A Conversational Interface for interacting with Machine Learning models

Davide Carneiro[1,2][0000−0002−6650−0388], Patrícia Veloso[1][0000−0002−0779−9076], Miguel Guimarães[1][0000−0003−0573−9122], Joana Baptista[1], and Miguel Sousa[1][0000−0001−8838−6419]

[1] CIICESI/ESTG, Politécnico do Porto, Portugal
{dcarneiro,pamv,8150520,8130144,8160204}@estg.ipp.pt
[2] Algoritmi Centre/Department of Informatics, Universidade do Minho

Abstract. As Machine Learning and other fields of Artificial Intelligence are increasingly used for automating the most diverse aspects of our day-to-day life, so too increases the scrutiny and accountability that these technologies are subject to. Many issues that were previously only attributed to Human decision-makers, such as prejudice or bias, can now also be seen in these automated means. To add up, these technologies are often harder to scrutinize and understand, and can take (eventually wrong) decisions at a much faster rate than Humans, thus having a far more potential impact. Trust, transparency, explainability, interpretability and accountability thus become desirable properties of AI systems. In this paper we start with an analysis of some Legal and Ethical considerations regarding the use of AI and related technologies. We then detail an approach whose main goal is to improve the ability of a ML system to explain its decisions, based on a conversational chatbot. The main goal is that the user can interact with and question the model regarding its predictions and, through this, gain an increased confidence on the model, and a better understanding of how it works.

Keywords: Machine learning · explainable AI · chatbot

1 Introduction

Over the past decade, Artificial Intelligence (AI) has assumed an unprecedented role in virtually all domains, quickly replacing or complementing Human decision-makers in their tasks. Jobs that were previously exclusive to Human experts are now becoming hybridized, as automated tools based on Machine Learning, Expert Systems, or others start to be increasingly used. Examples include the prediction of recidivism [23], the hiring of Human Resources [13], medical diagnosis [18], credit scoring [25] or, more recently, autonomous driving [12].

AI in general, and ML in particular, have thus gained a central role in many of today's activities. As a consequence, models, algorithms, tools and even AI developers must also have an increased responsibility over the consequences of the use of these algorithms or models [9].

One of the main causes for concerns comes from the automated nature of these approaches, combined with their far greater decision speed when compared with Humans. Indeed, if a Human practitioner is making a certain mistake in their decisions, they can only do so much harm, in the sense that Humans make decisions at a relatively slow rate. Automated tools, on the other hand, can make thousands or millions of decisions while a Human makes a single one, thus risking having much far-reaching consequences.

Moreover, Human decisions tend to be easier to scrutinize. Either because they are made at a slower pace (thus giving more time for an analysis), because they can easily come with a justification from the decision-maker, or because Human decision-makers are more restricted in their actions by legislation or other limits, which they consciously acknowledge.

Automated decisions are also often deemed *better* in the sense that they are thought to be free of prejudice or bad intentions. Indeed, there is no such thing as a good or bad intention or motivation in autonomous agents, as there may be in Human decision making. However, this does not mean that autonomous decisions are free from prejudice or other vices. Frequently, models end up showing just the same prejudices that we find in Human actions, as these were there in some point of the process, namely when collecting or selecting data. Recent examples include the existence of racism in the prediction of recidivism by the COMPAS model [23], or the sexist decisions of Amazon's hiring algorithm which favored male candidates over female ones [13].

All these challenges are becoming increasingly worrying in a time in which models grow in complexity. Indeed, most of today's models can be classified as black-box models, in the sense that they are so complex or opaque that it is nearly impossible for a Human to understand how a model works, to predict its behavior, or even to understand the reasons for a single prediction. In these conditions, it becomes increasingly difficult for a Human to debug a model and understand if and when its predictions have some prejudice, bias or any other undesirable property.

Given this current state of affairs, new methods are needed to better explain the inner workings of ML models, so that the quality of their decisions can be more efficiently monitored, not only in statistical terms but also under principles such as transparency, equality or justice.

This paper provides a brief analysis of the Ethical and Legal framework under which autonomous agents with decision responsibilities should operate. We then describe a new approach for ML interpretability and explainability that uses a proxy explainable model together with a conversational chatbot, to allow a Human to interact with an existing ML model and obtain explanations that allow to better understand its prediction.

While the proposed approach is independent of the domain or of the underlying model used, we describe a case study in which the current work is being developed, in the context of Tax Fraud detection in Portugal. In the context of the NEURAT project (Intelligent Digital Audit Knowledge Base Engine), an interactive Machine Learning system has been developed, that gradually learns

from the actions of the auditors (further described in Section 4). This is one of the many examples in which the user (Auditor) is generally not an expert in Machine Learning, but still will need to look at a prediction of a model and take a decisions (partially) based on it. The goal of this work is to increase the transparency of the system and, consequently, the trust of the user on the system, by improving the ability of the system provide explanations in a way that meet each users' information needs.

2 Legal and Ethical considerations

2.1 Ethical Principles

When it comes to Ethical considerations regarding AI, many different aspects can be analyzed. It can be argued that Ethic considerations should be present in every step of the process, from data creation to data storage and destruction. In this section we analyze ethical implications at three different levels: at the data level, at the model training level, and at the model evaluation level.

