WHITEPAPER

How Conversational XAI Makes AI More Responsible

Design and Implementation of an Explainable AI chatbot

Summary

The implementation of AI in critical decision-making processes is rapidly increasing. Explaining model outcomes in an understandable way to every stakeholder is important, but a difficult operation.

Every stakeholder has unique information needs and expertise. An explanation that is understandable and useful to one stakeholder, can be irrelevant to a second. A solution with a lot of potential for this issue is Conversational XAI. Through natural language dialogues, users can easily obtain the exact explanation they need, by “asking the model” why they made a particular decision, the same way they might question a colleague.

The application of Conversational XAI in the form of a chatbot as described in this whitepaper, yields the following benefits:

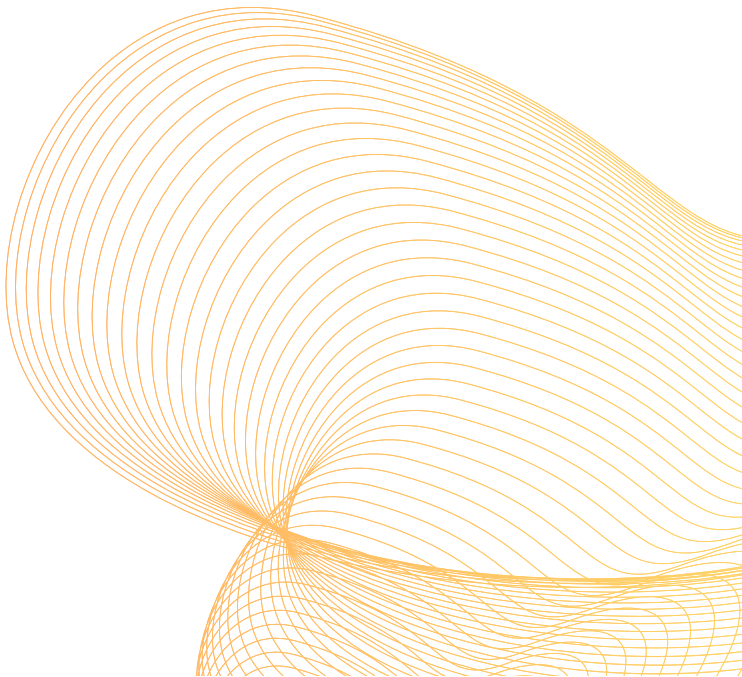
- Personalize explanations tailored to specific needs and only display the information you really need
- Interact through human dialogue that feels natural, even to those without a technical background
- Practice effective human oversight by allowing stakeholders to make well-informed evaluations

Meet Our Innovators

Nilay Aishwarya
Conversational XAI expert
naishwarya@deeploy.ml

Tim Kleinloog
CTO
tkleinloog@deeploy.ml

Table of Contents



Introduction

AI is rapidly evolving and becomes an integral part of our life. AI is already used in a wide range of critical decision-making processes across several domains. However, with the tremendous growth also comes the necessity to ensure that AI is being used responsibly. As a result, there is a rising interest in human-machine interaction, as can be seen by the number of articles being published about this topic. The upcoming EU AI Act is bolstering this development, enforcing effective human oversight for high-risk AI.

These factors have led to the development of multiple explainability techniques to translate machine model logic into something stakeholders can understand. However, stakeholders are from diverse backgrounds and may struggle to leverage explanations as it is hard to create a single explanation that fits all needs and requirements of the stakeholders.

This concern also hinders the practical implementation of human oversight in the form of feedback loops: where humans evaluate model decisions. By keeping a human in the loop, you can double-check model outcomes before using them. This also helps with the question about accountability of generated outcomes: there is always a human responsible for the decisions made in AI-aided systems. Additionally, feedback loops minimize the risk of losing control by helping keep track of model drift, and humans can steer algorithms in the right direction, by feeding them with feedback. However, the value of feedback loops is limited when stakeholders are unable to understand how a model came to a certain decision.

Why Conversational XAI?

