Explainability for machine learning models : from data adaptability to user perception

Julien Delaunay

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *INFO*

Par
Julien DELAUNAY

Explainability for Machine Learning Models : From Data Adaptability to User Perception

Strategies for Faithful and Understandable Explanations

Rapporteurs avant soutenance :

Marie-Jeanne LESOT	Professor, Univ. Sorbonne LIP6, France
Andrea PASSERINI	Associate Professor, Trento University, Italia

Composition du Jury :

Présidente :	Elisa FROMONT	Professor, Univ. Rennes, France
Examinateurs :	Pierre MARQUIS	Professor, Univ. Artois, France
	Katrien VERBERT	Professor, KU Leuven, Belgium
	Niels VAN BERKEL	Associate Professor, Aalborg University, Danemark
Dir. de thèse :	Christine LARGOUËT	Associate Professor, Institut Agro, Rennes, France
Co-dir. de thèse :	Luis GALARRAGA	Researcher INRIA/IRISA, Rennes, France

TABLE OF CONTENTS

Introduction **9**

Context and Motivation . . . 10

Lack of Transparency . . . 11

Explainable Artificial Intelligence . . . 12

Research Questions . . . 13

Publications . . . 15

1 Foundations of Explainability **17**

1.1 Explainable AI . . . 18

1.1.1 Self-Explainable vs Post-Hoc Explanations . . . 19

1.1.2 Global vs Local Explanations . . . 20

1.1.3 Model Dependent vs Model Agnostic . . . 21

1.2 Explanation Paradigms . . . 23

1.2.1 Notation . . . 23

1.2.2 Rule-based Explanations . . . 24

1.2.3 Feature-Attribution . . . 26

1.2.4 Example-based Explanations . . . 29

1.3 Evaluating Explanations Techniques . . . 31

1.3.1 Surrogate-Based Evaluation Criteria . . . 32

1.3.2 Instance-Based Criteria . . . 34

1.4 Outline of this Thesis . . . 36

I Good Explanation From Data Perspective **39**

2 Improving Anchor-based Explanations **40**

2.1 Context . . . 41

2.2 Anchors . . . 42
2.3 Impact of discretization on Tabular Data . . . 43
2.3.1 New Discretization Methods . . . 44
2.3.2 Experimental Evaluation . . . 44
2.3.3 Results . . . 45
2.4 Improving Anchors on Text . . . 46
2.4.1 Neighborhood Generation Strategies . . . 47
2.4.2 Pertinent Negatives . . . 48
2.4.3 Experimental Evaluation . . . 48
2.4.4 Results . . . 49
2.5 Conclusion . . . 50

3 When Should We Use Linear Explanations? 53
3.1 Context . . . 54
3.2 Preliminaries . . . 56
3.2.1 Problem Statement . . . 56
3.2.2 Linear Explanations . . . 57
3.2.3 Adherence and Fidelity . . . 57
3.2.4 Existing Methods . . . 58
3.3 Counterfactual Explanation Methods . . . 58
3.3.1 Counterfactuals for Tabular Data . . . 58
3.3.2 Counterfactual Techniques for Textual Data . . . 61
3.4 Adapted Post-hoc Explanations . . . 65
3.4.1 APE Oracle . . . 68
3.4.2 Linear Explanations . . . 70
3.4.3 Rule-based Explanations . . . 70
3.4.4 Illustrative Example . . . 71
3.5 Experiments . . . 73
3.5.1 Experimental Setup . . . 74
3.5.2 APE Oracle Evaluation . . . 76
3.5.3 Comparison with other Explanation Methods . . . 82
3.5.4 Ablation Study . . . 84
3.5.5 Summary and Discussion of Linear Suitability Results . . . 86
3.5.6 Counterfactuals Evaluation . . . 86

3.6 Discussion and Conclusion . . . 88

4 Explaining a Black Box Without Another Black Box 93
4.1 Context . . . 93
4.2 Counterfactual Techniques for Textual Data . . . 96
4.3 Comparative Study . . . 98
4.3.1 Complexity Spectrum . . . 98
4.3.2 Experimental Information . . . 100
4.4 Results . . . 102
4.4.1 Counterfactual Quality . . . 103
4.4.2 Method Quality . . . 105
4.5 Conclusion . . . 108

II Explanations Tailored to the User 111

5 User-Centered Evaluation of Explainability Methods 112
5.1 Evaluating Explanations with Users . . . 113
5.1.1 Evaluating Explainable AI Systems with Users . . . 113
5.1.2 Guidelines and Metrics to Conduct User Studies . . . 115
5.2 Method . . . 116
5.2.1 Methodological Framework . . . 116
5.2.2 Scales & Metrics . . . 118
5.3 Conclusion . . . 120

6 Explanation Techniques and Representations on Users 121
6.1 Explanation Techniques and Representations . . . 123
6.1.1 Datasets & AI models . . . 123
6.1.2 A Common Representation for Explanations . . . 124
6.2 Method . . . 127
6.2.1 Task . . . 128
6.2.2 Scales & Metrics (Illustration for One User) . . . 129
6.2.3 Participants . . . 132
6.3 Results . . . 134
6.3.1 Understanding . . . 134
6.3.2 Trust . . . 137

6.3.3 Additional Measurements . . . 139
6.3.4 Perception vs. Behavior . . . 140
6.3.5 Open Questions . . . 141
6.4 Discussion . . . 142
6.4.1 Impact of Explanation Technique . . . 143
6.4.2 Impact of Representation . . . 144
6.4.3 Limitations & Future Work . . . 145
6.5 Conclusion . . . 146

Conclusion and Perspectives **149**
Thesis Overview and Objectives . . . 149
Key Insights and Lessons . . . 151
Open Challenges and Opportunities for Future Research . . . 153
Explanations Adapted to Data . . . 153
Explanations Tailored to Users . . . 154
Envisioning the Future of Explainable AI . . . 156
Conclusion . . . 158

Bibliography **159**

A When Should We User Linear Explanations? **177**
A.1 Datasets and Classifiers . . . 177
A.2 LS vs. LS_{APE} . . . 178
A.3 Additional Linear Suitability Evaluation . . . 180
A.4 Growing Fields vs. Growing Spheres . . . 182

B Explanation Techniques and Representations on Users **183**
B.1 Code and Data Processing . . . 183
B.2 Questionnaire . . . 184
B.2.1 Satisfaction Scale . . . 184
B.2.2 Trust Scale . . . 185
B.2.3 Understanding Scale . . . 185
B.2.4 Question to verify user's validity . . . 186
B.2.5 Explanation Paragraph in Example Round . . . 187

Résumé en Français **195**
Context . . . 195
Besoin de transparence . . . 196
Expliquer les modèles d'apprentissage automatique . . . 197
Comprendre les différents types d'explications . . . 198
Les trois types d'explications courants . . . 201
Génération d'explications d'un point de vue des données . . . 203
Amélioration des explications basées sur Anchor . . . 203
Quand devrions-nous utiliser des explications linéaires ? . . . 205
Exploration du spectre des méthodes d'explication contrefactuel . . . 206
Conclusion . . . 206
Les perspectives utilisateurs dans la génération d'explications . . . 207
Évaluation des méthodes d'explications centrée sur l'utilisateur . . . 207
Impact des techniques et représentations sur les utilisateurs en XAI . . . 208
Conclusion . . . 208

INTRODUCTION

Contents

Context and Motivation 10
Lack of Transparency 11
Explainable Artificial Intelligence 12
Research Questions 13
Publications 15

In recent decades, the rapid advancement of artificial intelligence (AI), and particularly of machine learning (ML) models, has significantly impacted our daily lives. This remarkable progress can be attributed to the exponential growth in the availability of data and the enhanced accuracy of these models. As a result, AI and ML models have become capable of remarkable achievements such as providing medical diagnoses, generating coherent texts, and efficiently identifying environmental issues. These advancements have transformed numerous industries and have the potential to further revolutionize our society.

However, this progress has also led to an increase in complexity, which has turned ML models into black boxes. Their opaque nature makes it challenging to inspect their reasoning, conduct audits, or gain insights from them. The question then arises: Can we rely on these models in critical situations, even when we are unaware of their limitations and potential failures? In scenarios like predicting personal preferences for entertainment such as Spotify or Netflix, the consequences of model inaccuracies may be minor. But in cases like predicting natural disasters or making crucial decisions in areas such as medicine, job offers or justice, understanding the model's reliability and reasoning becomes paramount. Indeed, a lack of trust or misunderstanding in a model may lead to erroneous decisions. Moreover, these models have demonstrated vulnerabilities in the form of biases against minorities and adversarial attacks invisible to human eyes.

To address these issues, public discussions have brought these biases and drawbacks to light. The deployment of ML models has resulted in reported problems, as observed in media coverage of incidents. For instance, Amazon initiated a project aimed at automating their company's hiring process using an ML algorithm. However, this algorithm was found to exhibit gender bias, leading to discrimination against women and raising concerns about fairness and equity [1]. Similarly, in the case of Compas, the system used for assessing American prisoner recidivism, it has been observed that black prisoners are more likely to be classified as at risk of recidivism [2]. A more recent controversy concerns the Dutch government's allegations of welfare fraud against numerous families, many of whom have dual nationalities or immigrant backgrounds [3]. These accusations have resulted in unjust penalties and severe financial difficulties for these families. Consequently, questions arise about whether we, as a society, can fully trust ML models based solely on metrics such as accuracy or precision. To tackle these concerns, laws like the GDPR and the AI Act have been introduced to regulate and guide the usability of ML models. These regulations emphasize aspects such as model robustness, accuracy, and transparency [43, 78].

In response to these challenges, the research community has recognized the critical importance of explaining AI models' decision-making processes. The field of eXplaining Artificial Intelligence (XAI) has experienced a significant surge as a means to provide explanations for model predictions [68]. In this thesis, we focus on generating explanations to identify the primary factors influencing an ML model's predictions. By clarifying the decision-making process, these explanations aim to enhance user trust and improve the reliability and accountability of ML models.

Context and Motivation

Complex machine learning or black box models often lack interpretability, making it difficult to understand their decision-making processes and potential biases. Moreover, the lack of transparency in these models has elicited criticism from various legal structures, which recognize the importance of accountability and trust in automated decision-making systems [43, 78]. To address these issues, XAI has emerged as a key area of research. By providing interpretable

1. https://aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against
2. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

explanations for model predictions, explainability techniques aim to enhance transparency and facilitate a better understanding of the underlying mechanisms of complex models. These efforts not only address legal concerns but also promote ethical and responsible AI practices. In the following, we will delve deeper into the specific challenges posed by complex machine learning models, explore the implications surrounding their lack of transparency, and provide an overview of how explainability techniques can help overcome these challenges.

Lack of Transparency

One key issue in complex machine learning models is their lack of interpretability, which stems from their intricate structure. Deep neural networks, for instance, often consist of numerous hidden layers, each transforming the input data in a non-linear manner, resulting in an enormous number of parameters [4]. Understanding the specific reasons behind the classification decisions made by these models can be elusive. For example, comprehending why GPT4 [113] may translate "a nurse" with the female word "une infirmière" in French and "a doctor" with the male word "un médecin" may require exposing the complex transformations and learned features within the network [51].

The black-box nature of these models further intensifies the interpretability challenge. In ensemble models, multiple algorithms are combined to make predictions, adding a supplementary layer of complexity to understand their decision-making process. Each algorithm within the ensemble contributes with its own logic, making it challenging to trace the specific factors that influenced the final prediction. Consequently, the inner workings of ensemble models remain obscure, making it difficult to explain their outputs transparently and intuitively. The question may even arise for decision trees or linear models that are considered simple and transparent models. For example, when the depth of the tree or the number of coefficients of the linear model is too large [91].

From the lack of interpretability and the black-box nature of these complex models significant challenges in various domains arise [129]. In fields where explainability is crucial, such as healthcare, finance, or legal settings, it becomes essential to bridge the gap between model predictions and human understanding. Efforts to unravel the black-box nature of these models not only improve transparency and accountability but also cultivate trust and adoption in critical applications [75]. By making complex models more interpretable, users can gain a

4. BERT base: 110 million parameters; GPT4: 1.76 trillion parameters

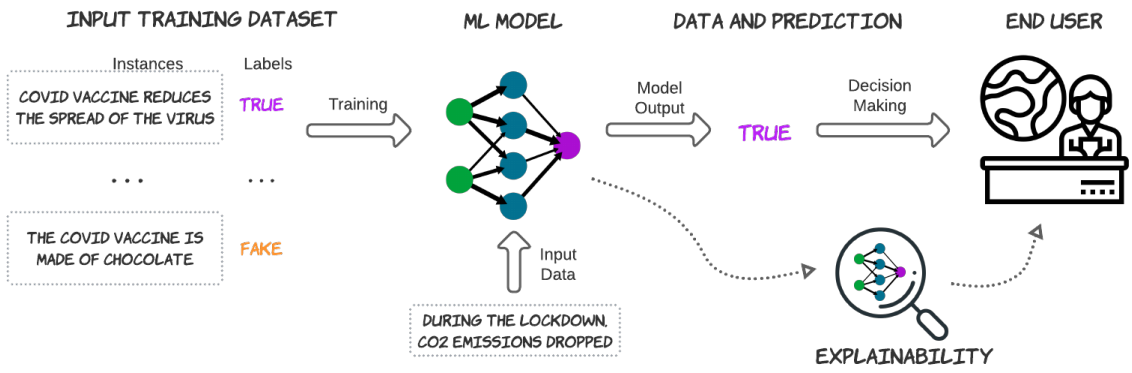


Figure 1 – Visual representation of a machine learning model designed to detect fake news within newspaper titles. This model is trained on diverse labeled examples to accurately classify news articles as either "fake" or "true". The aspect of explainability is added to the system to provide comprehensive and transparent insights into the model's predictions.

deeper understanding and confidence in the model's predictions and ultimately drive wider acceptance and responsible use of these powerful machine learning techniques [122].

Explainable Artificial Intelligence

The problem of generating transparent models has been a topic of research for many years. However, recent advancements in the field, in particular Deep Learning (DL) methods, have brought about a renewed focus on this area. This renewed interest arises from recognizing the limitations of complex models that have been widely applied without a comprehensive understanding. Researchers are actively exploring techniques and methodologies to enhance the interpretability of complex models [15, 68]. These approaches aim to shed light on the decision-making process of complex models, enabling users to gain insights into how specific inputs are transformed and influence the final prediction.