At the bottom there are the so-called Data Ethics, which can be seen as the analysis of the ethical implications that decisions at the data-level have [5].

In this level, one of the most frequently found problems is bias in data. Data bias happens when a given set of data is not representative of the actual phenomenon being studied. There may be multiple reasons for the occurrence of data bias [24]. It may happen due to the small size of the dataset, that may just not be large enough to capture the entirety of the phenomenon, and is thus unrepresentative. In these cases, only a part of the patterns will be represented in the data. Another frequent reason is that data collection/selection/curation processes are often controlled or defined by Humans. The prejudice or bias of Humans will often pass on to the collected data, most of the times unconsciously, because certain variables or instances were left out or deliberately removed. These forms of bias constitute what can be called as Selection Bias.

Nonetheless, other forms of bias exist. When data is obtained from online sources, namely social networks, the so-called Response Bias may occur [15]. This happens when data comes from few or unrepresentative data sources, like when a small group of users (and thus unrepresentative) produces most of the data.

While other sources and types of bias may exist, they result in the same problem: biased data will constitute a biased representation of the phenomenon, which will in turn result in biased decision models that may eventually be unethical, namely by discriminating against certain groups.

Data bias may go unnoticed for several reasons. Frequently, it is due to lack of transparency, caused by poor data quality or by leaving out certain variables due to privacy requirements. Thus, transparency often clashes with other equally important principles such as privacy and security. Other times, data is proprietary, which prevents third parties from accessing it and look for problems. In any case, data transparency is critical in the era of big data and the massive use of deep learning techniques [5].

Ethics are also paramount when training a model. One of the most fundamental issues is related with the use of complex models, which are difficult to explain and understand: the so-called black box models. Unfortunately, there is a trade-off between accuracy and explainability: the most accurate models (which also tend to be the most complex ones) are also the ones that are harder to explain. This means that when explainability is a requirement, accuracy is often sacrificed. The best example of this duality can be found in Deep Learning models [16].

Deep Learning predictions, as those of other models, can always be explained by providing all the computations that led to the model or to a particular prediction. The sheer complexity of this, however, prevents us from actually being able to use that information to understand the model or its predictions. A distinction must thus also be established between explainability and interpretability: the former being the ability of the model to explain itself, and the latter the ability of the Human to understand the explanation. One can only solve this problem by addressing both sides.

Understanding a model is however more difficult than understanding a single prediction. This complexity then becomes a matter of *trust*: how can we fully trust a model that we do not understand? This is sill more relevant in critical domains, such as with self-driving cars. These have been shown to be easily tricked by simple changes in images that would never trick a Human driver, for instance by placing small stickers on road signs. Moreover, the way their behavior is affected is completely unpredictable, because we do not understand how the models work, and autonomous cars have been shown to drive above speed limits, swerve into the wrong lane or simply ignoring street signs.

Finally, Ethical problems can also come from the way we evaluate models. Typically, a ML model is evaluated on a so-called hold-out dataset, that is, a set of data drawn from the same source but not used during training, to properly evaluate the ability of the model to generalize. Common metrics include the root-mean-square error, accuracy, precision, recall, f1-score, AUC curve, among others.

The hold-out dataset, however, can suffer from the same problems of the training data. Namely, it can also be biased or unrepresentative. Thus, the exact same model can be deemed *good* or *bad* (both statistically or Ethically) depending on the data it is evaluated on, or on the metrics selected (e.g. metrics such as accuracy are often misleading when others such as precision or recall are not considered).

Understanding the roots and biases of data, model and algorithms allows us to evaluate the idea of a transparency requirement. An effective transparency would need to offer an explanation that is very useful [8]. Improper use of data and algorithms may lead to discrimination, wrong decisions, and other adverse effects [5].

Transparency is also a predisposition for accountability [7]. But, the ideal of transparency raises several questions: what exactly needs to be revealed to the data subject? How detailed does the explanation have to be? [7].

In the context of AI systems, the main ethical principles are: (i) respect for human autonomy; (ii) prevention of harm; (iii) fairness; and (iv) explicability [10].

These principles must be translated into concrete requirements to achieve Trustworthy AI, so, we have [10]: (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, non-discrimination and fairness; (vi) societal and environmental well-being; (vii) Accountability, including auditability, minimisation and reporting of negative impact, trade-offs and redress.

AI technology can provide sufficient transparency in explaining how AI decisions are made [20]. Transparency can often be achieved through retrospective analysis of the technology's operations [20]. Sometimes, transparency can be more challenging, even limiting the use of some AI technologies such as neural networks [20].

Transparency concerns are driven by a certain logic: observation produces insights which create the knowledge required to govern and hold systems accountable [1]. The more facts revealed, the more truth that can be known through a logic of accumulation [1]. The more that is known about a system's inner workings, the more defensibly it can be governed and held accountable [1].

However, there are some limits of the transparency ideal, as follows: transparency can be disconnected from power; transparency can be harmful; transparency can intentionally occlude; transparency can create false binaries; transparency can invoke neoliberal models of agency; transparency does not necessarily build trust; transparency entails professional boundary work; transparency can privilege seeing over understanding; transparency has technical limitations; transparency has temporal limitations [1].