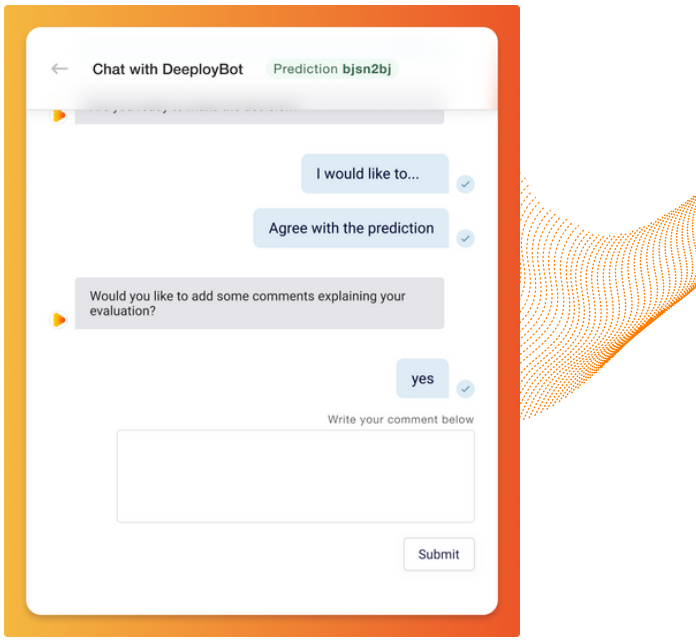
Decision-making is a diverse process that may vary from domain to domain, team to team, and person to person. Understanding the decision maker in their thought process, knowledgeability and background will ultimately help to deliver the information that is required / will ultimately support the decision-making process. However, across different domains, there is one common fact: the process of decision-making involves discussions between humans with mutual information exchange. Conversational XAI aims to facilitate mutual information exchange between humans and models, mimicking the human decision-making process.

We developed a conversational XAI chatbot in collaboration with Tim Kleinloog (Deeploy) Nilay Aishwarya, and Ujwal Gidarajuwe (TU Delft) to address the aforementioned concerns by providing stake-

holders the capability to obtain diverse explanations through a human-like dialogue system. Thus boosting stakeholders' capabilities to access different explanations and improving feedback loops.

Conversational XAI Chatbot

Our chatbot allows stakeholders to interact with an AI model through a natural language-based interface. After making a prediction, users can ask pre-defined questions about the prediction, such as "What would happen if you give the model different input?". The chatbot produces natural language responses to these questions. The chatbot boosts and stimulates stakeholders to check out different explanation approaches, ensuring they get several angles of the available explainability techniques, resulting in a better understanding of the prediction.



Once a user has a sufficient understanding of how the model came to a prediction, they can agree or disagree with the prediction directly through the chatbot. Furthermore, the users can describe their evaluation of the decision, and highlight key features that helped them come to their verdict. The chatbot gathers and saves such insights, aiding ML engineers in improving their models and explanations.

How does it work in practice? Meet Jerry and how he uses the chatbot in his work

Let's take a look at the life of Jerry, a Data Scientist at a major bank. Jerry is responsible for a machine learning model that predicts whether a loan application is creditworthy or not. In this case, the machine learning model suggests that the applicant is not creditworthy and the final decision has to be made by Jerry whether to agree or disagree with the prediction. Jerry opens our chatbot to investigate important factors (based on the SHAP score) that are leading to the current prediction outcome. In the conversation, The chatbot suggests that Jerry looks at the visual explanation to get an idea of what influenced the prediction.

Chat with DeeployBot Prediction bjsn2bj-snjsjb3-alkm-neb23

Minimum change needed to switch prediction?

In order to switch model prediction following minimum changes can be done to the input: Capital Gain was 2296.875

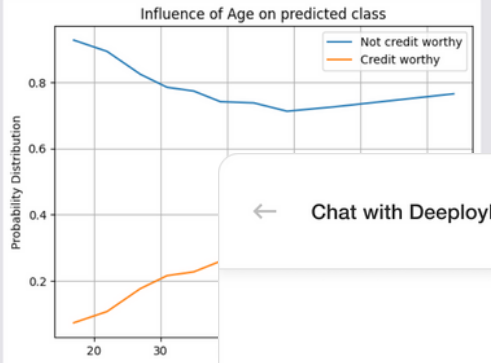
Do you want to see what is the correlation of a feature of your interest with the two possible predictions for original input?

Natural language explanation

Yes, for factor Age

Follow-up question

Here you go! This is how the prediction probabilities change for different output prediction classes as Age varies.