These interpretability methods can be applied to many domains, tasks, and data types [15]. To illustrate, consider the toy example in Figure 1, which portrays a model designed to detect fake news based on newspaper titles. This model is trained on existing newspaper headlines, each labeled as either fake or true. This training step enables the model to predict the label of a novel instance, such as a newspaper title in our example. The resulting prediction is then given to the end user, in this case, a journalist, who must make a decision based on this prediction. The explainability step of this process allows the journalist to investigate which elements from

the input article contribute to the model's decision. For example, an explainability method can indicate the specific words in the newspaper title that led the model to classify it as true. Alternatively, another explanation technique might identify sensitive words that if removed or replaced in a certain way, would have resulted in predicting "fake". These methods exhibit considerable diversity and can explain, for instance, why an image recognition model classified an object as a bird [25] revealing how specific parts of the object resemble features from known bird images.

However, relying solely on a single explanation method to investigate the significance of these elements comes with limitations. Various techniques may excel in explaining specific model types, may be adapted to different data formats, or may address specific instances, making it clear that no single method is universally effective across all scenarios. Consequently, it is mandatory to consider the data context, such as the model to explain or the type of instances before producing meaningful explanations. Moreover, providing identical explanations to diverse audiences and stakeholders can pose challenges and potential risks. For example, in recommendation systems, the expectations of the company developing the AI model may differ from those of consumers receiving recommendations. For example, an AI developer may be interested in debugging the model, whereas the consumer may require information to build trust in the system. Likewise, in healthcare systems, the questions and concerns of the company deploying the AI model may differ from those of doctors or hospitals utilizing the technology. Therefore, another important aspect of generating a good explanation for a model is to consider the person to whom the explanation is tailored.

Research Questions

At the time of the explainability boom, most of the research on generating explanations for AI models was conducted by machine learning researchers. Therefore, the first aspect tackled by the XAI community was to generate precise and reliable explanations for model prediction. In this context, global explainability's central goal is the understanding of the overall functioning of the model. However, in this thesis, our focus is the study of local explanations, which aim to explain the prediction of a model for a *target instance*. This instance may take various forms such as information about an individual or textual data as in Figure 1. The XAI community has developed numerous and various methods to explain the prediction of a model for a given instance [15, 68]. These methods span from simple decision rules [80, 124], linear models [52, 123], and showing similar inputs that convey different outputs from the model [82,

151]. A notable observation emerges: existing research has mostly focused on generating the best explanation techniques that should work on every instance, model, and user. However, as demonstrated in this thesis, the quest for such universal solutions may resemble a pursuit of the mythical El Dorado. No single technique can adapt to all data and user contexts.

Since there exists already a plethora of explanation techniques, in this thesis we aim to identify some limits of the existing explanation methods. As the quality of the explanation may be impacted by different aspects such as the kind of model to explain or the users' profile, this thesis is divided into two parts. Firstly, we focus on generating local explanations adapted to the data. Secondly, we study the impact of the chosen explanation techniques and representations on users. Therefore, our first research question is: **How to generate the best explanation from a data perspective?**. Conversely, while it is largely accepted that explanations should be tailored to the users receiving them, the users-centric aspect has been underrepresented in the literature [1, 4]. As such, the second part of our research seeks to answer: **How to generate the best explanation from a user perspective?**.

The research presented in this thesis focused on local explanations for supervised machine learning classification models trained on tabular and textual data. The manuscript is composed of seven chapters, including a preliminary of explainable AI and a conclusive summary. As the research included in this thesis addresses both of these research questions, we have chosen to structure the thesis into two parts, each focusing on one of these questions. The first part, devoted to the data perspective, initiates in Chapter 2 by studying the impact of an appropriate parametrization on the quality of explanations, specifically on rule-based explanations. Chapter 3 follows and proposes to adapt the explanation technique to the target instance. Finally, Chapter 4 studies the influence of the conversion space utilized for embedding input text before generating an explanation. In the second part, which concentrates on the user perspective, Chapter 5 introduces a methodological framework for the design of user studies that assess the impact of explanation techniques. Subsequently, Chapter 6 applied this framework to investigate the impact of different explanation techniques and their representations on users' trust and understanding. This two-part structure allows us to comprehensively explore the diverse facets of the explainability landscape and contribute with valuable insights to the field. These insights are supported by the publications produced during the Ph.D. and listed in the following section.

Publications

We conclude this section with a list of the articles published during my Ph.D. as well as those still under review:

Mentioned in this Thesis

- Published:
  - Improving Anchor-based Explanations. Julien Delaunay, Luis Galárraga, and Christine Largouët, in: Proc. International Conference on Information and Knowledge Management *CIKM*, 2020.
  - When Should We Use Linear Explanations? Julien Delaunay, Luis Galárraga, and Christine Largouët, in: Proc. International Conference on Information and Knowledge Management *CIKM,* 2022.
  - Adaptation of AI Explanations to Users' Roles. Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels van Berkel, in: *Conference on Human Factors in Computing Systems CHI, workshop on Human-Centered Explainable Artificial Intelligence HCXAI*, 2023.
- Under review:
  - Impact of the Explanations Techniques and Representations on Users' Trust and Understanding. Julien Delaunay, Luis Galárraga, Christine Largouët, and Niels van Berkel, *under review in: Conference on Human Factors in Computing Systems CHI*, 2024.
  - Explaining a Black Box without a Black Box. Julien Delaunay, Luis Galárraga, and Christine Largouët, *under review in: Conference of the North American Chapter of the Association for Computational Linguistics NAACL,* 2024.

The following publications, have been conducted through collaborations during my PhD, but are not mentioned in this manuscript.

Other Works (joint collaboration)

— s-LIME: Reconciling Locality and Fidelity in Linear Explanations. Romaric Gaudel, Luis Galárraga, Julien Delaunay, Laurence Rozé, and Vaishnavi Bhargava, in: *Proc. IDA,* 2022.

— On Moral Manifestations in Large Language Models. Joël Wester, Julien Delaunay, Sander De Jong, Niels van Berkel, in: *Proc. CHI Workshop on Moral Agents,* 2023

— "Honey, Tell Me What's Wrong", Explicabilité Globale des Modèles de TAL par la Génération Coopérative. Antoine Chaffin and Julien Delaunay, in: *Proc. Le Traitement Automatique des Languages Naturelles,* 2023.

CHAPTER 1

FOUNDATIONS OF EXPLAINABILITY

Contents

1.1 Explainable AI 18
1.1.1 Self-Explainable vs Post-Hoc Explanations 19
1.1.2 Global vs Local Explanations 20
1.1.3 Model Dependent vs Model Agnostic 21
1.2 Explanation Paradigms 23
1.2.1 Notation 23
1.2.2 Rule-based Explanations 24
1.2.3 Feature-Attribution 26
1.2.4 Example-based Explanations 29
1.3 Evaluating Explanations Techniques 31
1.3.1 Surrogate-Based Evaluation Criteria 32
1.3.2 Instance-Based Criteria 34
1.4 Outline of this Thesis 36

The precise definitions of many fundamental terms used in the explainable AI literature continue to vary among different authors, indicating a lack of consensus [91]. For instance, the terms interpretability and explainability are often used interchangeably since they have been

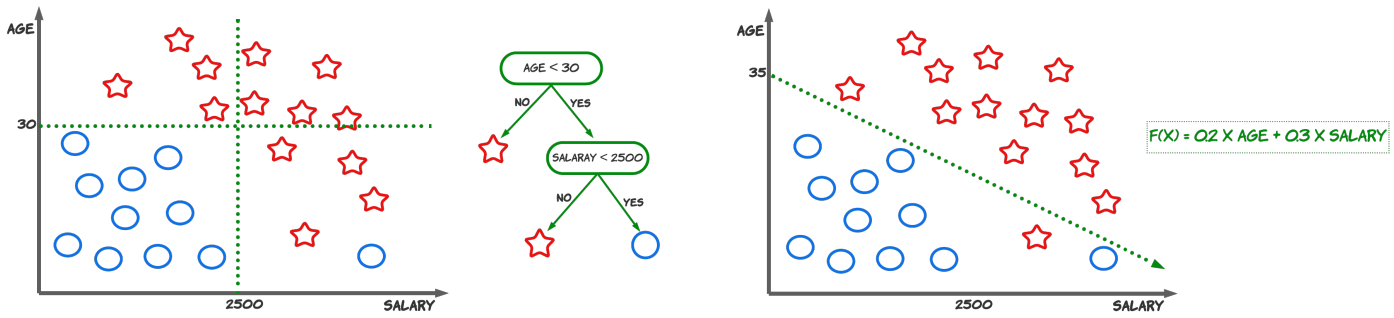
introduced in a close time gap. As the field of eXplainable AI (XAI) has grown, researchers have tried to unify the definitions of the terms employed. While some researchers have proposed or preferred to use interchangeably the terms interpretability, explainability, and transparency to name just a few of them, Lipton [91] proposed a definition of interpretability and how it differs from transparency. This definition has been well-accepted in the research community, amassing over 4000 citations in 2023. It suggests that a model should be considered transparent if a person can read it in its entirety – without necessarily understanding it. On the other hand, Lipton defined a model as interpretable if a human can understand it or can take input data together with the model parameters and calculate the model's prediction in a reasonable amount of time. Finally, in the realm of XAI, the concept of explainability is centered on the process of extracting information from the model and translating this information to the user.

In this thesis, the term "interpretability" refers to the capacity of a user to understand the inner workings of a machine learning model. In contrast, "explainability" is used to describe the methods that are put on top of a machine learning model to elucidate its functioning to users. Lastly, "transparency" characterizes methods through which the model's mechanisms are directly observable. Transparency indicates how exactly the model works by presenting details about the model's inner workings, parameters, etc.

In this chapter, we begin in Section 1.1, by categorizing machine learning explanation techniques and defining the taxonomy used in this thesis. Subsequently, in Section 1.2, we define formally some notations and terms to enhance the clarity of this manuscript. Additionally, we present the three families of explanation methods. Finally, Section 1.3 discusses methods for evaluating the performance of explanation techniques.

1.1 Explainable AI

This section delves into the realm of explanation methods for AI systems, specifically focusing on three fundamental dimensions defined by the community [6, 15, 39, 57]: self-explainable vs. post-hoc explanations, global vs. local explanations, and model-dependent vs. model-agnostic explanation methods. Each dimension plays a crucial role in shaping the interpretability landscape, and understanding their distinctions is crucial to comprehend the stakes and the scope of this thesis.



(a) Decision tree

(b) Linear model

Figure 1.1 – Illustration of two models, a linear model and a decision tree, employed to predict loan approval likelihood based on applicants' age and salary. Blue circles represent loan rejections, while red stars represent loan approvals.

1.1.1 Self-Explainable vs Post-Hoc Explanations

The distinction between self-explainable and post-hoc explanations sets the foundation for the taxonomy. Self-explainable models are constructed in a way that inherently incorporates transparency and intelligibility as integral features. These models possess inherent interpretability and can provide easily understandable insights into their decision-making processes without relying on external explanation techniques. These approaches resort to interpretable-by-design models such as decision trees [80] and simple linear models [134]. To illustrate these concepts, we provide an illustrative example using a dataset that considers loan approval decisions based on individuals' age and salary in Figure 1.1. In this context, Figure 1.1a, introduces a decision tree model, while Figure 1.1b shows a simple linear model. Decision trees are naturally interpretable as they allow humans to follow the conditions outlined within the tree structure until they reach a terminal leaf, which represents a model prediction or decision. This step-by-step navigation simplifies the comprehension of decision logic. Specifically, in the decision tree shown in Figure 1.1a, the explanation reveals that if an individual is younger than 30 years old, their loan application is declined; conversely, if their monthly salary exceeds $2500, the loan is accepted. On the other hand, when it comes to linear models, users can gain insights by inspecting the numerical coefficients associated with each feature. As a clear example, we observe that, as depicted in Figure 1.1b by the linear function: $f(x) = 0.2 \cdot \text{age} + 0.3 \cdot \text{salary}$, salary is the most important factor influencing the loan decision. This function indicates that increasing both age and salary is important for approval.

Post-hoc explanation methods, in contrast, come into play after a complex model, commonly a black-box algorithm, has been trained to generate predictions. This is useful when we do

not have access to the model's internal mechanisms. This inaccessibility can happen due to privacy constraints, for example. Consequently, these techniques employ various approaches to analyze the model's behavior and provide explanations, as reported in various surveys [57, 15]. The most common approach involves utilizing the complex model to label training data and subsequently training a simplified model or surrogate on this new dataset [123, 59, 124]. It is worth noting that this approach has faced criticism for potentially oversimplifying complex models, which can result in the omission of crucial information [129]. Nonetheless, post-hoc explanations offer a key advantage by enabling the interpretation of models without requiring retraining, which proves valuable in terms of time and energy consumption. This can prove particularly valuable for critical production models that are integral to a business and cannot be replaced from one day to another. Therefore, this thesis focuses on post-hoc explanation techniques that enable explainability for already deployed black-box models.

1.1.2 Global vs Local Explanations

Global explanations provide a holistic understanding of a model's behavior across its entire decision space (see e.g., [80, 134, 140]). These explanations are valuable for gaining high-level insights into the model's overall behavior and for identifying patterns or biases. Some common global explanation techniques include decision rules, feature importance scores, partial dependence plots, and decision boundary visualization [104].

In contrast, local explanations zoom in on individual predictions, offering precise insights into why a specific instance received a particular output from the model. In other words, it seeks to answer the question, "Why did the model predict this outcome for this specific input?". Local explanation methods are especially valuable when dealing with complex models that lack inherent interpretability, such as deep neural networks or ensemble models. They aim to shed light on the black-box nature of these models by highlighting the specific features or input characteristics that had the most significant impact on the model's decision for that particular instance. Techniques like Local Interpretable Model-agnostic Explanations (LIME) [123] and SHapley Additive exPlanations (SHAP) [94] stand as two popular and widely-used post-hoc local explanation methods [68]. LIME for instance is a method that explains a complex model by learning a linear surrogate on the outputs of the original model.

Recently, hybrid approaches have combined various local explanations and synthesized them into a global explanation [123] to obtain more comprehensive insights. Thus, researchers have proposed methods such as Black box model Explanations by Local Linear Approximations (BELLA) [121] or Natively Interpretable t-SNE [12]. BELLA is a method that combines linear

models on specific neighborhoods to explain a regression model. Similarly, Natively Interpretable t-SNE generates the best set of linear explanations and their associated coverage for dimensionality reduction. These approaches aim to improve coverage and accuracy in global explanations.