Among the obstacles to algorithmic transparency, we have: (i) technical obstacles; (ii) intellectual property obstacles; and (iii) state secrets and other confidential information of state authorities [7].

Fundamental rights are a basis for Trustworthy AI. In this particular, we refer to: respect for human dignity; freedom of the individual; respect for democracy, justice and the rule of law; equality, non-discrimination and solidarity; as well as citizens' rights [10]. The fundamental rights upon which the EU are founded and directed towards ensuring respect for the freedom and autonomy of human beings [10].

2.2 Explainability

Explanation has been a central feature of AI systems for legal reasoning since their inception [4]. Paradigms underlying this problem fall within the so-called eXplainable AI (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models [3].

AI researchers and practitioners have focused their attention on explainable AI to help them better trust and understand models at scale [2]. Explainability is a prerequisite for building trust and adoption of AI systems in high stakes domains requiring reliability and safety such as healthcare and automated

transportation, and critical industrial applications with significant economic implications such as predictive maintenance, exploration of natural resources, and climate change modeling [2].

In a legal dispute, there will be two parties and one will win and one will lose, then, losers have a right to an explanation of why their case was unsuccessful [4]. Given such an explanation, the losers may be satisfied and accept the decision, or may consider if there are grounds to appeal [4].

The challenges for the research community include [2]: (i) defining model explainability; (ii) formuating explainability tasks for understanding model behavior and developing solutions for these tasks; and (iii) designing measures for evaluating the performance of models in explainability tasks.

The right to explanation is viewed as a promising mechanism in the broader pursuit by government and industry for accountability and transparency in algorithms, artificial intelligence, robotics, and other automated systems [11].

2.3 General Data Protection Regulation

The EU General Data Protection Regulation (GDPR), which comes into force in EU Member States in May 2018, modernizes a European data protection regime that dates back a quarter century [6]. A number of provisions in the GDPR seek to promote a high degree of transparency in the processing of personal data [6].

For example, where personal data are obtained from the data subject, Article 13(2)(f) requires data controllers to provide data subjects with information about "the existence of automated decision making, including profiling and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." [6]. The GDPR also introduces an explicit accountability principle that was arguably only implicit in the former Data Protection Directive [6].

The GDPR distinguishes general profiling, decision-making based on profiling and solely automated decision-making about individuals including profiling. "Profiling" is defined in Article 4(4) as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" [6].

Besides, there are further restrictions where decisions are based on special categories of personal data. Consent might appear to offer a strong mechanism for legitimating automated decision making and profiling [6]. "Consent" is defined as "any freely given, specific, informed and unambiguous indication of a data subject's wishes by which he or she by a clear affirmative action signifies agreement" [6].

Rules of the General Data Protection Regulation on automated decision making in the age of Big Data and to explore how to ensure transparency of such decisions, in particular those taken with the help of algorithms [7].

The core of data protection law consists of eight principles, which can be summarised as follows [26]: (a) personal data may only be processed lawfully, fairly and transparently; (b) such data may only be collected for a purpose that is specified in advance, and should only be used for purposes that are compatible with the original purpose; (c) organisations should not collect or use more data than necessary; (d) organisations must ensure that such data are sufficiently accurate and up to date; (e) organisations should not store the data for an unreasonably long time; (f) organisations must ensure data security; (g) the organisation that determines the purposes and means for processing (the "controller") is responsible for compliance.

GDPR does not, in its current form, implement a right to explanation, but rather what we term a limited "right to be informed" [11].

3 Architecture

In a traditional Machine Learning setting, a user interacts directly with a model to obtain predictions. The main disadvantage of most of current ML models, as already addressed, is the lack of an explanation to accompany the prediction. This makes it harder for a Human to understand the reasons for the prediction, prevents a deeper understanding of the model and of the phenomenon being studied, hides eventual bias and/or prejudice problems, and ultimately decreases the trust of the user in the system.

In this paper we propose a new approach that introduces two key components: a conversational interface and an explainable model (Figure 1). The conversational interface is implemented in the form of a chatbot and constitutes the main point of interaction of the user with the system. The explainable model is a proxy model, that may be very different from the predictive model (the "main" model), that is used to build explanations for the predictions provided by the system. Thus, the user does not interact directly with the predictive model anymore. Instead, it interacts with the chatbot to obtain predictions, which come accompanied by a basic explanation. The user can then interact with the chatbot to drill down into the explanation, obtaining further detail as needed.

Both the predictive and the explainable models interact with the conversational chatbot through a REST API. This abstracts the models' functionality, makes integration easier, and allows for easily swapping the Machine Learning frameworks used (e.g. scikit-learn, H2O) for training models without changing the conversational interface. Methods for training or updating the predictive model are not exposed to the conversational interface as these are intended to be used by a user with a different role (e.g. ML Engineer). To the extent of this paper, a user is thus viewed as the client of the system, who uses it to obtain predictions/explanations for a given context, and who is not necessarily an expert on Machine Learning. The goal is indeed that any user can obtain predictions and explanations just by writing questions in natural language.