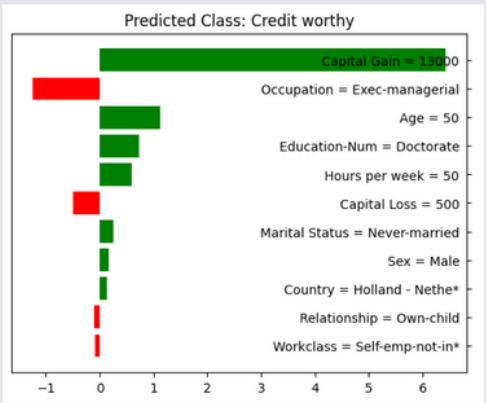
Chat with DeeployBot Prediction bjsn2bj-snjsjb3-alkm-neb23

What are the most important factors for the prediction?

Here are the five most important factors(features) that the model uses to make this prediction: Capital Gain, Occupation, Age, Education-Num, Hours per week

The relative importance of all the features can be seen in the figure below (based on SHAP) where higher magnitude means higher importance.

Visual explanation



Do you want to see what is the correlation of a feature of your interest with the two possible predictions?

Interactive

Yes No

Additionally, The chatbot also allows and suggests to Jerry the minimum changes needed in the loan application for it to be found creditworthy.

Jerry's colleague, Rudy, is a Compliance Officer at the bank and wishes also to understand why certain predictions have been made by the ML model. However, Rudy is not a Data Scientist and cannot understand technical machine learning metrics. The chatbot also addresses these concerns by having the capability to personalize its information output for the Compliance Officer to understand.

How Conversational XAI supports AI-aided decision making

Our XAI chatbot helps promote the responsible use of AI for decision-making in high-risk industries, by making it easier for stakeholders to interpret model outcomes, specifically:

Get personalized explanations tailored to your specific needs, by requesting only the information you really need

Interact through human dialogue that feels natural, even to those without a technical background

Practice effective human oversight by allowing stakeholders to make well-informed evaluations

Developing cutting-edge technology

We involved users in every step of the development of the chatbot by following a Design Thinking process and using methodologies such as user interviews and extensive prototyping.

Empathize

Interviewed stakeholders to understand challenges associated with XAI

Define Problem

Current XAI solutions lack the dialogue of human on human explanations

Ideate

Brainstormed solutions and decided on the idea of a chatbot

Prototype & Test Idea

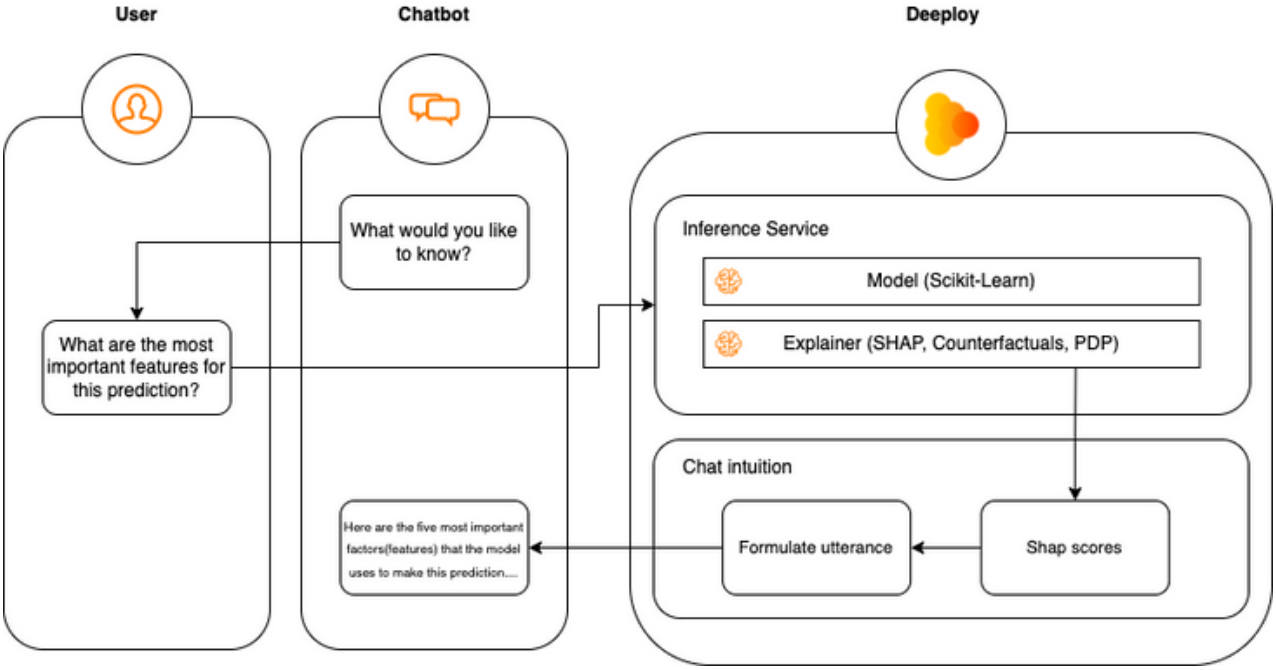
Explore conversational paths and used Wizard-of-Oz prototyping with positive results