A critical challenge in designing explanation techniques lies in finding the right balance between fidelity to the complex model and simplicity or understandability. Global explanations may lack fidelity to the complex model, as simpler methods like decision trees or logistic regression may not fully approximate and explain every output variation of a model with a vast number of parameters. On the other hand, explaining the decision boundary locally for complex models allows explanation techniques to retain high fidelity to the model in a local context while remaining highly readable.

Throughout this thesis, our focus centers on generating post-hoc and local explanations, to reveal the underlying mechanisms of intricate models and enhancing their interpretability, transparency, and applicability in real-world scenarios.

1.1.3 Model Dependent vs Model Agnostic

Lastly, we discuss the distinction between model-dependent and model-agnostic explanation techniques. Model-dependent methods are specifically tailored to explain the outputs of a particular model or family of models. These techniques leverage the internal characteristics and structure of the model to provide insights into its decision-making process. The explanations derived from model-dependent methods tend to be more faithful, as they may leverage control over the model's training [157] or exploit the unique architecture of the model, such as in neural networks [138] or tree ensembles [8].

On the other hand, model-agnostic methods are not tied to any specific model architecture and can be applied universally to various algorithms. This flexibility enables post-hoc explanations to be applied across a wide range of models, domains, and scenarios. These techniques prioritize transparency and generalizability, enabling a more versatile and inclusive approach to model interpretability. Furthermore, model-agnostic explanations provide a consistent and standardized approach to understanding model behaviors, making them invaluable in scenarios where the deployment of different models is commonplace. However, it is essential to note that while model-agnostic methods provide this flexibility, they may introduce some trade-offs, such as potential loss of explanation fidelity or increased computational complexity. Despite these challenges, model-agnostic techniques empower researchers and practitioners to gain insights into the decision-making processes of various models without the need for specialized

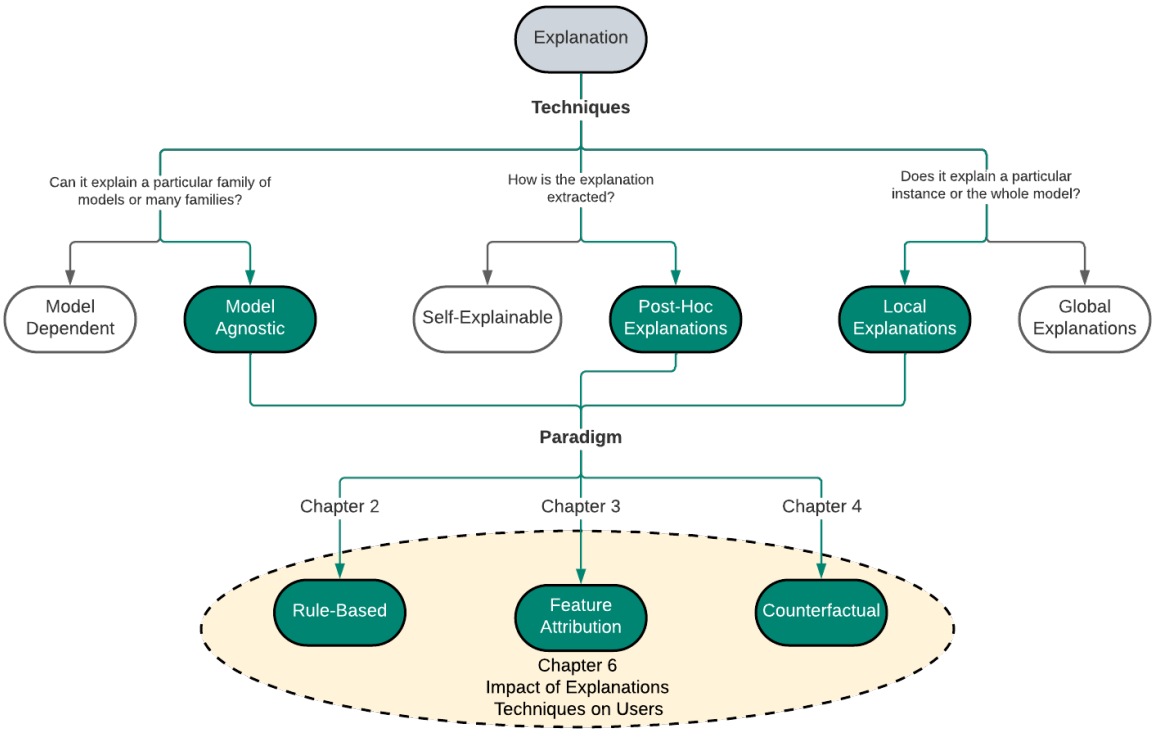


Figure 1.2 – Taxonomy of the explanation techniques. Paths in green represent the explanation techniques studied in this thesis.

adaptations or modifications of the explanation methodology. During this thesis, I co-proposed Therapy [24], a global explanation method for textual data [1], which leverages constraint generation to produce representative text instances representing the most different classes. Therapy stands out as the first method capable of generating accurate explanations while being both model and data-agnostic.

Figure 1.2 illustrates the taxonomy and highlights the specific categories that are the primary focus of this thesis. Our emphasis is placed on model-agnostic methods due to their inherent complexity compared to model-dependent explanations. The latter explanations are tailored to specific model architectures and may not be easily transferable to other models or domains. In the subsequent section, we introduce the distinct explanation paradigms employed to interpret a model. Each of these paradigms assumes a central position in each of the following chapters, ultimately contributing to our user study presented in the second part.

1. not presented in this thesis

1.2 Explanation Paradigms

In this section, we provide a review of three major types of explanation techniques, commonly employed in the domain of tabular and textual data [15, 57, 161]. Before delving into these different families of paradigms, we introduce some notation used throughout the thesis to enhance readability and comprehension. This notation will serve as a helpful tool for readers to navigate within the content effectively.

1.2.1 Notation

Classifier and Instances. In the context of this thesis, we employ standardized notation to describe our problem domain. Our objective centers around explaining the rationale behind the predictions made by a black-box classifier $f : X \rightarrow Y$, for a target instance represented as $x = (x_1, \cdots, x_d) \in X$. This instance is defined differently depending on the data type. For example, it may take the form of a set of features or attributes in the case of tabular data or a sequence of words in textual data. Depending on the domain, these features are represented as either numerical or categorical values (in tabular data) or words/tokens (in textual data). Additionally, Y denotes a set of classes, which can represent categories such as low-risk versus high-risk or newspaper topics.

Surrogate and Neighborhood. To explain the reasoning behind $f(x) = y$, we may leverage surrogate explanation methods. These methods train a white-box surrogate model g, which can be a linear model or decision tree, among others. The surrogate model g approximates the behavior of the classifier f within a local vicinity of the instance x. This locality is determined by a function $\nu_x : X \rightarrow \{0, 1\}$ such that $\nu_x(x') = 1$ if an instance x' is considered a neighbor of x and 0 otherwise. The complete set of all possible neighbors of x is then denoted as $\Phi_x = \nu_x(X) = \{x' \in X \mid \nu_x(x') = 1\}$. It is important to note that the specific implementation of ν_x may vary depending on the chosen explanation method.

Most surrogate methods adopt a common approach to generate local explanations. They train a surrogate model g using a sample of instances created through a generative process. This process produces what is referred to as *artificial instances* denoted as $z \in Z \subset \Phi_x$ within the neighborhood of the instance x [59, 123, 124]. Additionally, if available, these methods also consider the presence of *real instances* that belong to the same neighborhood. These real instances are represented as $t \in T \cap \Phi_x$.

Counterfactuals and Friends. Within this context, we introduce the terms *counterfactual* or *enemy* [82] to describe any instance denoted as $e \in E \subset X$ where $f(e) \neq f(x)$. Conversely, when the classifier assigns the same label to an instance x' as it does to the target instance x (*i.e.*, $f(x') = f(x)$), we refer to x' as a *friend* of x. Counterfactual instances that are close to the target instance x can serve as informative contrastive explanations for $f(x)$.

Symbol	Definition	Symbol	Definition
$f(\cdot)$	Black-box classifier	$g(\cdot)$	Surrogate
X, x	Input domain, target instance	Y	Output domain
T, t	Input dataset, instance	F	Target's friend instances
E, e	Target's enemies, enemy	Φ, $\nu_x(\cdot)$	Locality, Locality function
Z, z	Artificial instances, instance	R	Feature-attribution ranking
$m(\cdot)$	Adherence metric	τ	Adherence threshold

Table 1.1 – Notation used in the thesis.

In the following, we will further describe the three categories to generate local explanations. Examples of these explanation methods are available for tabular data in Table 1.2.

Instance x	family=False, age=18, monitor=True, meals='Low', high-c='No', (y=30)
Expl. Technique	**Explanation**
Feature attribution	(family=False) $\rightarrow -6$, (meals $\geq$ 'Low') $\rightarrow -5$
Rule	If age $\leq 20 \wedge$ monitor=True $\Rightarrow$ non-obese
Counterfactual	meals='Sometimes', high-c ='Yes', $(y = 70)$

Table 1.2 – Explanations for a classifier f computing the risk of obesity $y \in [0, 100]$ with the outcome of 'non-obese' if $y \leq 50$. The attributes consist of the patient's family's obesity antecedents (family), age (age), monitoring calorie consumption (monitor), consumption of food between meals (meals), and high-caloric food (high-c).

1.2.2 Rule-based Explanations

Logic rules are widely recognized for their interpretability and have a rich history of research. Consequently, rule extraction stands as an attractive approach for interpreting complex models. The rule-based methods determine the necessary conditions on the target instance's features

that make the AI predict a particular outcome. These conditions take the form of one or multiple decision rules applied to the input features and are commonly represented as:

$$\text{If } P, \text{ then } Q.$$

Here, P is referred to as the antecedent, while Q serves as the consequent, which, in our context, signifies the prediction of a classifier, such as a class label. Generally, P is a combination of conditions related to various input features. Additionally, it is worth noting that explanation rules take on various forms, including propositional rules, first-order logic rules or fuzzy rules [142].

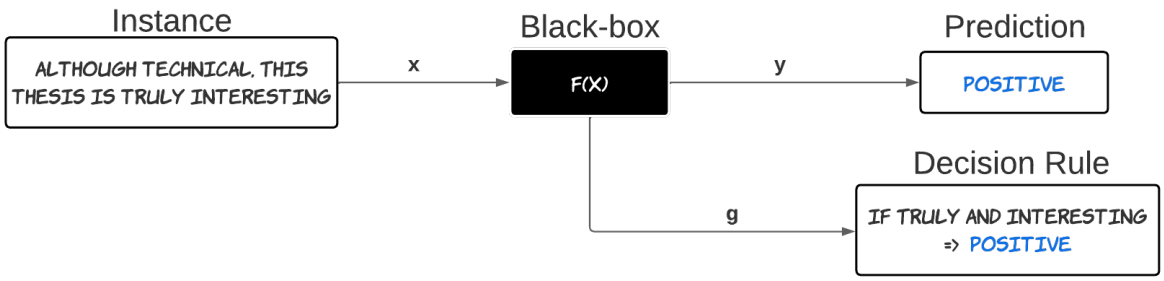


Figure 1.3 – Rule-based explanation for a sentiment classification model. This rule specifies that the model predicted positively due to the presence of the words 'truly' and 'interesting'.

As an example, in Table 1.2 on the obesity dataset, a rule-based explanation specifies that "if an individual who is under the age of 20, monitors their calorie consumption, then our example random forest classifier predicts no risk of obesity". Another example of a rule-based explanation, shown in Figure 1.3, indicates that a classifier assigns a positive classification to a text document if it contains the words "truly" and "interesting". These rule-based explanations offer a structured and interpretable way to understand the AI model's decision-making process based on specific feature conditions [129, 80]. It is important to note that while there may be differences between rules, decision trees, and their respective extraction techniques, we do not make a clear distinction in this context. This is because both rules and decision trees provide similar types of explanations since a decision tree can be seen as a collection of decision rules.

Anchors [124] generates random artificial instances to learn a rule-based explanation. It resorts to a multi-armed bandit exploration that generates those instances gradually. Anchors then computes a single general and accurate decision rule that mimics the black box's behavior on the target instance. The instances are used by a breadth-first-search rule mining procedure that favors shortest rules. Anchors mines for the shortest decision rule as such a rule will cover

more instances [93]. The antecedent of the rule consists of conditions on the input features, e.g., $salary > 50k \ \wedge \ status =$ '$single$', which are used to predict the class assigned to the target instance by the black box. This assignment to a class is made with a high level of confidence, over a given threshold, often set at 95%.

Other methods such as LORE (LOcal Rule-based Explanations) [59], xSPELLS (eXplaining Sentiment Prediction generating Exemplars in the Latent Space) [111], and ABELE (Adversarial Black box Explainer generating Latent Exemplars) [58] rely on decision trees. These techniques have been proposed by Guidotti et al. [59, 111, 58] and extract rules from a decision tree trained on artificial instances that resemble the target instance. They benefit from the tree structure to propose rule-based explanations. Indeed, starting from the leaf in which the instance falls, these methods generate rules by going up until the root of the tree. Alternatively, they can also search for paths in the tree that lead to a leaf node associated with a different black-box prediction, providing this contrastive path as a counterfactual explanation. These methods work differently depending on the data type. For instance, LORE uses a genetic algorithm to create similar instances to the target, whereas xSPELLS and ABELE employ a variational autoencoder to encode the target instance into a latent space and make slight perturbations, specifically tailored to textual and image data, respectively.

In their work, Dhurandhar et al. [35] introduce a novel concept known as "pertinent negative and positive explanations". These explanations are designed to be rule-based explanations and provide insights into why a certain input x is classified as class y based on the presence of certain features $f_i, \cdots, f_k$, and the absence of some other features $f_m, \cdots, f_p$. The researchers achieve this by identifying small, sparse perturbations that either preserve the same prediction when applied to the original input or change the prediction when applied to a target input.