From the conversational interface's point of view, the predictive model's API exposes methods for getting predictions for an instance or group of instances, as

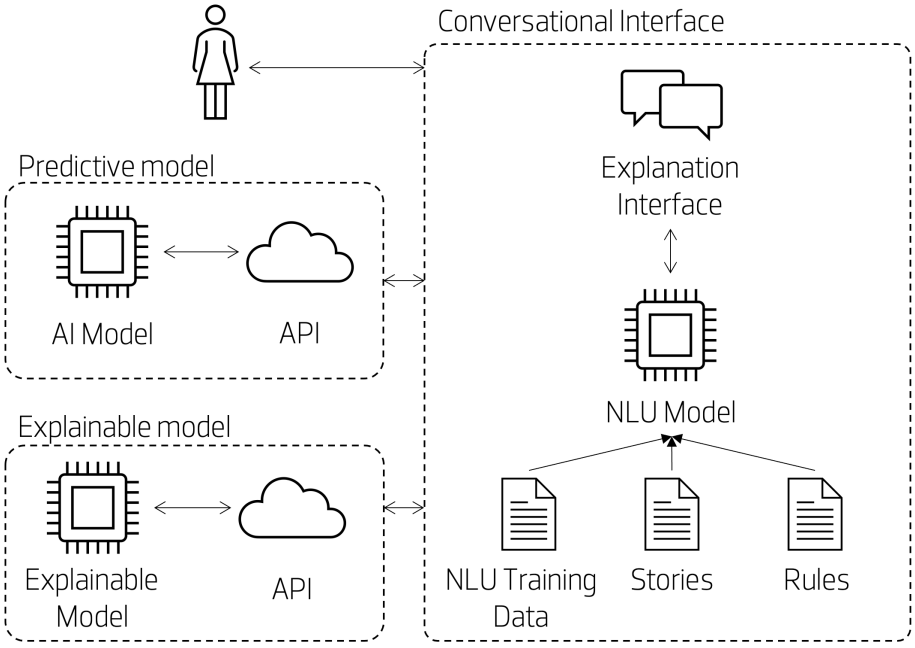

Fig. 1. Architecture of the proposed solution.

well as for getting meta-data about the model (e.g. performance metrics, training date). The explainable model, on the other hand, exposes services for getting explanations, which may include different elements further detailed below (e.g. textual explanations, feature relevance, statistical measures).

In its current form, the user interacts with the conversational interface through a console. In a future version, however, the chatbot will be integrated at the application level (Figure 5). When a value that results from a prediction is shown to the user, a basic explanation will also be shown. Then, the user will be able to open the chatbot by clicking a button, and will be able to interact with it using natural language to drill down on the explanation and obtain additional detail that meets her/his information needs.

The two key elements of the proposed system (i.e. the explainable model and the explainable interface) are further detailed in Sections 3.1 and 4, respectively.

3.1 Explainable Proxy Model

The main goal of the explainable model is to provide explanations for the predictions of another model. In that sense, it acts as a proxy model. An explainable model is trained for each predictive model whose predictions need to be explained. That is, the explainable model does not actually explain the predictive model, but rather each of its predictions. Moreover, it can explain predictions

from any type of model, including black box models, as it is independent of their internal structure.

Types of Explanations The explainable model is trained from a modified version of the CART algorithm [22]. This is a Decision Tree algorithm. In any Decision Tree, each node of the tree contains boolean rules about the observations (e.g. if feature X is greater than y) and each leaf contains the result of the prediction for a given path in the tree. While the tree is being built, the training set is increasingly split at each node, leading to smaller and better grouped subsets of the data. This splitting process ends when one or more stopping criteria are met, which may include a minimum size of the split or a minimum degree of variance/purity.

In this context, variance denotes how much the values for the dependent variable of a split are spread around their mean value, in regression tasks. Purity, on the other hand, considers the relative frequency of classes. If all classes have roughly the same frequency, the node is deemed "impure". The Gini index is used in the CART algorithm to measure impurity [14]. In terms of decision-making, an impure node (or one with high variance) represents a low level of confidence. That is, the tree provides a prediction but one that is based on data that does not have a clear tendency.

The relationship between the outcome y of a Decision Tree and a given feature x can be described by Formula 1[17]. Each instance of the training set is attributed to a single leaf node (subset R_m). $I\{x \in R_m\}$ is a function that returns 1 if x is in the subset R_m or 0 otherwise. In a regression problem the predicted outcome $\widehat{y} = c_l$ of a leaf node R_1 is given by the average value of the instances in that node.

$$\widehat{y} = \widehat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{1}$$

Given this, it can be stated that a Decision Tree is naturally explainable from its internal structure. That is, when one travels down the tree in order to make a prediction, one can look at the nodes and their thresholds and build an explanation based on the features traversed and their values. However, this may be difficult for a Human user when trees are very complex, either in terms of depth or number of features. Moreover, there is much more information that can be generated, namely of statistical nature, that is implicit during the training of the tree and that may be useful for supporting Human decision processes. This section describes all the elements that are generated by the explainable model, and how they can be used to generate explanations.

Explainable elements are generated as the tree is being built, i.e., whenever a new split is created, and are stored in the tree's internal structure, in the form of json objects (one for each node or leaf). These objects can then be accessed, when traversing the tree for making predictions, and used to build explanations.

