computer scientists when they develop explainable systems, as well as the lawyers' expectations towards such systems. Thus, it can serve as a background work that guides how to ensure taxpayers' rights and increases tax morale in the AI-reliant areas of tax law. This goes on to show that transparent and explainable AI contributes to a more equal distribution of that knowledge. Moreover, this can facilitate the creation of new user-centric and domain-specific XAI methods, a challenge in its own right.

Funding Open access funding provided by European University Institute - Fiesole within the CRUI-CARE Agreement. The study was partly funded by IBM-Notre Dame University Tech Ethics Lab. In a smaller part, the study was also funded by the Amsterdam Centre for Tax Law (ACTL) research project "Designing the tax system for a cashless, platform-based and technology-driven society" (assistant professorship of Błażej Kuźniacki) and Fundacion Carolina (doctoral scholarship of Marco Almada)

References

Adibi J, Cohen PR, Morrison CT (2004) Measuring confidence intervals in link discovery: a bootstrap approach. Proceedings of the ACM Special Interest Group on Knowledge Discovery and Data Mining (ACM-SIGKDD-04)

Agrawal T, Agrawal T (2021) Hyperparameter optimization using scikit-learn. Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient, 31–51

Almada, M (2019) Human intervention in automated decision-making: toward the construction of contestable systems. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19, 2–11. Montreal, QC, Canada: ACM Press. https://doi.org/10.1145/3322640.3326699.

Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making:'automation bias' and 'selective adherence' to algorithmic advice. J Public Administr Res Theory 33(1):153–169

Amnesty International. 2021. "Xenophobic Machines: Discrimination through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal." Amnesty International. October 25, 2021. https://www.amnesty.org/en/documents/eur35/4686/2021/en/.

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115

Aytekin C (2022) Neural networks are decision trees. arXiv. http://arxiv.org/abs/2210.05189.

Babic, B, Glenn Cohen I (2023) The algorithmic explainability'bait and switch. Minnesota Law Review 108

Bardutzky, S (2022) Duty to provide reasons. Oxford Public International Law. 2022. https://doi.org/10.1093/law-oeeul/e57.013.57

Barenstein M (2019) ProPublica's COMPAS Data Revisited." *arXiv Preprint* arXiv:1906.04711.

Barysė D and Sarel R (2023) Algorithms in the court: Does it matter which part of the judicial decision-making is automated? Artif Intell Law, 1–30

Bell A, Solano-Kamaiko I, Nov O, Stoyanovich J (2022) It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, 248–266

Bench-Capon T. (1993) Neural Networks and open texture. In: Proceedings of the 4th International Conference on Artificial Intelligence and Law, 292–297

Bibal A, Lognoul M, De Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. Artif Intell Law 29:149–169

Bordt S, Michèle Finck, Raidl E and von Luxburg. U (2022) Post-Hoc explanations fail to achieve their purpose in adversarial contexts. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 891–905. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533153.

Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, Pfaff M, Liao B (2021) Scalable and explainable legal prediction. Artif Intell Law 29(2):213–238

Busuioc M, Curtin D, Almada M (2023) Reclaiming transparency: contesting the logics of secrecy within the AI act. Eur Law Open 2(1):79–105. https://doi.org/10.1017/elo.2022.47

Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N (2023) A survey on XAI and natural language explanations. Inf Process Manage 60(1):103111

Chang, Felix, Erin McCabe, and James Lee. 2020. "Mining the Harvard Caselaw Access Project." *Available at SSRN 3529257*.

Collosa, Alfredo. 2021. "Use of Big Data in Tax Administrations." September 1, 2021. https://www.ciat.org/use-of-big-data-in-tax-administrations/?lang=en.

Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, Holmes JH (2022) A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 133:102423

Craig P (2020) Legality: six views of the cathedral 233–256

Creel KA (2020) Transparency in complex computational systems. Phil Sci 87(4):568–589. https://doi.org/10.1086/709729

Curtin D (2020) The EU automated state disassembled. Essays in Honour of Paul Craig. Oxford University Press, In The Foundations and Future of Public Law

Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif Intell 299:103525. https://doi.org/10.1016/j.artint.2021.103525

De Cooman, Jerome. 2023. "Outsmarting Pac-man with artificial intelligence, or why ai-driven cartel screening is not a silver bullet. J Eur Compet Law Pract lpad017.

Demková S (2021) The decisional value of information in European semi-automated decision-making. Rev Eur Administr Law 14(2):29–50. https://doi.org/10.7590/187479821X16254887670874

Durkan C, Bekasov A, Murray I, Papamakarios G (2019) Neural spline flows. Adv Neural Inf Process Syst 32

Esposito E (2022) Transparency versus explanation: the role of ambiguity in Legal AI. J Cross-Disciplinary Res Computat Law 1 (2)

Ferrario A, Loi M (2022) How explainability contributes to trust in AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1457–1466

Fink M, Finck M (2022) Reasoned A (I) dministration: explanation requirements in EU law and the automation of public administration. Eur Law Rev 47(3):376–392

Górski Ł, Ramakrishna S (2021) Explainable artificial intelligence, lawyer's perspective. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 60–68. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3462757.3466145

Górski Ł, Ramakrishna S, Nowosielski JM (2020) Towards Grad-CAM Based Explainability in a Legal Text Processing Pipeline. Extended Version. In AI Approaches to the Complexity of Legal Systems XI-XII, 154–68. Springer

Hadwick D (2022) Peer reviewed articles: 'behind the one-way mirror: reviewing the legality of EU tax algorithmic governance. EC Tax Rev 31(4).

Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(4):e1312

Kaminski ME, Urban JM (2021) The right to contest AI. Columbia Law Rev 121(7):1957–2048

Kolkman D (2022) The (in) Credibility of Algorithmic models to non-experts. Inf Commun Soc 25(1):93–109

Kuźniacki B, Hadwick DRG (2023a) (Non)Natural Born Killers of Xai in Tax Law: The Roadmap toward Holistic Explainability. Kluwer International Tax Blog. https://kluwertaxblog.com/2023/09/15/nonnatural-born-killers-of-xai-in-tax-law-the-roadmap-toward-holistic-explainability/.

Kuźniacki, Błażej, Marco Almada, Kamil Tyliński, and Łukasz Górski. (2022) Requirements for Tax XAI Under Constitutional Principles and Human Rights. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, 221–38. Springer

Kuźniacki B, Almada M, Tyliński K, Górski Ł, Winogradska B, Zeldenrust R (2022) Towards eXplainable Artificial Intelligence (XAI) in Tax Law: The Need for a Minimum Legal Standard. World Tax J 14

Kuźniacki B (2023b) (Non)Natural Born Killers of Xai in Tax Law: Trade Secrecy, Tax Secrecy and How to Kill the Killers. Kluwer International Tax Blog. https://kluwertaxblog.com/2023/09/12/nonnatural-born-killers-of-xai-in-tax-law-trade-secrecy-tax-secrecy-and-how-to-kill-the-killers/.

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI Research. Artif Intell 296:103473. https://doi.org/10.1016/j.artint.2021.103473

Mehdiyev, Nijat, Constantin Houy, Oliver Gutermuth, Lea Mayer, and Peter Fettke (2021) Explainable Artificial Intelligence (XAI) Supporting Public Administration Processes–on the Potential of XAI in Tax Audit Processes. In: Innovation Through Information Systems: Volume I: A Collection of Latest Research on Domain Issues, 413–28. Springer

Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63(2):81

Molnar C (2020) Interpretable Machine Learning. Lulu.com. https://christophm.github.io/interpretable-ml-book/index.html#summary.