1.2.3 Feature-Attribution

Feature-attribution techniques compute the contribution of a black box's input features to the classification of a target instance. The magnitude of the contribution tells us the importance of the feature for a particular prediction outcome, which can correlate positively or negatively with the answer provided by the black box. For instance, consider the information in Table 1.2, which suggests that in the case of a random forest classifier predicting the risk of obesity using factors like family history and daily routines, a decreased risk of obesity is linked with habits such as low food consumption between meals and having non-obese parents. In another case, for a sentiment prediction model as the one in Figure 1.4, the word 'truly' pushes the model

towards positive predictions with a positive weight of 0.8. Conversely, the word 'technical' makes a negative prediction more likely with a weight of 0.4.

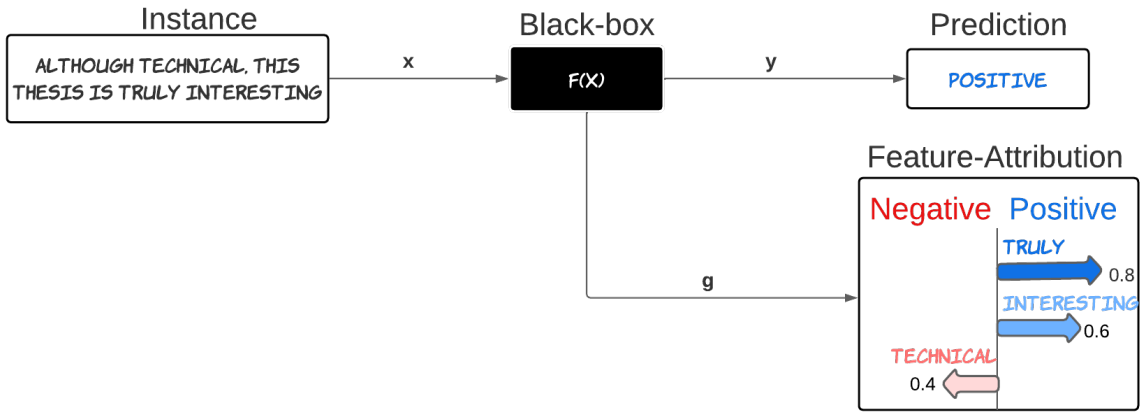


Figure 1.4 – Feature-attribution explanation for a sentiment classification model. The length of each bar represents the extent to which the presence of a specific word in the sentence influences the model's prediction toward the corresponding class (positive or negative).

Several well-known feature-attribution methods, such as LIME [123], SHAP [94], Integrated Gradient (IG) [143], among others, are the most frequently employed in the field of XAI [107, 68]. Among these methods, LIME (Local Interpretable Model-Agnostic Explanations) [123] is the most prominent approach to compute surrogate explanations [68]. Its core concept involves the training of a linear model to approximate the complex model's behavior around a specific instance.

LIME initiates the explanation process by creating an artificial neighborhood in the locality of the target instance. The specifics of this process depend on the data type under consideration. For tabular data, LIME perturbs the numerical attributes of the target according to a μ-centered and σ-scaled normal distribution, where μ and σ are the attribute's mean and standard deviation in the training set. For categorical attributes, LIME uses the empirical distribution of the attribute values. For the textual data, LIME randomly 'hides' part of the input instance (words) to generate an artificial neighborhood.

Subsequently, LIME assigns weights to these artificial neighbors based on their proximity to the target. The weighting is done using an exponential kernel that considers the l_2-distance between neighbors and the target instance. This weighting scheme ensures that closer neighbors are given more importance. LIME then trains a linear model on this weighted neighborhood. The coefficients associated with each input feature in this model form the basis of the explanation.

Additionally, LIME resorts to a regularization term that limits the number of features used by the linear model. This regularization helps simplify the complexity of the explanation.

The effectiveness of LIME has led to a significant body of research exploring the impact of the different components and parameters of LIME on the quality of the resulting explanations. To provide a comprehensive overview of the landscape of LIME extensions, we survey some of these extensions.

While SHAP and LIME produce the same type of explanation, the semantics of their explanations are different [3]. Indeed, if properly parametrized, LIME approximates the instantaneous gradient of the black box w.r.t. the input features [49]. Conversely, SHAP takes root from game theory and computes – or rather approximates – the Shapley values [136]. These values quantify the feature contributions to the difference between the model's answer on a baseline instance and the target. The baseline depends on the use case, e.g., a single-color image, or an empty sentence. This contribution is measured by iteratively replacing some parts of the target instance with values from the baseline and observing its impact on the classifier prediction. The model-agnostic version of SHAP (KernelSHAP) computes the Shapley values by (i) generating artificial instances as in LIME but concentrating the training weights on the closest and farthest instances in the neighborhood, (ii) dropping the regularization term used in LIME to reduce the complexity of the linear surrogate. SHAP's explanations offer interesting theoretical guarantees such as local accuracy, i.e., the surrogate is always accurate on the target instance [94].

Among the local explanation techniques, it has been demonstrated that applying LIME within a neighborhood defined by the classifier's decision boundary can lead to more locally faithful explanations [83]. In that vibe, the Local Surrogate approach centers the generative process not on the target instance but on its closest enemy, which by itself provides a complementary explanation for the model prediction. Local Surrogate then constructs a linear surrogate within a hyper-sphere centered at the closest counterfactual, further enhancing the fidelity and stability of explanations. This method will be further developed in Chapter 3.

Additionally, various extensions of LIME have emerged to address the inherent instability of the original LIME algorithm [135, 148, 160]. In the original LIME algorithm, two executions with the same input may yield different explanations due to randomness in different algorithmic steps. These extensions offer diverse approaches to tackle this instability issue. For example, the authors of Optimized Local Interpretable Model Explanation (OptiLIME) [148] delve into the relationship between the bandwidth parameter, the adherence, and the instability in LIME. OptiLIME underscores the critical role of selecting an appropriate bandwidth value for each instance and uncovers an inverse correlation between bandwidth and explanation instability.

Building upon this insight, OptiLIME introduces a method to select the optimal bandwidth value that strikes the ideal balance between adherence and stability. Throughout this thesis, I collaborated on Smoothed LIME (s-LIME) [50], an extension of LIME that specifically examines the impact of this bandwidth parameter on the quality of generated explanations. s-LIME achieves this by generating neighbor instances in a continuous space. The size of the neighborhood defined by those instances and the magnitude of the perturbation depend on the distance defined by the bandwidth parameter. Thus, s-LIME generates a neighborhood that is more nuanced and allows for higher faithful linear explanations.

1.2.4 Example-based Explanations

Example-based explanations provide users with similar instances or examples that are either classified as the target instance (prototype) or differently (counterfactual). Prototypes are instances representative of a given class [61]. Conversely, counterfactuals illustrate the minimum changes in the target instance necessary to modify the AI's prediction. They, therefore, identify the most *sensitive* features in the AI agent's decision process. Counterfactual explanations have gained prominence for two main reasons. Firstly, they align more closely with the way humans naturally explain concepts, drawing from cognitive processes and human reasoning [21, 102, 147]. Secondly, they fulfill legal requirements for explaining prediction, particularly in accordance with regulations such as the GDPR [151].

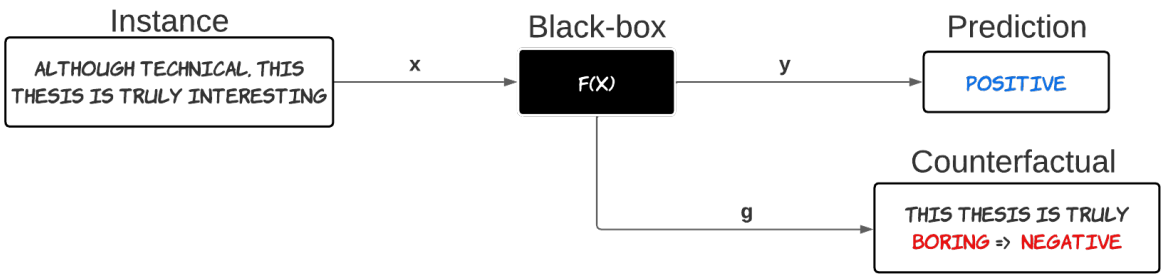


Figure 1.5 – Counterfactual explanations for a sentiment classification model that predicts a text as positive. This explanation illustrates that by substituting the word 'interesting' with 'boring', the text would have been classified as negative. Hence, the word 'interesting' is indeed important for the positive class.

We provide in Figure 1.5, an example of a counterfactual explanation for a model that initially predicts a text as positive. This explanation highlights how replacing the word 'interesting' with

'boring' alters the model's prediction. Table 1.2 offers another instance of a counterfactual explanation, this time for a random forest classifier predicting an individual as having no risk of obesity. This explanation shows that changing an individual's frequency of food consumption between meals from 'low' to 'sometimes', along with consuming high-caloric food, would result in the model predicting a high risk of obesity for that individual.

In recent times, a multitude of innovative counterfactual explanation techniques have emerged, leading to an expansion of the field. This rapid growth is confirmed by the fact that, within just one year, more than 50 new methods were proposed between the surveys conducted by Bodria et al. [15] and the one from Guidotti [55]. Therefore, among the example-based explanations, we first differentiate between the "non-synthetic" methods which return training samples [118] and the methods that generate the instances [147]. We find among the non-synthetic category, methods such as Nearest Unlike Neighbor (NUN) [30], which is derivative from the nearest neighbors method and looks for the nearest element that belongs to a different class. Additionally, we have methods like MMD-critic [76] and Nearest-CT [86] which select prototypes and counterfactuals from the original data points.

However, in this thesis, we focus on methods that generate artificial instances to produce counterfactual explanations because this generating process may be utilized to construct linear or rule-based explanations, as demonstrated in Chapter 3. One of these methods is Growing Spheres [82], which produces the closest valid counterfactual based solely on the target instance and the model to explain. This method, detailed further in Section 3.3, iteratively perturbs the target instance until it identifies an artificial instance that elicits a different prediction from the complex model. Growing Spheres generates instances within a hypersphere centered on the target instance. This hypersphere radius is extended until the closest counterfactual is found. Another approach, Reinforcement Learning Agent eXplainer (ReLAX) [27] leverages reinforcement learning (RL) techniques to generate counterfactual instances. The primary goal of ReLAX is to find a counterfactual instance that is at the same time close to the original instance and classified differently from it. ReLAX does this by formulating the task as a sequence of actions that an RL agent must take to maximize its expected reward. In this context, the optimal policy represents the objective of finding an instance that is both differently classified and close to the original instance. The RL agent in ReLAX operates iteratively, making decisions about which features to modify and by what magnitude. Until the closest counterfactual is found, the agent, at each iteration, selects from the set of features that have not yet been modified and determines the extent to which a chosen feature should be increased or decreased.

Other methods, such as Diverse Counterfactual Explanations (DiverseCF) [106] aim to obtain the highest feasibility and diversity. Thus, DiverseCF achieves diversity by leveraging a mathematical concept known as determinantal point processes. This concept has previously been used in solving problems involving the selection of diverse subsets. In this context, determinantal point processes help ensure that the counterfactuals generated cover a wide range of possibilities. In addition to diversity, DiverseCF incorporates various constraints to make the counterfactual explanations more meaningful and applicable. These constraints include proximity to the instance being explained, sparsity in the generated counterfactual compared to the target, and user-defined constraints to enhance trustworthiness.

Furthermore, it is worth noting that some of the methods we explored in rule-based explanations can also be considered as counterfactual explanations. For example, Pertinent Negative [35] identifies elements that must be absent to maintain the prediction for the target. On the other hand, xSPELLS [111], LORE [59], and ABELE [58] use decision trees to generate explanatory rules, with the paths within the decision tree that lead to different classifications being utilized as contrastive explanations.

1.3 Evaluating Explanations Techniques

The evaluation of explanation techniques is a crucial aspect of assessing their effectiveness and utility. While the fundamental goal of generating explanations is to make them comprehensible to humans, the evaluation of most existing methods has predominantly relied on performance criteria [38, 125]. This focus on performance evaluation can be attributed, in part, to the significant presence of machine learning researchers within the eXplainable AI (XAI) community. Leveraging metrics commonly employed in their respective domains, these researchers quantify the quality and performance of explanation techniques.

As surprising as it may sound, certain researchers have presented arguments against using human validation to assess the accuracy of explanation methods [107]. Indeed, it is pertinent to note that while an explanation may appear plausible to a person, it does not guarantee that it accurately reflects the underlying reasoning of the model [69]. Conversely, when confronted with models that learn nonsensical correlations to derive predictions, explanations that appear implausible should not be penalized, as their purpose is to reflect the model's intrinsic logic. Furthermore, the evaluation of explanations can be significantly influenced by the user and the specific context. The provision of examples that seem reasonable can be a way to evaluate

explanations, but there exists a risk of cherry-picking examples to pass a face-validity test, potentially undermining the evaluation process [37].

In this section, we explore a series of criteria that illustrate the effectiveness of explanation methods. We differentiate between two fundamental approaches: surrogate evaluations and instance-based assessments due to their distinct nature and the specific focus of their evaluation. In the case of surrogate-based explanations, the evaluation process tends to provide more general insights, offering a comprehensive summary of the model's behavior across multiple instances. These evaluation standards often encompass comparisons with a ground truth or assessments of the stability of explanations across various instances. In contrast, example-based explanations are renowned for their specificity, offering insights into individual instances and how they can be altered to achieve different predictions. When it comes to assessing example-based techniques, the process primarily involves evaluating the instances themselves rather than the explanation models.

1.3.1 Surrogate-Based Evaluation Criteria

This section examines a variety of criteria that provide insights into the performance of surrogate explanation methods. These metrics establish performance criteria capable of objectively gauging the effectiveness of these methods in providing comprehensible explanations. Some prominent criteria include fidelity, adherence, uncertainty, stability, and complexity. This focus on performance-based criteria not only aligns with the research community's background but also acknowledges the need for objective assessment in a field where the subjectivity of human interpretation can often be intricate.

1.3.1.1 Adherence and Fidelity

The **adherence** of an explanation refers to the degree to which the explanation accurately represents the decision-making process of the underlying model. This property is crucial and often the first aim of explanation techniques. It is often measured by comparing the prediction of the explanation surrogate and those of the classifier to explain [59, 123]. This comparison may take the form of measuring the accuracy or precision between the two predictions on a set of artificial or real instances. A similar criterion, sometimes confounded with adherence is the **fidelity**. This standard also named agreement, measures whether the features involved in the explanation are effectively important for the model to explain. One way to measure the fidelity of an explanation is to resort to a transparent or glass-box model. By doing so, we can control

the features involved in the prediction and compare them with those indicated as important in the explanation [123]. Another approach involves inserting or deleting elements in the target indicated as important in the explanation. This helps assess the impact of these elements on the classifier's prediction for this target. The intuition behind deletion is that removing the "cause" will force the model to change its decision [117]. Similarly, adding an element indicated by the explanation as important for another class should lower the confidence of the model in the original class.