The following listing details the explainable elements that are stored in each node or leaf:

- The name of the feature on which the split was made, the threshold (value), and the type of condition (e.g. $>$, $>$ or $==$) (e.g. $X > 5$). This element is not generated for leafs;
- The prediction $\widehat{y}$ based on the split (either the average or the most frequent value, depending on the type problem), although this is generally not used since the prediction is given by the predictor model;
- Measures of confidence, based on the dispersion/purity of the split (e.g. variance, standard deviation, Gini index): the lower the dispersion or the higher the purity, the higher the confidence on the decision is;
- A measure of support, based on the number or percentage of instance in the split;
- An index of all the instances in the split;

These elements can then be used to build different explanations. The simplest one is to build a string, in natural language, that explains a given decision based on the features and their values. For instance, "The prediction is y because X is greater than a and Z is equal to $'bcd'$". This string is built by traversing the tree, accessing the features, conditions and values in each node, and concatenating them to build the string in an appropriate format and language. A pagination mechanism is implemented so that the user can control the depth of the explanation. For instance, the initial explanation may be based on the first 3 levels of the tree, and afterwards the user may request additional levels until the end of the explanation is reached (the leaf in the tree).

This also provides the user with a sense of feature relevance: the features that are mentioned first are relatively more important than the ones that come next. Moreover, a feature that is not used in an explanation tells the user that that feature is not significant for the decision-making process.

Every prediction also comes with measures of confidence and support. That is, the user is always provided with the number of instances (and its percentage) on which the explanation is being based, and the dispersion and purity metrics of the last node used for building the explanation.

In some of the levels, these strings may come with a warning that states something like: "But the prediction would be z if the value of X changed by a", in which a may be a positive or negative number, or a label (in which case the text reads "... if the value of X was a"). This happens when the value for a given feature is very close to the threshold value and the prediction would change if that threshold value was crossed. Essentially, this aims to provide the user with a metric for risk assessment, in the sense that a small variation in a given feature would significantly change the prediction, so a decision taken under these conditions may be more risky.

These warning messages, that are intertwined in the explanation, are also one of the ways for the user to perform counterfactual analyses in the sense that they provide the user with a notion of "what would happen" if one of the facts changed. The user can also explicitly do a counterfactual analysis by asking the chatbot questions such as "What would happen if the value of X was a?". In these cases, the tree will be traversed using the case provided by the user (which

is a modified or entirely new instance of data) and provide an explanation for that scenario (along with a prediction by the predictor model). Thus, explanations become interactive and the user can create or simulate "what-if" scenarios and gain a notion of how the predictions would change if the data changed.

Finally, another useful explanation is built by using the index of instances in the stopping node. Essentially, this allows to present the user with a list of past similar cases or instances. That is, the user can look at past similar cases and analyze some of them to have a better notion of what happened in the past. In an Audit scenario, this means that the user is looking into past cases audited by himself or other Auditors, into its characteristics and the outcome, and that can be used as a form of precedent, to help justify his current decisions. Instances can be sorted by different criteria, including date and similarity to the current instance being predicted. Similarity is calculated based on a weighted sum of differences, given by the euclidean distance for numerical variables and by the cosine similarity for the vector of nominal data (if any).

All these explanations are dependent on the pagination mechanism, that is, how deep the user wants to drill in the explanation. As the user moves down in the explanations, splits become smaller (lower support) but confidence increases. It is up to the user to decide how far down to travel: an early stop may lead to a more general explanation (with a tendency to high support and low confidence), while going further down will lead to low support but high confidence.

Building the Explainable Model The process of creating an instance of the explainable model is dependent on whether or not there is access to the data that were originally used to train the predictive model. If the original dataset is public and accessible, the process followed is the one detailed in Figure 2.

In this case this is a very straightforward process in which our modified version of the CART algorithm is trained on the dataset. This results in the Decision Tree that is used as the explainable model. The use then gets the predictions from the predictor model (eventually through the explanation interface), and the explanations from the explainable model, through the explanation interface.

If, on the other hand, the original dataset is not available, for instance due to proprietary or privacy issues, the process detailed in Figure 3 is followed. In this case there are some preliminary steps. First, the input to the process is no longer the original dataset but meta-data describing the original dataset. This meta-data includes the names of the features, their type, and their domains.

Based on this meta-data we create random instances of data. These instances are then submitted to the predictive model, for classification. The output provided by the model (predictions) is then added to the random input, thus constituting a Synthetic Dataset. The idea is that this dataset resembles, to the extent possible, the original dataset that was used to train the predictor model. This synthetic dataset is then used to train the explainable model, and the rest of the process is the same as the previous one.