Oswald M (2018) Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. Phil Trans Royal Soc Math Phys Eng Sci 376(2128):20170359

Reiling A Dory (2020) Courts and artificial intelligence. In IJCA, 11:1. HeinOnline

Rissland EL, Ashley KD, Loui RP (2003) AI and law: a fruitful synergy. Artif Intell 150(1–2):1–15

Robbins S (2019) A misdirected principle with a catch: explicability for AI. Mind Mach 29(4):495–514. https://doi.org/10.1007/s11023-019-09509-3

Savelka J, Ashley KD (2021) Discovering explanatory sentences in legal case decisions using pre-trained language models. arXiv Preprint arXiv:2112.07165

Schauer F (1994) Giving reasons. Stan l Rev 47:633

Schmitz GPJ, Aldrich C, Gouws FS (1999) ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. IEEE Trans Neural Netw 10(6):1392–1401

Schweighofer E (2022) Rationale discovery and explainable AI. In Legal Knowledge and Information Systems: JURIX 2021: The Thirty-Fourth Annual Conference, Vilnius, Lithuania, 8–10 December 2021, 346:225. IOS Press

Semenova L, Rudin C, Parr R (2022) On the existence of simpler machine learning models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1827–1858.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–86. New York NY USA: ACM. https://doi.org/10.1145/3375627.3375830

Sourdin T (2022) What if judges were replaced by AI? Turkish Policy Quarterly

Speith T (2022) How to evaluate explainability?-a case for three criteria. In 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW), 92–97. IEEE

Villegas Héctor B (2001) Curso de Finanzas, Derecho Financiero y Tributario. *Buenos Aires-Argentina*

Vreš D and Robnik-Šikonja M (2022) Preventing deception with explanation methods using focused sampling. Data Mining Knowl Discov 1–46

Wagner B (2019) Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. Policy Internet 11(1):104–122

Walker, Vern R, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning." In *ASAIL@ ICAIL*.

Wang X, Yin M (2021) Are explanations helpful? A comparative study of the effects of explanations in Ai-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces, 318–328

Yu J, Cristea AI, Harit A, Sun Z, Aduragba OT, Shi L, Moubayed NA (2022) Interaction: a generative XAI framework for natural language inference explanations. In: 2022 International Joint Conference on Neural Networks (IJCNN), 1–8. IEEE

Zandstra T, Brouwer E (2022) Fundamental Rights at the Digital Border—The Digital Constitutionalist. June 28, 2022. https://digi-con.org/fundamental-rights-at-the-digital-border/.

Zareapoor M, Shamsolmoali P (2015) Application of credit card fraud detection: based on bagging ensemble classifier. Procedia Comput Sci 48:679–685

Zerilli J, Bhatt U, Weller A (2022) How transparency modulates trust in artificial intelligence. Patterns

Authors and Affiliations

Łukasz Górski[1,2] · Błażej Kuźniacki[3,4,5] · Marco Almada[6] · Kamil Tyliński[7,15] · Madalena Calvo[8] · Pablo Matias Asnaghi[9] · Luciano Almada[9] · Hilario Iñiguez[10] · Fernando Rubianes[11] · Octavio Pera[11] · Juan Ignacio Nigrelli[12,13,14]

✉ Marco Almada
Marco.Almada@eui.eu

Łukasz Górski
lgorski@icm.edu.pl

[1] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland

[2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University in Toruń, Toruń, Poland

[3] Łazarski University, Warsaw, Poland

[4] Centre for AI and Data Governance (CAIDG), The Singapore Management University, Bras Basah, Singapore

[5] PwC Netherlands, Amsterdam, The Netherlands

[6] ASPIRE Programme, Department of Law, European University Institute, Florence, Italy

[7] DLT Science Foundation, London, UK

[8] Amazon Web Services, London, UK

[9] Buenos Aires City Tax Authority, Buenos Aires, Argentina

[10] Tax Management Systems, Buenos Aires, Argentina

[11] Datopia, Buenos Aires, Argentina

[12] Universidad Mayor de San Andres, La Paz, Bolivia

[13] Tax Law , University of Buenos Aires, Buenos Aires, Argentina

[14] Criminal Law and Criminology, Buenos Aires City Tax Authority, Buenos Aires, Argentina

[15] Centre for Blockchain Technologies, University College London, London, UK