1.3.1.2 Stability and Uncertainty

When generating an explanation for a model, it is important to assess the extent to which it can be relied upon. Therefore, significant efforts have been made to measure the **stability** of explanation techniques [2, 149]. Robustness serves as a metric to quantify the stability of explanation techniques. Highly robust explanations should exhibit minimal changes in response to slight perturbations for the instance to explain. This indicates that the explanation may serve as a reliable substitute for the complex model within the vicinity of the explanation. Alvarez and Jaakkola [2] proposed a novel metric that relies on the concept of local Lipschitz continuity. This metric perturbs artificial instances located within a ball centered on the target instance and then measures the ratio between the variance observed in the original feature space and the explanation space. In essence, it quantifies the ratio between the distance of (a) the perturbed instances from the target instance, and (b) the original explanation and the modified one. Similarly, Visani et al. [149] introduced two metrics known as the Variables Stability Index (VSI) and Coefficients Stability Index (CSI) to assess the stability of linear explanations. Since linear explanations resort to random perturbations to generate artificial instances, the explanations they produce may vary with each execution. VSI tracks the top features indicated by successive execution of the explanation module, evaluating whether these variables or features consistently appear, while CSI assesses whether the associated attribution scores remain similar across repeated runs.

Another aspect that contributes to users' trust in explanations is the measurement of **uncertainty**. In the work presented in [52], the authors proposed the use of bootstrapping to generate a sample of different explanations and measure the stability of contribution values. This approach involves first selecting a model along with an explanation and generating random samples with varying values. Subsequently, explanations are generated for each of these new samples, and uncertainty is measured as the variation in contribution values between these explanations. Moreover, noteworthy findings from research [74, 89] have highlighted the positive

relationship between improving a model's explainability and the robustness of explanation methods. This implies that enhancing the transparency and interpretability of a model can lead to more stable and reliable explanation techniques.

1.3.1.3 Simplicity and Conciseness

One facet of explanation quality involves assessing the **simplicity** and **conciseness** of the explanations. Simple, concise explanations are more accessible to users and facilitate their understanding of the model's decisions. Various metrics and qualitative assessments can be employed to gauge this aspect of complexity effectively. One quantitative approach involves examining the length of explanations, which can be measured by counting the number of terms or conditions contained within the explanation. For instance, in rule-based explanations, the complexity could be quantified through the number of predicates in the rule [124]. Similarly, for feature-attribution explanations, the number of coefficients associated with input features can serve as a measure of complexity. In the case of decision trees, the depth is a significant factor in assessing complexity. The shorter the path within a tree, the simpler and more concise the explanation it provides [59].

Furthermore, an alternative approach to evaluating conciseness is to employ ground-truth models tailored to different explanation types (e.g., linear or rule-based). These transparent models serve as benchmarks [56], enabling comparisons between the identified important feature coefficients in feature-attribution explanations and their true values. In the case of rule-based explanations, the features included in the paths of decision trees can be assessed for conciseness by comparing them to the ground truth. This approach helps quantify the level of alignment between the explanation and the actual importance of features, shedding light on the simplicity and conciseness of the explanations provided.

1.3.2 Instance-Based Criteria

The instance-based explanation approaches are designed to find optimal instances based on one or multiple desired criteria. In the context of counterfactual explanation techniques, Guidotti [55] has outlined eight criteria, including validity, similarity, sparsity, diversity, actionability, causality, plausibility or realisticness, and discriminative power.

1.3.2.1 Validity and Similarity

A counterfactual is deemed **valid** if its classification by the classifier differs from the prediction on the target instance. This notion of validity is inherent to every counterfactual. Moreover, we consider a counterfactual to be **similar** to the target when the distance between them, determined by a distance function, is low [102, 151]. This criterion is also referred to as "minimality".

1.3.2.2 Sparsity and Diversity

The **sparsity** condition indicates that the counterfactual should possess the fewest distinct features when compared to the target [82]. This principle of sparsity ensures that no other valid counterfactual exhibits fewer different attribute value pairs [151]. Conversely, when generating multiple counterfactuals, we expect **diversity** among them, meaning they should be close to the target but distant from one another. For instance, consider two counterfactuals proposing that an individual should be younger to be classified as having no risk of obesity. In this scenario, the insights gained might be limited compared to two explanations that highlight different attributes such as monitoring calorie consumption and increasing vegetables intake. Indeed, the former explanation fails to provide a nuanced understanding of the model's decision [106].

1.3.2.3 Actionability and Causality

Actionability aims to increase users' trust in the explanation. These counterfactuals ignore unchangeable features such as age or gender. Thus, an actionable counterfactual only differs from the target in other attributes [120]. **Causality** is connected with these two properties, as a causally generated counterfactual validates actionability and plausibility by maintaining any causal relationship between features.

1.3.2.4 Realisticness and Discriminative Power

A counterfactual is considered **realistic** if its attribute values align with those of instances from a broader population. In other words, a realistic counterfactual does not deviate significantly from typical instances in the original dataset. The **realistic** metrics serve as a means to assess whether generated examples lie within the data manifold and remain coherent with the underlying data distribution. Such metrics prevent artificial instances from being considered outliers [147]. One practical approach to measuring realisticness involves calculating the average distance of the generated counterfactual to the k-closest instances from a dataset [147].

This metric gauges whether a generated instance aligns with the original data distribution, helping identify potential outliers. Moreover, Laugel et al. [84, 85] have contributed to this area with their work on the Local Risk Assessment (LRA) metric. LRA provides a means to evaluate whether a counterfactual explanation is justified by verifying the existence of an ϵ-chain between the counterfactual instance and an instance from the training dataset. An ϵ-chain is established by generating intermediate instances between the counterfactual and the real instance. This chain effectively serves as a path connecting two instances, where the model's prediction remains consistent for every artificial instance along this path.

Finally, the **discriminative power** principle stipulates that changes should be discernible and comprehensible to humans. It is important to note that even small alterations, such as modifying a few pixels in an image, may not be noticeable by humans but can significantly impact the model's prediction [72]. This aspect is vital for distinguishing between adversarial attacks which aim to fool a model by slightly perturbing the input and counterfactuals which explain why the model made a prediction [72].

1.4 Outline of this Thesis

After having introduced the key concepts of explainability and defined how they are used in the rest of this thesis, this chapter presented an overview of three kinds of explanation techniques: rule-based, feature-attribution, and example-based. Finally, it presented the evaluation process for these techniques, differentiating between surrogate and instance-based methods.

As we have seen through this overview, numerous aspects are important to generate an explanation adapted to the situation. Such aspects encompass the type of explanation (rules, feature-attribution or counterfactual), the criteria they aim to maximize (fidelity, stability, realisticness) and the model to explain (model agnostic or model dependent). Consequently, this thesis tackles two essential dimensions depicted in Figure 1.6 to generate adapted explanations. First, it delves into the data-centric perspective, with each chapter focusing on one explanation paradigm:

— In the first chapter of this section (Chapter 2), we introduce two improvements to the widely used rule-based explanation method, Anchors [124]. These improvements address the inherent limitations on tabular data by relying on better discretization techniques and broadening the search space of rules for textual data.

— Chapter 3 proposes an approach to determine when it is appropriate to employ linear explanations. This decision is driven by a study of the decision boundary's shape in

the locality of the instance to explain. Within this chapter, Section 3.4 introduces a framework for identifying scenarios in which a linear explanation is most suitable and proposes a rule-based alternative in other cases.

— More and more counterfactual explanation techniques are developed with the goal of generating more and more accurate, faithful, and plausible explanations. However, similarly to the original model that explainability aims to open, these methods are becoming more and more complex. Therefore, in Chapter 4, we study the benefits of using complex black boxes to explain other black boxes.

Moving beyond the data perspective, a crucial aspect to consider when generating an explanation is the person who receives it. As a consequence, the second part of this thesis studies how to generate the best explanation from a user perspective. We thus conducted user studies on laypersons with diverse explanation representations and techniques as illustrated in Figure 1.6.

— In the first chapter of this second part (Chapter 5), we argue for the need for user studies. We thus provide an overview of how existing researchers have evaluated explanation techniques with users. Then, we introduce a methodology to conduct user studies and delve into the various scales and metrics employed to assess the impact of explanations on user perception and behavior.

— In Chapter 6, a comparative analysis evaluates the effectiveness of three distinct explanation techniques: rule-based, feature attribution, and counterfactual. These techniques are evaluated across two distinct representations: graphical and textual. The chapter examines the influence of these techniques on users' trust and comprehension, presenting the merits and limitations of the methods and representations.

This thesis concludes with Chapter 6.5 where all the works of the past three years are summarized. Furthermore, this chapter introduces some open challenges and highlights opportunities for future work.

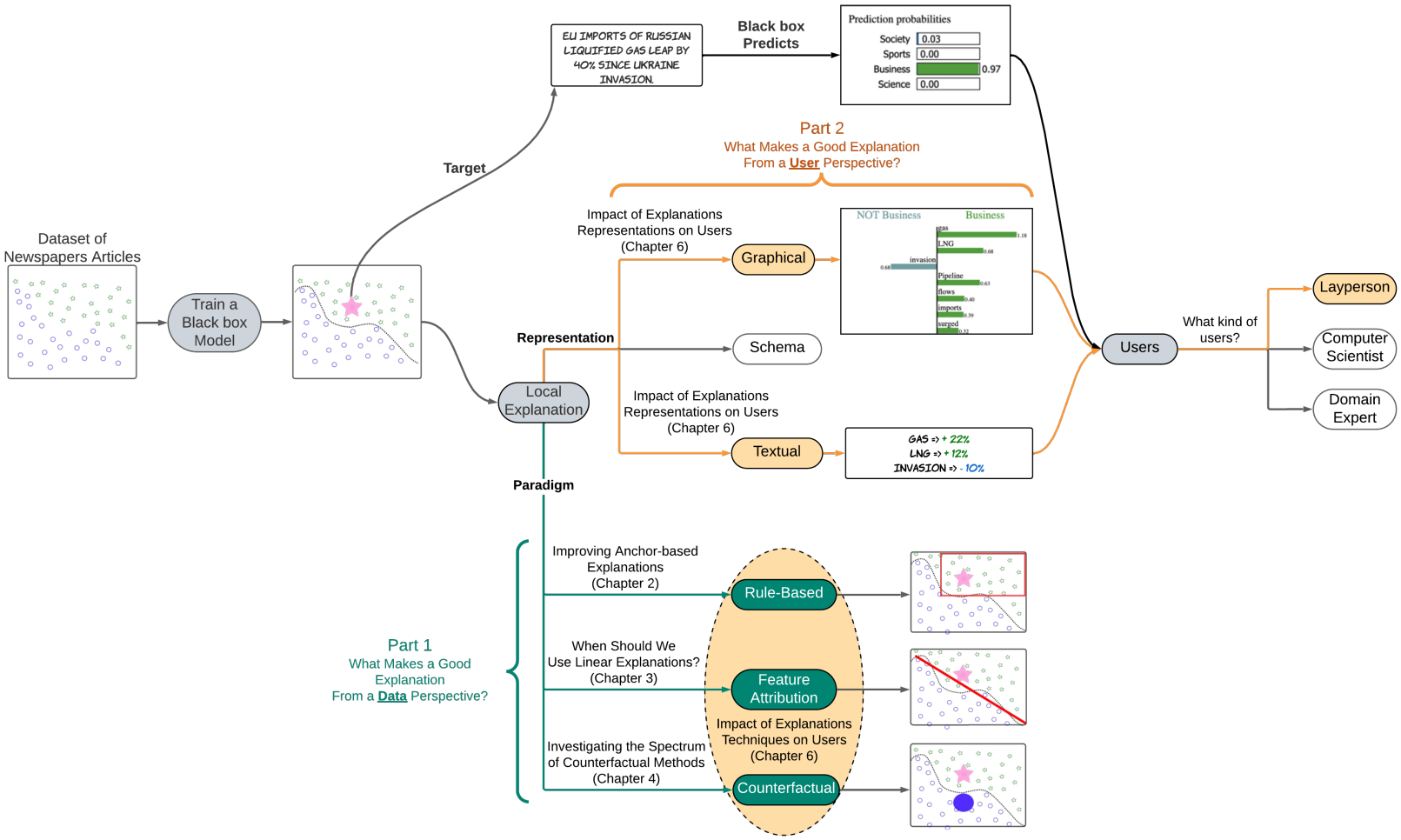


Figure 1.6 – Diagram depicting the aspects of explainability studied in this thesis. The first part depicted by the green boxes, mainly focuses on the different explanation techniques. The focus of the second part is more on the users' aspects and is represented by the orange ovals.

PART I

WHAT MAKES A GOOD EXPLANATION FROM A DATA PERSPECTIVE?

CHAPTER 2

IMPROVING ANCHOR-BASED EXPLANATIONS

Contents

2.1 Context 41

2.2 Anchors 42

2.3 Impact of discretization on Tabular Data 43

2.3.1 New Discretization Methods 44

2.3.2 Experimental Evaluation 44

2.3.3 Results 45

2.4 Improving Anchors on Text 46

2.4.1 Neighborhood Generation Strategies 47

2.4.2 Pertinent Negatives 48

2.4.3 Experimental Evaluation 48

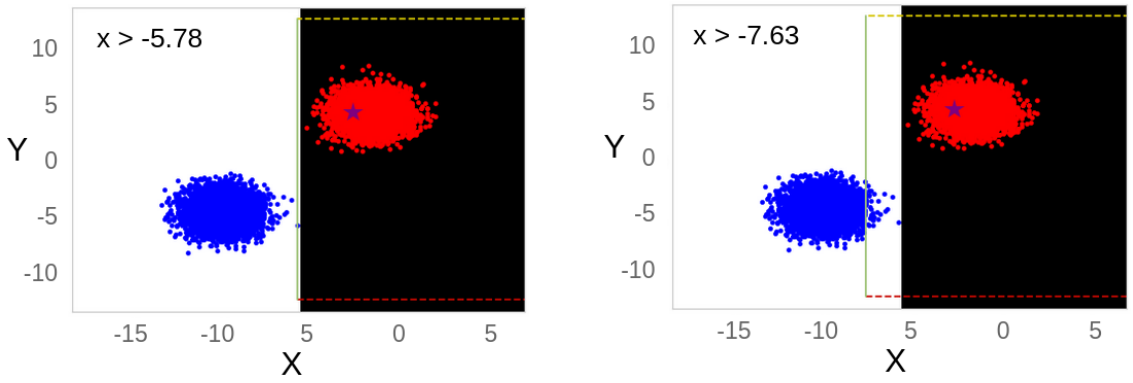
2.4.4 Results 49

2.5 Conclusion 50

2.1 Context

Explanations based on logical rules are a popular strategy to explain the logic of complex black-box machine learning (ML) classifiers [102, 151]. However, approximating a complex model with human-readable rules incurs an inevitable trade-off: Fidelity can only be achieved at the expense of complexity, and complex explanations miss the whole point of explainable ML. For this reason, recent approaches, such as Anchors [124], focus on explanations of local scope. These are if-then rules – also called *anchors* – that mimic the black box in the vicinity of a target instance. This strategy relies on the assumption that the black-box classifier is simpler to approximate when we focus on a particular region of the space.