In any case, whether there is access to the original data or not, this process produces a Decision Tree model that can be used to explain the predictions

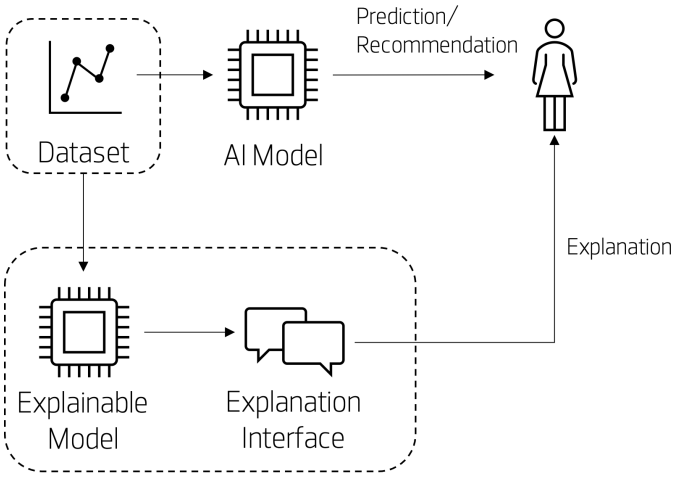

Fig. 2. Process followed for training an instance of the explicable model when the original dataset is accessible.

of any another model, independently of its internal structure. The explainable interface then interacts with this model to provide the user with Human-friendly interactive explanations.

3.2 Explainable Interface

The Explainable Interface was implemented in the form of a chatbot, so that the user can interact with it in a natural manner, using her/his own language. Specifically, the chatbot was implemented in Rasa, an open source machine learning framework for automated text and voice-based conversations.

There are two main components in the chatbot. The first is the NLU (Natural Language Understanding model) and the second is the conversational interface itself, that interprets the current context of a conversation and decides the next action in the conversation.

The main goal of the NLU model is to extract structured information from user messages, which are written in natural language. This includes generally two main aspects: user intent (i.e. what is the user's goal with a given message) and entity extraction (i.e. specific words that have meaning in a given conversation). Based on these two elements, the chatbot will then retrieve a response in order to continue the conversation.

The NLU model is trained based on the NLU training data. Training data consists of sample sentences, similar to those that the user would do, categorized by the corresponding intent. For instance, the sentences "Until next time!" and "Thank you for your help." could be associated to an intent called "goodbye". Responses are defined in very much the same way, by providing samples of utterances that could be used in a given point in the conversation.

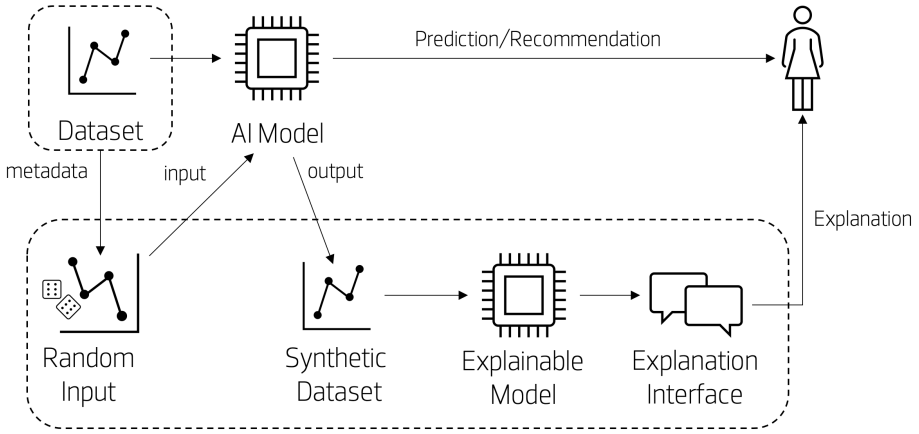

Fig. 3. Process followed for training an instance of the explicable model when the original dataset is not accessible.

Entities can be defined using regular expressions or, alternatively, by training a ML model. Entities are defined bearing in mind the concepts or information that the user would need to accomplish a given goal. For instance, in the utterance "Show me the 3 most similar cases", the chatbot would need to detect the intent "view past cases", and parse the number "3" as an entity, so that it could ask the API for the 3 cases most similar to the current one, to be shown to the user.

It is also possible to specify Stories. Stories are a type of training data as well, used to train the model that decides how the dialogue is message. That is, at a given point in the conversation, what should the next action be? How should the conversation develop? Stories are defined using sequences of intents and actions. Their main goal is to have conversational models that generalize better to unseen conversation paths.

Finally, Rules are similar to Stories in the sense that they are also constituted by intents and actions, but they are generally shorter and they are strict. That is, they will always follow the same path. Thus, conversations based on Rules are unable to generalize in the same way that those based on Stories do.

4 An application to Tax Fraud Detection

The approach described thus far in this paper is independent of the domain of application and of the predictive model used. Nevertheless, explanations always make reference to the domain by being mostly based on the features. In that sense, they are easily interpreted by Human experts.

We now describe a use-case of the proposed approach, that inspired it. This is an application in the domain of financial fraud detection, in the context of

the Neurat funded project (31/SI/2017 - 39900). The main goal of the Neurat project is to build a cooperative system in which Machine Learning models and Human experts work together to increase the efficiency of tax audit and fraud detection processes.

The project tackles two main challenges. First, it seeks to improve a previously existing rule-based audit tool. This rule-based tool has the disadvantage of producing many false positives, often very similar between them. Auditors must go through all of them, labeling them appropriately, which becomes very repetitive and time-consuming.