Prototype & Test MVP

Created chatbot MVP running on Deeploy and performed successful user testing

A Scalable Solution

The chatbot is built on a flexible, modular foundation to tackle future developments and innovation. Upon the user asking a question, the chatbot interprets user intent and then communicates with the Deeploy infrastructure and inference services. The chatbot combines various standard explanation techniques, such as SHAP, counterfactuals, and Partial Dependency Plots. The explanation instances are managed on Deeploy's platform. The chatbot generates an utterance using the output of the explanation technique and presents it to the user as an answer to their question.



Statistical Deepdive

The Partial Dependency Plot

The idea of a partial dependency plot is to get the average effect of feature values on the outcome prediction by marginalizing all other features. Hence this explainer uses the entire dataset as a reference, providing a global explanation of the model's dependency on features. This is done in the following manner:

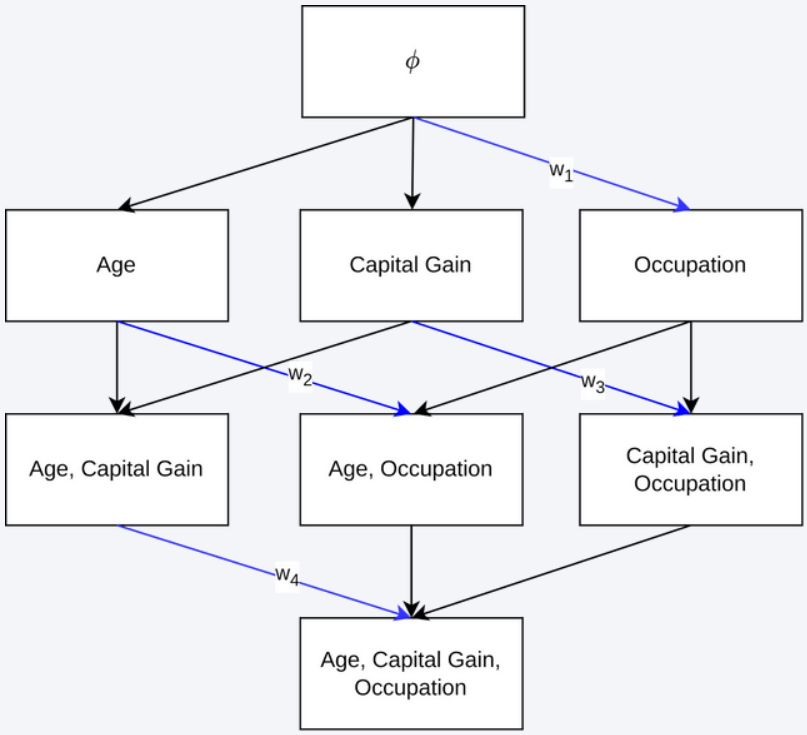
1. Train a model on the training dataset.
2. Choose a feature of interest: e.g. "Occupation".
3. Replicate the dataset excluding the feature of choice (Occupation here) as many times as a distinct value of Occupation exists in the dataset. So for example, if we have *n* distinct Occupation values in a dataset containing *m* entries then the total replicated dataset would be *n* x *m*.
4. Get predictions on this new dataset.
5. Average out the prediction probabilities for each of the *n* distinct values and plot them for every *n* value of Occupation.

SHAP Principles

Shapley value[1] (SHAP) principles from Game Theory are used to obtain the impact of features on the predicted outcome. For simulating the game and obtaining relevant feature importance it defines the following:

1. Game: Reproducing the predicted value of the model
2. Players: The players are the features of the model.

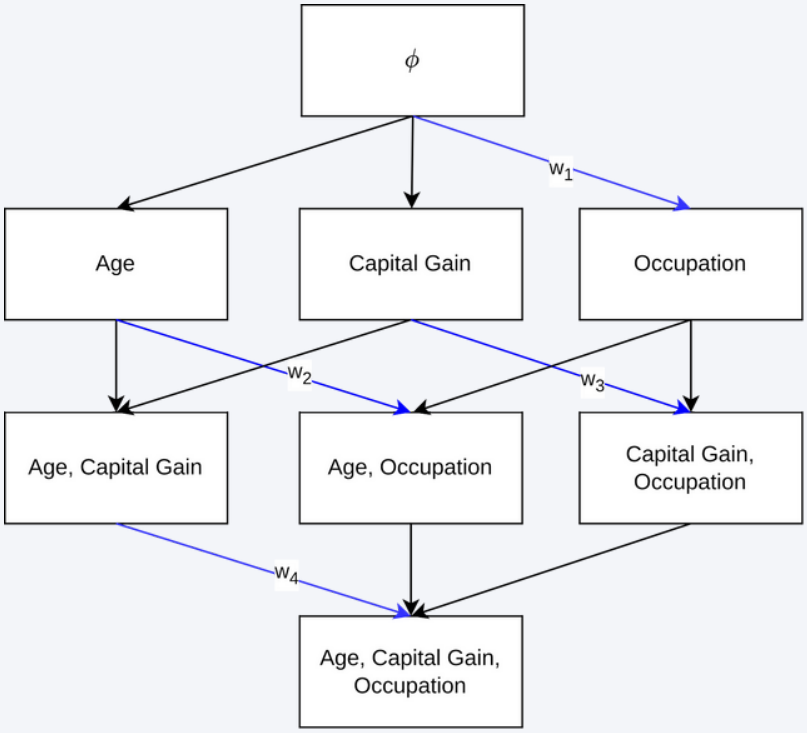
We quantify the contribution of every "player" in the "game" i.e. the feature on the predicted outcome value. A point to remember is that this is a local explanation i.e. for current input. The idea is that each possible combination of players should be used to understand the importance of a single player.
For example let's take a simple linear regression model which predicts the Credit value of an applicant using 3 features: Age, Capital Gain, and Occupation.



The powerset of the given features like in this example it would be *2^n = 2^3 = 8* possible combinations. SHAP trains a distinct identical predictive model for each combination in the power set and the difference between these models is the number of features the model uses.

[1] SHAP: https://arxiv.org/abs/1705.07874

Looking at the figure above here each of these nodes are model, and every node down the levels is different from other nodes at the same level by a feature. Now each of these edges brings a marginal contribution of a feature to the model. For example, the null set model uses just training observations to make predictions without any features. When we add a feature to the model (Occupation) we want to see how much the predicted value changed upon the addition of the feature. For the overall effect of Occupation, we need to see its contribution in all models where Occupation is present.



SHAP(Occupation)[1] = w1* Model(Occupation) + w2* Model(Occupation, Age) + w3* Model(Occupation, Capital Gain) + w4* Model(Occupation, Age, Capital Gain) where w1 + w2 + w3 + w4 = 1

Two assumptions are made:-

1. The sum of the weights of all the marginal contributions to 1-feature-models should equal the sum of the weights of all the marginal contributions to 2-feature-models equal to the sum of the weights of 3-feature-models. I.e. w1 = w2 + w3 = w4
2. All the weights of marginal contributions to same level feature-models should be equal to each other. I.e. in this case w2 = w3

[1] SHAP: https://arxiv.org/abs/1705.07874

Going forward similar process for all features. Once all SHAP[1] values are obtained for every feature the sum of SHAP values yields the difference between the prediction of the model and the null model. Thus allowing us to interpret the contribution of each feature to the outcome prediction.

Counterfactual Explainer Deep Dive

A counterfactual[2] explanation is an instance of the input data that, if it were true, would cause the model's output to change to a desired value. The counterfactual explainer works by finding the smallest possible change to the input data that results in a different output from the model. It does this by minimizing a loss function that penalizes the distance between the original input and the counterfactual explanation.

[1] SHAP: https://arxiv.org/abs/1705.07874
[2] Counterfactual: https://arxiv.org/abs/2205.15540

What's Next:

The current implementation is just the starting point for Deeploy, as we continuously innovate with and for our customers. Our chatbot lays the foundation for future projects in the space of human-centric explainability with:

Advanced Tailoring

Customized utterances and visual explanations tailored to different types of stakeholders

Natural Conversation

Improved utterances and enhanced natural flow of conversation with more complex Large Language Models

More frameworks

Increased number of supp-orted XAI frameworks to account for a wide variety of model types

TALK TO US:

(0031) 6 344 711 54
marketing@deeploy.ml