While local rule-based explanations yield simple and locally faithful explanations, their quality can still be very sensitive to some design factors. One of such factors is the discretization of the numerical attributes for tabular data. Figure 2.1 illustrates the anchors obtained for the same dataset with two discretization methods. When running Anchors with a suitable discretization method on the left-hand side of the figure, we obtain the anchor $x > -5.78 \Rightarrow Red$ that matches the black box's behavior more faithfully than the anchor obtained by the discretization method on the right-hand side.



(a) A good discretization

(b) A suboptimal discretization

Figure 2.1 – Two anchors (depicted as green lines) learned with different discretizations of the numerical features. The target instance is marked as a violet star

Another factor that can impact the quality of an anchor is the training set used to learn the explanation. Anchors [124] generates training samples by perturbing the instance of interest according to a neighborhood generation strategy. Figure 2.2 shows the average anchor length (number of conditions on the rule's antecedent) and precision across 10 instances of three explanations learned with different neighborhood generation methods. The strategy in dark blue

(*pertinent negatives* explained later) provides the explanation with the best trade-off between rule length and precision.

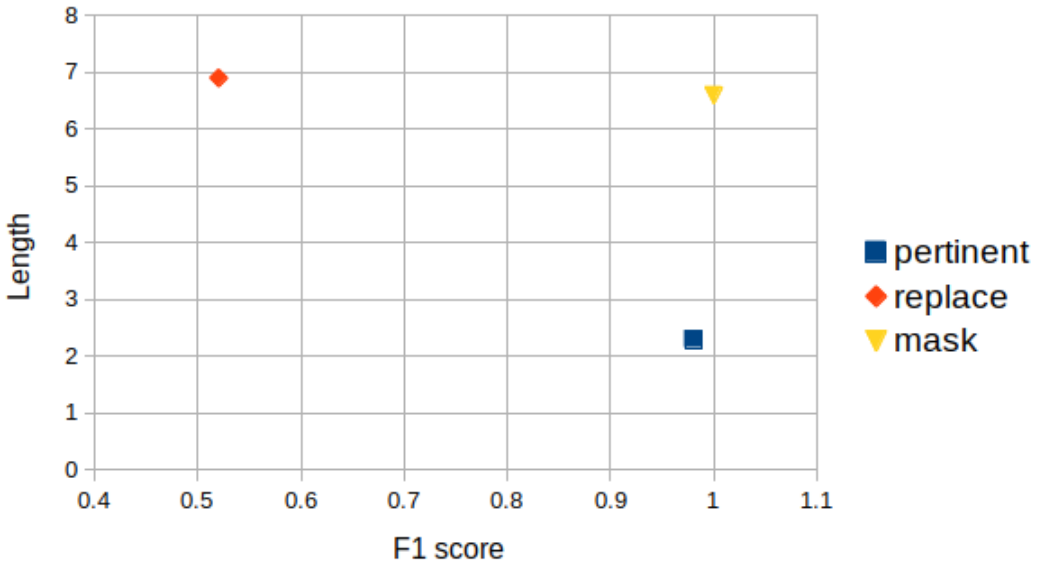


Figure 2.2 – Trade-off between F1 score and anchor length for three neighborhood generation methods

In this chapter, we study the impact of discretization and neighborhood generation on different metrics that define the quality of anchor-based explanations. Our contributions focus on the tabular and text variants of Anchors and include (i) the application of Minimum Description Length Binning (MDLP) [44] to discretize the numerical attributes on tabular data, and (ii) the definition of *pertinent negatives* on text classifiers. Before elaborating on our contributions, we provide a proper introduction to Anchors in the next section.

The research discussed in this chapter formed the basis of the paper entitled: Improving Anchor-based Explanations, which was published at the CIKM 2020 conference [31].

2.2 Anchors

Ribeiro et al. [124] define an *anchor* as a logical rule R that explains a black box f around a target instance x. This instance x is a vector of d attributes that can be either categorical or numerical, and $x[j]$ denotes the value of x for the j-th attribute. The anchor operates on a surrogate interpretable space defined via a conversion function $\eta : \mathcal{X}^d \rightarrow \{0,1\}^{d'}$. The

generated rule has the form:

$$R : \mathbf{B} \Rightarrow f(\eta^{-1}(z)) = f(x) \qquad \text{with } \mathbf{B} = \bigwedge_{j \in F \subseteq \{1,\ldots,d'\}} z[j]$$

The left-hand side (or antecedent) of the rule is a conjunction of conditions that predicts $f(x)$, i.e., the class of the target instance x according to the black box. An example is the rule $age \in [28, 37] \wedge$ *workclass*="*private*" $\Rightarrow$ "*well-paid*" (for simplicity we write only the predicted value on the right-hand side). For tabular data, the interpretable space can be obtained by discretizing the numerical variables – to turn them categorical – and then binarizing the resulting conditions. For text classifiers, the surrogate space is usually defined as the presence or absence of words.

The method proposed by Ribeiro et al. [124] learns rule-based explanations from the answers of the black box f on a randomly generated neighborhood $\mathcal{Z} \subseteq \{0, 1\}^{d'}$ constructed around $z = \eta(x) \in \mathcal{Z}$. Anchors applies principles of depth-first search and multi-armed bandit theory to output the shortest anchor with the largest coverage that satisfies the precision guarantee $prec(R) = P(f(\eta^{-1}(z)) = f(x) \mid \mathbf{B} \wedge z \in \mathcal{Z}) \geq \tau$ for a user-defined precision threshold τ. The coverage of an anchor is the ratio of instances in $\mathcal{Z}$ that match the anchor's antecedent, i.e., $cov(R) = P(\mathbf{B} \mid z \in \mathcal{Z})$.

The next two sections present our contributions that highlight some of the limitations of Anchors and propose improvements.

2.3 Impact of discretization on Tabular Data

The variant of Anchors for tabular data assumes we have access to the black box's training dataset $\mathcal{D} \subseteq \mathcal{X}^d$. Tabular Anchors uses $\mathcal{D}$ to discretize the numerical attributes properly according to the data distribution. It supports three discretization methods. Two of them, *decile* and *quartile*, are based on classical quantile discretization. In contrast, the *entropy* discretization method splits the domain of an attribute j in a dataset $\mathcal{D}$ (denoted by $\mathcal{D}[j]$) so that the information entropy of $f(x)$ – for $x \in \mathcal{D}$ and black box f – is minimized. Anchors' entropy-based discretization outperforms quantile-based discretization in terms of coverage, precision, and anchor length. However, as we show later in this section, it can still lead to relatively long anchors. On these grounds, we investigate the performance of two new discretization methods.

2.3.1 New Discretization Methods

2.3.1.1 K-means

We propose a baseline discretization method based on the k-means clustering algorithm [70]. This method splits the domain $\mathcal{D}[j]$ of an attribute into k clusters that minimize the intra-cluster distance while maximizing the inter-cluster distance. Distance is based on the absolute difference of the values in $\mathcal{D}[j]$. Therefore, and unlike the entropy-based method, our adaptation of k-means does not make use of the labels provided by the black box f. The parameter k is chosen using the Elbow method [144].

2.3.1.2 MDLP Discretization

Fayyad and Irani [44] proposed a method for discretization of continuous-valued attributes into multiple intervals based on the Minimum Description Length Principle (MDLP). Continuous-valued attributes are discretized using the information entropy minimization heuristic. Intuitively, MDLP returns the minimal number of "pure" intervals needed to separate instances from distinct classes. Compared to a traditional entropy-based method, MDLP focuses on compression minimality, hence it outputs as few intervals as possible. Its key heuristic lies in the selection of the "best" cut points.

2.3.2 Experimental Evaluation

We evaluate the quality of Anchors when used with five discretization methods. This includes the three methods already supported, i.e., quartile (Q), decile (D), entropy (E), and the methods proposed in this work, i.e., MDLP (M) and k-means (K). The precision threshold τ (Section 2.2) is set to the default value, that is, $\tau = 0.95$.

2.3.2.1 Metrics

The quality of an anchor is defined by its *coverage*, *precision*, and *length*. Precision, as previously discussed in Section 1.3 refers to the degree of adherence, while the length of the rule gauges the succinctness of the explanation. Thus, shorter anchors with high coverage and precision are preferred. We highlight that the coverage of an anchor is almost analogous to the *recall*: It co-exists in trade-off with the precision. However, unlike recall, coverage can never reach 1, as this would mean that every instance is classified as the target instance. To account

for this issue, we define the normalized coverage $ncov(R)$ of an anchor R as:

$$ncov(R) = \frac{cov(R)}{P(f(\eta^{-1}(z)) = f(x) \mid z \in \mathcal{Z})}.$$

That is, the standard coverage is now normalized by the maximal attainable coverage of an anchor that explains the class given by $f(x)$. With this formulation, we can define the F1 score of an anchor as the harmonic mean of the precision and the normalized coverage. This score provides a trade-off between coverage and precision.

$$F1(R) = \frac{2}{ncov(R)^{-1} + prec(R)^{-1}}$$

2.3.2.2 Datasets

We use three synthetic and two real datasets for our evaluation. The synthetic datasets were generated by randomly drawing 10k instances with the functions `make_blobs`, `make_moons`, and `make_circles` available in scikit-learn [1]. The real datasets comprise (i) *Titanic* [2], where the goal is to predict if a passenger of the Titanic survived based on her age, sex, class, etc., and (ii) *Adult* [3] where we aim at predicting if a person earns more than 50k USD also based on personal characteristics.

2.3.2.3 Black-box models

We tested our contributions on a variety of black-box classifiers, namely logistic regression, support vector machines, multi-layer perceptron, and random forests.

2.3.3 Results

Table 2.1 summarizes the F1 performance of Anchors for a set of instances [4] of each dataset for all the studied black-box models and discretization methods. The labels K, M, D, Q, and E denote k-means, MDLP, decile, quartile, and entropy respectively. The different discretization methods exhibit similar performance on the artificial datasets because all black boxes behave equivalently on those datasets (e.g., they all achieve 95% accuracy). Thus, the difference

1. `https://scikit-learn.org`
2. `https://www.kaggle.com/c/titanic/data`
3. `https://archive.ics.uci.edu/ml/datasets/adult`
4. 10k for the synthetic datasets, 100 for Titanic and Adult.

	Support Vector Machines					Logistic Regression				
	K	M	D	Q	E	K	M	D	Q	E
Blobs	0.89	**1**	0.84	0.85	**1**	0.89	**1**	0.84	0.85	**1**
Circles	0.45	**0.87**	0.43	0.46	0.77	0.45	**0.87**	0.43	0.46	0.77
Moons	0.6	0.7	0.66	**0.72**	0.68	0.6	0.7	0.66	**0.72**	0.68
Adult	0.66	0.66	0.66	**0.98**	0.66	0.66	0.96	0.66	**0.98**	0.66
Titanic	0.42	**0.93**	0.33	0.42	0.42	0.42	**0.93**	0.33	0.42	0.42
MR	3	**1.4**	3.6	2	2	3.2	**1.4**	3.8	2	2.2
	Multilayer Perceptron					Random Forest				
	K	M	D	Q	E	K	M	D	Q	E
Blobs	0.89	**1**	0.84	0.85	**1**	0.89	**1**	0.84	0.85	**1**
Circles	0.45	**0.87**	0.43	0.46	0.77	0.45	**0.87**	0.43	0.46	0.77
Moons	0.6	0.7	0.66	**0.72**	0.68	0.6	0.7	0.66	**0.72**	0.68
Adult	0.66	0.96	0.66	**0.98**	0.66	0.66	0.65	0.66	**0.98**	0.66
Titanic	0.42	**0.93**	0.33	0.42	0.42	0.42	**0.93**	0.33	0.42	0.42
MR	3.2	**1.4**	3.8	2	2.2	3	**1.6**	3.6	2	2

Table 2.1 – F1 score for Anchors using different discretization methods. MR denotes the mean rank of the method.

between each model is very slight. All discretization methods obtain good performance on the highly structured dataset *Blobs* (depicted in Figure 2.1). We observe that overall, MDLP achieves the best F1 followed by quartile and entropy. In some cases, however, MDLP is worse than quartile (e.g., for the *Adult* dataset). In particular, MDLP and quartile split the domain of attributes into fewer intervals, leading to less specific conditions with potentially higher coverage. The use of black-box labels for binning usually gives MDLP a significant advantage over a simple quartile discretization. Besides, the focus on compression minimality makes MDLP output fewer intervals than the *entropy* strategy. Table 2.2 confirms our intuitions as we observe that MDLP yields on average the shortest anchors. An example of an anchor using MDLP in the *Adult* dataset is $age \leq 22 \wedge relationship =$ *"own-child"* $\Rightarrow < 50kUSD$.