The main goal is thus to implement a ML-based approach, that can learn from the feedback and interactions of the Auditors. However, the use of Machine Learning, and in particular of supervised methods, requires vast amounts of labeled data. This is the second challenge that is being tackled by the Neurat project.

The problem is that data can only be labeled by Human experts (Auditors) and, in this case, it comes at a high cost: auditors must undergo extensive training and their time is very limited in face of all the instances they must audit. As a consequence, they are able to review only a small portion of the instances, usually by sampling, and thus contribute only with a small amount of labeled data.

To deal with these challenges, an Active Learning (AL) approach [21] is being followed to implement the Neurat project (Figure 4). Generally, AL approaches aim to make ML less expensive by reducing the need for labeled data. To achieve this, a so-called *Oracle*, which may be a Human expert or some automated artifact, is included in a cycle in which a ML model is improving over time by training and re-training on a growing pool of labeled data.

However, we introduce two major changes to the "traditional" AL process (Figure 1). First, we consider a pool of models rather than a single model [19]. New models are trained and added to the pool, which constitute a voting/averaging ensemble whose weights are continuously optimized by a Genetic Algorithm. Over time, models with a smaller weight are removed from the ensemble. This allows the system to converge while using relatively simple models, trained with partial data, instead of a very large and complex one, that would be very-hard to re-train. Moreover, we deal with multiple ML problems simultaneously. For instance, ML models are used to predict the risk of fraud, as well as to predict the value of user-defined features. That is, high-level or abstract features, that are not extracted from the raw data, but are instead provided by the Auditors. Since these cannot be derived from the raw data, they are predicted (and explained) by specifically trained models once there is enough data.

Secondly, we add another input to the Oracle, which in this case is the Human auditor. The auditor has access to the selected instance i, which is now accompanied by a prediction p and an explanation e. Now, when the auditor receives the instance to label (that is, when the auditor performs an audit action), he also receives the label proposed by the system as well as an intelligible expla-

nation for it, tailored for this specific domain. This applies to several features, including fraud risk and the previously-described user-defined features.

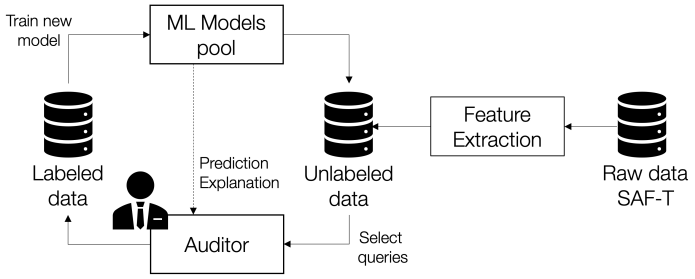

Fig. 4. High-level view of the Neurat system: the chatbot is placed between the model pool and the auditor.

The explainable interface that is now being implemented is placed between the model pool and the Auditor. That is, it will be used whenever the auditor requires a prediction from the system, which will now come with an explanation. The Auditor interacts with the audit system through an interface similar to that of Figure 5. The difference is that this UI is used for demonstration purposes as well as for validation, so it has additional debugging information and actions. In any case, the Auditor has a list of instances to audit, which he can analyze in detail by clicking in one, having access to more information. For each instance, the system provides suggestions regarding the risk of fraud and the values for the user-defined features, based on the ML models. The Auditor can then interact with the chatbot to request and refine explanations. The auditor does any changes deemed necessary and then saves them, marking the case as validated by a Human user. When this is done, all its data is moved to the labeled dataset and can be used as additional training data from now on.

We believe that the integration of this explanatory interface in applications such as this will contribute to increased transparency and trust in ML systems by Human users, especially due to its interactive nature and the use of a mostly symbolic approach based on domain-relevant features.

5 Conclusions

As AI-based applications gain increased relevance and control over our day-to-day living, so too must their level of responsibility and scrutiny grow. Legal or Ethical requirements mandate that automated decisions are transparent, interpretable, fair and trustworthy. However, this is often easier said than done.

Currently, one of the main issues stems from the use of the so-called black box models: models that are so complex or intricate that they become virtually

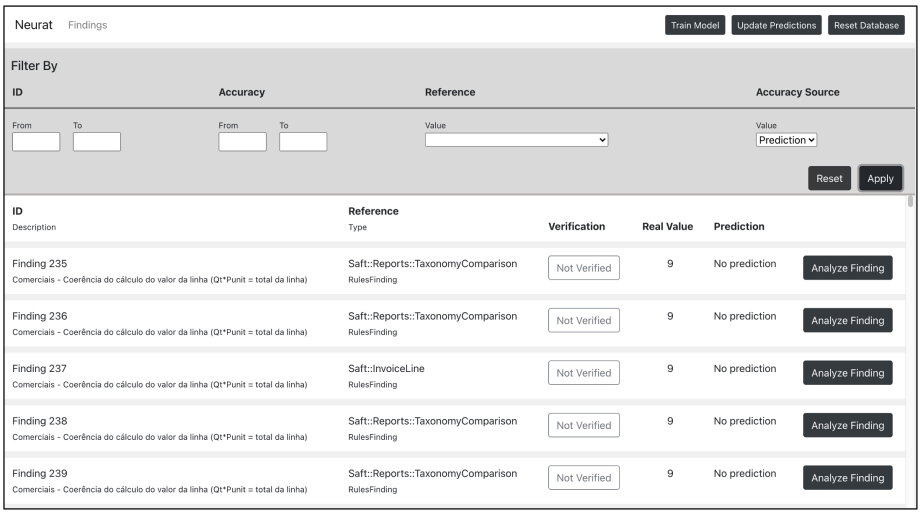

Fig. 5. Main UI used by the auditor to access the list of instances to audit, and provide feedback.

impossible to explain in a way that a Human can understand not only a particular decision but the behavior of the model itself. Deep Learning models are probably the best examples of this.