2.4 Improving Anchors on Text

In some of our experiments with Anchors on text data, it was impossible to attain the default precision threshold $\tau = 0.95$. This phenomenon makes Anchors output rules with a precision smaller than τ. We argue that the maximal attainable precision of Anchors depends on (i) the distribution of the training neighborhood and (ii) the expressiveness of the rule language. In this section, we study the performance of Anchors for two different neighborhood

	Support Vector Machines					Logistic Regression				
	K	M	D	Q	E	K	M	D	Q	E
Blobs	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Circles	8.14	**2.39**	4.79	3.69	3.69	8.14	**2.39**	4.79	3.69	3.69
Moons	2.47	2.46	**1.95**	2.49	2.72	2.47	2.46	**1.95**	2.49	2.72
Adult	9.37	7.43	7.6	**6.48**	8.54	9.16	8.03	7.89	**6.41**	8.23
Titanic	4.18	**2.58**	4.72	3.52	3.52	4.18	**2.58**	4.72	3.52	3.52
MR	3.2	**1.4**	2.4	2	2.8	3.2	**1.6**	2.2	2	2.4
	Multilayer Perceptron					Random Forest				
	K	M	D	Q	E	K	M	D	Q	E
Blobs	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Circles	8.14	**2.39**	4.79	3.69	3.69	8.14	**2.39**	4.79	3.69	3.69
Moons	2.47	2.46	**1.95**	2.49	2.72	2.47	2.46	**1.95**	2.49	2.72
Adult	8.99	7.26	8.34	**6.67**	8.49	9.11	7.21	8.26	**6.77**	8.84
Titanic	4.18	**2.58**	4.72	3.52	3.52	4.18	**2.58**	4.72	3.52	3.52
MR	3.2	**1.4**	2.4	2	2.8	3.2	**1.4**	2.4	2	2.8

Table 2.2 – Anchor length using different discretization methods. MR denotes the mean rank of the method.

generation strategies and propose an extension of the rule language by considering negated conditions, known as pertinent negatives in the explainable AI literature.

2.4.1 Neighborhood Generation Strategies

The variant of Anchors for text classification converts a textual instance into a surrogate binary vector where each entry defines the absence or presence of a word of the target phrase. Consider, for example, a black-box classifier f for sentiment analysis and the target instance "This is a good book". Anchors will convert this instance into a five-component vector, i.e., 11111, and generate neighbors by randomly toggling off bits of this binary representation. Examples are the instances 10101 or 11101. An anchor is induced from that set of neighbors and their class labels according to the black box f. However, f operates in a different space than Anchors. Hence, the inverse conversion function η^{-1} must map the generated neighbors to actual text instances. The strategy called *mask words* (MW) does so by replacing the words of each zero component with a neutral wildcard unseen before by the black box. In our example the neighbor 11101 becomes "This is a $\mathbb{W}$ book" for wildcard $\mathbb{W}$. The strategy called *replace words* (RW), on the other hand, replaces toggled-off words with random words that have the same syntactic role, i.e., they would be assigned the same part-of-speech tag. For instance, the neighbor 11101 could become "This is a **great** book".

2.4.2 Pertinent Negatives

We highlight that anchors are defined on conjunctions of non-negated conditions. For text data, this entails conditions on the presence of words in phrases. This design decision guarantees simpler rules while keeping the search space under control. On the downside, it imposes limits on the expressiveness of explanations. Inspired by the work presented by [35], we propose to change the language of Anchors and provide explanations on the absence of words. Those words are known as *pertinent negatives* (PN) and can be seen as counterfactual explanations, i.e., words whose presence would change the answer of the black box.

Considering the absence of all possible words in the corpus makes the search space for anchors prohibitively large. Hence, we apply two mechanisms to alleviate this fact. First, we focus on a limited set of words. This set consists of the top k most frequent words that co-occur next to the words of the target instance, stopwords excluded. For our example "This is a good book", our algorithm would consider words such as *scientific*, *interesting*, or *very* as they may often appear with "book" and "good". Second, we set an upper bound p in the number of pertinent negatives allowed in explanations.

It follows that a neighborhood generation method purely based on pertinent negatives represents a phrase as a vector of $m + p$ components where m is the number of words in the target phrase and p is the number of pertinent negatives. The target instance is mapped to a vector where the first m elements are set to 1 and the remaining are set to 0. Neighbors are then generated by randomly toggling on the zero entries of the pertinent negatives, which instructs Anchors to add the word to the phrase. Our goal is to show the potential and viability of pertinent negatives in Anchors, thus we leave as future work the implementation of a hybrid approach that combines pertinent negatives with one of the classical strategies for neighborhood generation based on present words.

2.4.3 Experimental Evaluation

We evaluate the discussed neighborhood generation strategies using the F1 measure and the anchor length as quality criteria. For pertinent negatives, we use $p = 20$.

	Support Vector Machines			Logistic Regression			Multilayer Perceptron			Random Forest		
	RW	MW	PN	RW	MW	PN	RW	MW	PN	RW	MW	PN
Tweets	5.1	**4.3**	8.1	**4.6**	6.4	9.1	**2.1**	4.6	7.2	**3.7**	5.4	4.2
Polarity	8.3	7.7	**4.9**	6.7	**3.6**	4.5	6.9	6.6	**2.3**	**2**	**2**	**2**

Table 2.3 – Length of textual Anchors for different neighborhood generation strategies.

	Support Vector Machines			Logistic Regression			Multilayer Perceptron			Random Forest		
	RW	MW	PN	RW	MW	PN	RW	MW	PN	RW	MW	PN
Tweets	0.63	0.41	**0.82**	0.56	0.31	**0.91**	0.85	0.44	**0.95**	0.57	0.27	**0.95**
Polarity	0.35	**1**	0.87	0.35	**0.98**	0.86	0.52	**1**	0.98	0.47	**1**	0.79

Table 2.4 – F1 score of textual Anchors for different neighborhood generation strategies.

2.4.3.1 Datasets.

Our experimental datasets comprise (i) *Polarity*[5], a set of movie reviews for sentiment analysis, and (ii) *Tweets*[6], a set of tweets used for a multi-class classification task focused on predicting the occurrence of emojis.

2.4.3.2 Black-box models.

We use the same black-box models as in Section 2.3.2.3. Those models were trained on a vector representation of the phrases based on word counts and provided by the class `CountVectorizer` of scikit-learn. In this representation, a phrase is converted into a sparse vector such that each component is associated with a word seen in the corpus – the set of phrases used for training – and stores the frequency of the word in the phrase.

2.4.4 Results

We summarize the aggregated results for the F1 measure and the anchor size among 10 randomly selected instances in Tables 2.3 and 2.4. We first observe that the *replace words* strategy (RW) lies far behind the pertinent negative (PN) and mask words (MW) for the *Polarity* dataset. While it usually produces anchors of high precision, the coverage of those anchors is very low, in other words, it generates overly specific explanations. This intuition is confirmed by Table 2.3, where we can observe that *replace words* yields, on average, longer anchors than the other strategies. These are a consequence of the neighborhood generation strategy. By

5. `http://www.cs.cornell.edu/people/pabo/movie-review-data/`
6. `https://competitions.codalab.org/competitions/17344#learn_the_details-data`

replacing toggled-off words with other words of the same syntactic role, the neighbor instances become very unstable: the addition of a single word can change the meaning of the phrase as well as the black box's answer. We observe this phenomenon to a lesser extent when using the strategy *mask words* in the *Tweets* dataset. This happens because replacing a word with the wildcard forces the black box to decide based on fewer words. We observe that pertinent negatives lead to short and still fairly accurate anchors on *Polarity*, and achieve the best F1 score on *Tweets* – at the expense of length – while it returns more complex rules on Tweets and a slightly less F1 score on *Polarity*. These long anchors are due to the fact that the training set for PN consists of many more features. This is also aggravated by the large number of classes as *Tweets* is a multiclassification problem with 20 different emojis to predict. An example anchor with PN is $\neg caring \wedge downpur \Rightarrow$ 😂 for the tweet "Totally worth getting caught in this evening's downpour. #jacquelineonassisreservoir". The addition of the word "caring" may have resulted in the classifier predicting a different emoji.

2.5 Conclusion

In this chapter, we delved into the core aspects that influence the effectiveness of anchor-based explanations for machine learning models. Specifically, our investigation centered on two critical components: the discretization of numerical features in tabular data and the strategy employed for generating neighborhoods in text data.

Our findings reveal that a meticulous adjustment of these components yields substantial enhancements in the precision, coverage, and length of anchor-based explanations. This illustrates the potential of finding more accurate and comprehensive explanations for other interpretability methods in machine learning models.

We identify two promising directions for future research. Firstly, we envision assessing the impact of discretization on other explanation methods, such as LIME, thereby broadening the scope of our insights. Secondly, we aim to explore post-hoc discretization techniques, aiming to expand the coverage of Anchors. A possible approach is to expand the boundaries of each variable within an anchor until precision begins to decline. This approach has the potential to increase the coverage of the explanations.

Furthermore, we plan to investigate the synergy between pertinent negatives and mask word strategies, aiming to create more expressive and accurate anchors that encompass both the presence and absence of specific words. This holistic approach holds the promise of further

enriching the explanatory power of anchor-based explanations. The code, data, and experimental results are available at `https://github.com/juliendelaunay35000/anchors`.

In this chapter, we assumed that a user would apply Anchors without considering its suitability for a given use case. This is because Anchors was proposed as a general, model-agnostic approach. However, as we saw, Anchors is not always optimal on all black-box classifiers and datasets. In the next chapter, we put to test the question of whether a one-size-fits-all approach for surrogates is ideal. In this chapter, we emphasized refining the components that underpin the effectiveness of rule-based explanations. Our exploration now revolves around linear surrogates, which, while widely used for their simplicity and fidelity, may not always be the most suited for delivering unambiguous, faithful explanations. This leads us to introduce a novel method that characterizes black-box classifier decision boundaries and identifies the ideal scenarios for deploying linear models as reliable explanations. This shift in focus represents our quest to make interpretability more nuanced, adaptable, and data-driven.

CHAPTER 3

WHEN SHOULD WE USE LINEAR EXPLANATIONS?

Contents

3.1 Context 54
3.2 Preliminaries 56
3.2.1 Problem Statement 56
3.2.2 Linear Explanations 57
3.2.3 Adherence and Fidelity 57
3.2.4 Existing Methods 58
3.3 Counterfactual Explanation Methods 58
3.3.1 Counterfactuals for Tabular Data 58
3.3.2 Counterfactual Techniques for Textual Data 61
3.4 Adapted Post-hoc Explanations 65
3.4.1 APE Oracle 68
3.4.2 Linear Explanations 70
3.4.3 Rule-based Explanations 70
3.4.4 Illustrative Example 71
3.5 Experiments 73
3.5.1 Experimental Setup 74
3.5.2 APE Oracle Evaluation 76

3.5.3 Comparison with other Explanation Methods 82
3.5.4 Ablation Study 84
3.5.5 Summary and Discussion of Linear Suitability Results 86
3.5.6 Counterfactuals Evaluation 86
3.6 Discussion and Conclusion 88

3.1 Context

One way to explain a black-box model in a post-hoc manner is to learn a surrogate white-box model that mimics the black box. Even though there may be multiple surrogates or paradigms to explain the verdict of a black box for a given scenario, this decision is rarely based on the particularities of the use case. In the previous chapter, we focused on how to improve the quality of rule-based explanations, which are fairly popular. In this chapter, we shift the focus to address the nuanced question of selecting the most suitable explanation method in a given context. Linear surrogates, appreciated for their simplicity and faithfulness, are another popular choice for locally approximating black-box models [68]. Despite the popularity of local linear explanations, they may not always be the most adapted method to explain a black-box outcome.

To illustrate this, consider the two scenarios depicted in Figure 3.1. In Figure 3.1a, the instance of interest lies in a zone where there is clearly a single local linear approximation for the black-box classifier. In contrast, the target instance in Figure 3.1b depicts a scenario where three possible linear explanations are possible. Since these approximations exhibit different slopes, the attribution scores assigned to the input features are obviously contradictory – a situation that would harden interpretation. While we could provide one of the explanations for Figure 3.1b, that would tell an incomplete story.

Based on the aforementioned arguments, this chapter proposes APE, which stands for Adapted Post-hoc Explanations, a novel method to determine *a priori* whether a black-box classifier and a target instance admit a faithful and unambiguous local linear explanation. When this is not the case, APE does not stop but recommends a different explanation paradigm. Building upon the knowledge acquired in the preceding chapter, we investigate the use of rule-based explanations in our experimental work. Through our empirical evaluation, we demonstrate that these explanations serve as a valuable complement to linear explanations. APE operates by characterizing the classifier's decision boundary, which is achieved by identifying the target's

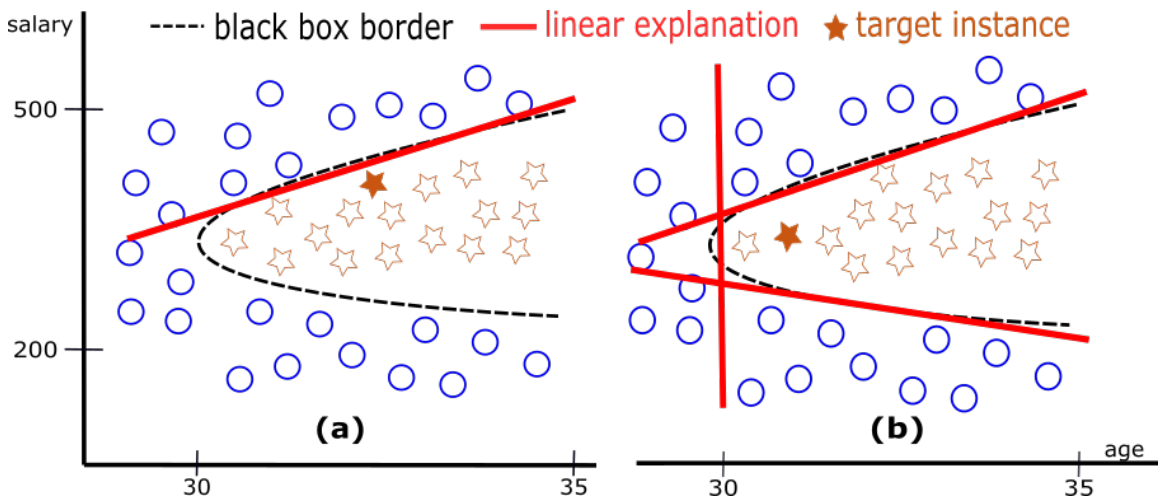


Figure 3.1 – Two explanation scenarios for a classifier and a target instance (the filled star): (a) a suitable single linear explanation; (b) three contradictory linear explanations.

closest counterfactual instance. We recall that counterfactual instances, also called enemies, are instances that are close to the target instance but are classified differently by the black box. Such instances can be used as explanations that highlight the minimal changes required on the target instance to change the classifier's outcome. All in all, our contributions are:

— A definition of *suitability* for explanations based on local linear surrogates. This definition builds upon existing notions such as *adherence* and *locality*, which we also define formally.