The main problem when we do not understand how a model behaves, is that it becomes unpredictable. For instance, we cannot fathom what the behavior of the model will be when exposed to unseen instances of data. This is particularly worrying in critical domains, in which the lives of Human beings depend on the decisions of models.

The need to devise more Human-friendly models is thus evident. Models that are naturally easier to explain, such as Decision Trees, do exist. However, they generally tend to have lower accuracy. There is thus a trade-off between how good a model is at making predictions, and how good it is at explaining them.

In this paper we proposed an approach for trying to have the best of two worlds. On the one hand, we still use the accurate original model for making predictions. On the other hand, we use a secondary explainable model that is responsible for generating several explainable elements, that can then be used to construct a wide range of Human-readable explanations. Moreover, this system can be used even if there is no access to the original dataset, either due to privacy concerns, to the data being proprietary, or to any other reason.

Finally, we use a conversational interface to provide access to the explanations, so that the user can actually interact with the model, ask it questions, get explanations, refine those explanations at will by drilling up or down, simulate scenarios for counterfactual and what-if analyses, among others.

All in all, we believe that this kind of systems may significantly improve the transparency of ML applications, and consequently the trust that Human decision-makers place on them.

6 Acknowledgments

This work was supported by the Northern Regional Operational Program, Portugal 2020 and European Union, trough European Regional Development Fund (ERDF) in the scope of project number 39900 - 31/SI/2017, and by FCT - Fundação para a Ciência e a Tecnologia, through project UIDB/04728/2020.

References

1. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media & society **20**(3), 973–989 (2018)
2. Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020)
4. Atkinson, K., Bench-Capon, T., Bollegala, D.: Explanation in ai and law: Past, present and future. Artificial Intelligence p. 103387 (2020)
5. Bertino, E., Kundu, A., Sura, Z.: Data transparency with blockchain and ai ethics. Journal of Data and Information Quality (JDIQ) **11**(4), 1–8 (2019)
6. Blacklaws, C.: Algorithms: transparency and accountability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170351 (2018)
7. Brkan, M.: Ai-supported decision-making under the general data protection regulation. In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. pp. 3–8 (2017)
8. Buiten, M.C.: Towards intelligent regulation of artificial intelligence. European Journal of Risk Regulation **10**(1), 41–59 (2019)
9. Campolo, A., Crawford, K.: Enchanted determinism: Power without responsibility in artificial intelligence. Engaging Science, Technology, and Society **6**, 1–19 (2020)
10. Comission, E.: Ethics Guidelines for Thrustworthy AI. Tech. rep., European Comission (04 2019)
11. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)
12. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (2020)
13. Hong, J.W., Choi, S., Williams, D.: Sexist ai: An experiment integrating casa and elm. International Journal of Human–Computer Interaction **36**(20), 1928–1941 (2020)
14. Lerman, R.I., Yitzhaki, S.: A note on the calculation and interpretation of the gini index. Economics Letters **15**(3-4), 363–368 (1984)

15. Lloyd, E.P., Hugenberg, K.: Beyond bias: response bias and interpersonal (in) sensitivity as a contributors to race disparities. European Review of Social Psychology pp. 1–34 (2021)
16. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report **49**(1), 15–21 (2019)
17. Molnar, C.: Interpretable machine learning. Lulu. com (2019)
18. Pendyala, V.S., Figueira, S.: Automated medical diagnosis from clinical data. In: 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). pp. 185–190. IEEE (2017)
19. Ramos, D., Carneiro, D., Novais, P.: evorf: An evolutionary approach to random forests. In: International Symposium on Intelligent and Distributed Computing. pp. 102–107. Springer (2019)
20. Reed, C.: How should we regulate artificial intelligence? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170360 (2018)
21. Settles, B.: From theories to queries: Active learning in practice. In: Active Learning and Experimental Design workshop In conjunction with AISTATS 2010. pp. 1–18 (2011)
22. Singh, S., Gupta, P.: Comparative study id3, cart and c4. 5 decision tree algorithm: a survey. International Journal of Advanced Information Science and Technology (IJAIST) **27**(27), 97–103 (2014)
23. Soares, E., Angelov, P.: Fair-by-design explainable models for prediction of recidivism. arXiv preprint arXiv:1910.02043 (2019)
24. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
25. Zhao, Z., Xu, S., Kang, B.H., Kabir, M.M.J., Liu, Y., Wasinger, R.: Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Systems with Applications **42**(7), 3508–3516 (2015)
26. Zuiderveen Borgesius, F.J.: Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights **24**(10), 1572–1593 (2020)