— Three novel algorithms for counterfactual exploration are introduced: *Growing Fields* (GF), *Growing Language* (GL), and *Growing Net* (GN). These techniques build upon the Growing Spheres (GS) algorithm [82]. Notably, Growing Fields is tailored for tabular data and handles categorical attributes. Growing Fields also accounts for the distribution of input features, employing the standardized Euclidean distance as a robust metric. Growing Language and Growing Net are specifically designed to generate counterfactual explanations for textual data. They rely on external knowledge to explore the search space and restrict it to pertinent words.

— The APE Oracle is a linear suitability test that tells users whether a black-box classifier can be locally approximated by a single and faithful linear surrogate. To do so, APE characterizes the distribution of the instances around the decision boundary.

— The APE algorithm that returns a linear explanation if suitable. Otherwise APE proposes a rule-based explanation. In all cases, APE computes complementary counterfactual explanations.

The chapter is structured as follows. After formulating the problem and introducing preliminary concepts in Section 3.2, Section 3.3 introduces our novel counterfactual explanation techniques. Section 3.4 elaborates on the APE approach before we evaluate it on a handful of datasets and classifiers in Section 3.5. This is followed by a discussion of our insights.

Most of the work presented in this chapter was the subject of the paper: When Should We Use Linear Explanations?, published at the CIKM 2022 conference [32]. The results on textual data are novel works.

3.2 Preliminaries

3.2.1 Problem Statement

Given a black-box classifier $f : X \to Y$ trained on a dataset $T \subset X$, and a target instance $x = (x_1, \cdots, x_d) \in X$, our goal is to construct an Oracle that tells us whether a linear surrogate g learned on a locality $\Phi_x \subset X$ (defined below) is suitable to explain $f(x)$. By "suitable" we mean that two *contradictory* linear explanations g, and g' may not have the highest adherence in Φ_x – the adherence being the outcome agreement between f and g. In this formulation, Φ_x is a region of the space that (i) covers x, (ii) is traversed by f's decision boundary, and (iii) is maximal, otherwise stated, the surrogate g cannot attain the quality guarantee $m(g) \geq \tau$ in any locality $\Phi'_x \supset \Phi_x$ for some adherence metric m.

Requirement (i) guarantees that the target instance x is included in the surrogate's training set. Moreover, requirement (ii) ensures that this training set contains both instances inside and outside the class $f(x)$. Consequently, the minimal locality satisfying these two requirements should be centered on the decision boundary – more precisely on x's closest counterfactual. This alignment places the target instance x directly on the boundary, as illustrated by the inner dotted circle in Figure 3.2. Incorporating requirement (iii) implies that the scope Φ_x might be expanded if the surrogate g maintains a strong adherence. Thus, Figure 3.2 illustrates two localities represented by distinct dashed circles. The blue circle denotes the initial locality, with a radius equivalent to the distance between the target instance and its associated closest counterfactual. The orange circle represents the largest possible locality to approximate a black box model with a linear surrogate while preserving adherence between the linear model and the complex model above a given τ threshold. In such a scenario, the explanation generalizes to broader regions of the data space, as depicted by the larger dashed circle.

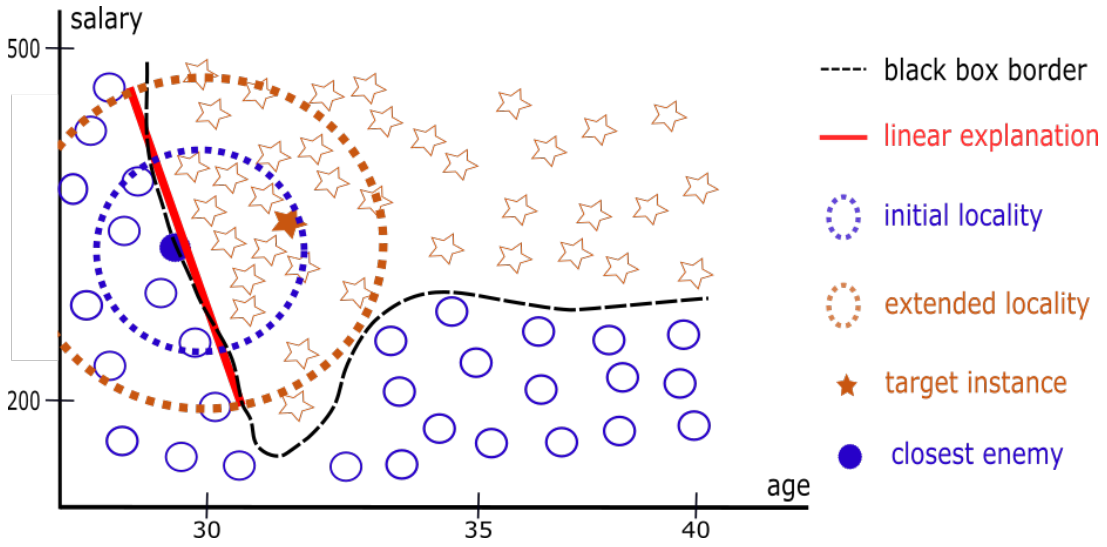


Figure 3.2 – A linear explanation for a classifier and a target instance x. The inner circle (dotted in blue) is the minimal locality Φ_x that covers x and is traversed by the decision boundary. Locality can be extended (orange circle) and still provide an equally good linear approximation for the black box. Friends F of x are represented by yellow stars, and enemies E by blue circles.

3.2.2 Linear Explanations

In order to explain the outcome $f(x)$ of a classifier f on a target instance x, methods such as LIME [123] or Local Surrogate [83] provide a signed feature-attribution ranking $R(g)$ that consists of ordered sets of features $R^+(g)$ and $R^-(g)$. The features in $R^+(g)$ contribute positively to predicting the class $f(x)$, whereas the features in $R^-(g)$ push towards predicting a different class. The ranking is based on the coefficients of a linear surrogate g that approximates f in a locality or neighborhood around x. We say that two linear explanations g and g' for $f(x)$ are contradictory if they induce different attribute rankings, more formally, if $R(g) \neq R(g')$.

3.2.3 Adherence and Fidelity

The quality of a surrogate model g for a black-box classifier f is evaluated through the notions of adherence and fidelity. The *adherence* of a surrogate model g for a black-box model f is the degree of agreement between f's and g's outcomes. The *fidelity*, on the other hand, assesses the surrogate's ability to identify the features truly employed by the black-box model. When f is a true black box, users can only rely on adherence to estimate the quality of explanations.

3.2.4 Existing Methods

LIME [123] stands out as the predominant method for computing local linear explanations. It has been shown by Laugel et al. [83] that we can learn more locally faithful explanations if we apply LIME on a neighborhood traversed by f's decision boundary. In that vibe, the Local Surrogate (LS) approach [83] centers the generative process not on the target instance x but on its closest enemy e – which by itself constitutes a complementary explanation for $f(x)$. LS then learns a linear surrogate on a neighborhood defined by a hyper-sphere centered at e, as depicted by the inner circle in Figure 3.2.

3.3 Counterfactual Explanation Methods

To inspect the decision boundary of a classifier f in the vicinity of a target instance x, the first step is to find its closest enemies. Growing Spheres (GS) [82] is a method that searches for enemies of x by drawing instances uniformly within the volume of a l_2-sphere of radius r centered at x. The value of r is adjusted so that the resulting sphere traverses f's decision boundary and encompasses enemies of x lying close to the border. While Growing Spheres is known for its simplicity and versatility, it comes with several inherent limitations. Notably, it does not handle categorical attributes and does not consider the variance within feature distributions. Additionally, Growing Spheres is primarily designed for generating counterfactuals in tabular data settings, making it unsuitable for data types like text. Therefore, we developed three enhancements of the Growing Spheres algorithm that we call Growing Fields (GF), Growing Language (GL), and Growing Net (GN). We first develop in this section on Growing Fields, the tabular extension which proceeds likewise, but tackles some of the limitations of Growing Spheres as explained next. We then present Growing Language and Growing Net, our two extensions for textual data.

3.3.1 Counterfactuals for Tabular Data

Growing Fields generates instances by slightly perturbing the target instance until it identifies an instance classified differently by the black box, thus providing a counterfactual explanation. The instances generating process of Growing Fields is outlined in Algorithm 1, while the main algorithm is detailed in Algorithm 2. Growing Fields extends Growing Spheres through two key enhancements: firstly, it incorporates the distribution of the input features, and secondly, it deals with categorical attributes.

Algorithm 1 The $\mathcal{F}$ instance generation process

Require: a dataset $T \subset X$, a radius $r \in (0, 1]$, an instance $x = (x_1, \ldots, x_d) \in X$
Ensure: An artificial instance $z = (z_1, \ldots, z_d)$

for $i \in 1 \ldots d$ **do**
  if x_i is numerical **then** // $A_i = max_i - min_i$
    $a = min(0, r \times A_i(T) - \sigma_i(T))$
    $b = a + \sigma_i(T)$
    $z_i \leftarrow x_i + \rho_k$ with $\rho_k \sim \mathcal{U}(a, b)$
  else
    $z_i \leftarrow (x_i$ with prob. $1 - \rho_k)$ with $\rho_k \sim \mathcal{U}(0, r)$
  end if
end for
return z

3.3.1.1 Attribute-dependant perturbations

By drawing instances uniformly in a l_2-sphere, Growing Spheres assumes that all numerical attributes should be perturbed at the same rate. In reality, the attributes may have different amplitudes, variances, and distributions. Consequently, in Growing Fields the perturbation added to a numerical attribute x_i follows a uniform distribution that depends on both the radius r and the attribute's domain amplitude $A_i(T)$, and at the same time preserves the attribute's std. deviation in the input dataset T – denoted by $\sigma_i(T)$. Therefore, Growing Fields generates uniformly perturbed instances depending on the radius r of the field and follows a uniform distribution whose variance is $\sigma_i(T) = \frac{(b-a)^2}{12}$ given lower and upper bound parameters a and b. Growing Fields computes a and b by solving:

$$\begin{aligned} v &= \sqrt{12\sigma_i(T)} \\ a &= min(0, r - v) \\ b &= a + v. \end{aligned}$$

The product $\rho \times v$ where $\rho \sim \mathcal{U}(0, r')$, $v \sim \mathcal{N}(0, \sigma_i(X))$, and $\sigma_i(X)$ is the standard deviation of the i-th attribute of x in the input dataset. This implies that the vicinity generated by Growing Fields around an instance x is not anymore a sphere, but rather a volume or, as we call it, a *field*. The actual shape of this field depends on the distance function. We highlight that taking into account the data distribution guarantees a data-aware exploration of the space, which results in a speed-up of up to 2 orders of magnitude w.r.t. Growing Spheres as shown in the results section.

Another limitation of Growing Spheres is that all attributes have the same impact when computing the distance between two instances. That said, a salary "distance" of 30 EUR is insignificant compared to an age "distance" of 30 years. On those grounds, Growing Fields normalizes the contribution of attribute i using the mean μ_i and standard deviation σ_i in the training set, which boils down to the standardized Euclidean distance[1]:

$$dist(x, x') = \sqrt{\sum_{i=1}^{d} (\frac{(x_i - \mu_i) - (x'_i - \mu_i)}{\sigma_i})^2} \tag{3.1}$$

Equation 3.1 assumes that the categorical attributes have been one-hot encoded. We normalized the distance by dividing it by the maximum distance between the target instance and the instances in the dataset. This normalization guarantees that the generated instances do not exceed the distance of real instances when their distance is below 1. In contrast, Growing Spheres lacks a predefined upper limit on the radius value r, allowing for unconstrained variations. This enables us to establish a meaningful distance between the counterfactual and the target instance, one that can be compared to the rest of the dataset.

3.3.1.2 Support for categorical features

The original Growing Spheres algorithm does not support categorical attributes, such as the gender or the marital status of a person. We can now handle those attributes by treating them as random continuous variables distributed uniformly within the range of $[0, r]$. Consider a field with radius $r = 0.5$ and a target instance with the attribute $sex = F$. If by drawing a random value in $[0, 0.5]$ we obtain, for example, a value of 0.2, we interpret it as throwing a biased coin that keeps the sex of the target instance with probability $1 - 0.2 = 0.8$. If the attribute defines more than two categories, e.g., {single, married, divorced, widowed} and we have to change the category, we employ the re-adjusted empirical probabilities of the other categories in the input dataset T to randomly choose the new category. Note that our way of handling categorical attributes requires a parameter r, the radius of the sphere that lies in $(0, 1]$.

Algorithm 1 details the resulting generation process, called $\mathcal{F}$ (which stands for field), used to draw random artificial instances with both numerical and categorical attributes. The result of integrating $\mathcal{F}$ into Growing Spheres gives rise to the Growing Fields algorithm detailed in Algorithm 2. Growing Fields starts with an initial field of radius r_0 and reduces it until no

1. This is a special case of the Mahalanobis distance when the covariance matrix is diagonal.

Algorithm 2 GROWING FIELDS (GF)

Require: a dataset $T \subset X$, a target instance $x = (x_1, \ldots, x_d) \in X$, a classifier $f : X \to Y$,
Hyper-parameters: $r_0 = 0.1$, $\theta = 1.8$, $n = 2000$ as defined by Growing Spheres [82]
Ensure: Set Z of instances; resulting field radius r
1: $r \leftarrow r_0$
2: $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
3: **while** $\exists\ e \in Z \mid f(e) \neq f(x)$ **do**
4: $r \leftarrow r/2$
5: Update $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
6: **end while**
7: **while** $\nexists\ e \in Z \mid f(e) \neq f(x)$ **do**
8: $r \leftarrow \min(1, \theta \times r)$
9: Update $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
10: **end while**
11: **return** $Z, \arg\min_e \{dist(x, e) \mid e \in Z$ **and** $f(e) \neq f(x)\}$

enemies are found (lines 3-6). In the second stage, the field is *gradually expanded* until the decision boundary is crossed and close counterfactual instances can be reported (lines 7-10). The algorithm then returns x's closest counterfactual.

3.3.2 Counterfactual Techniques for Textual Data

The Growing Fields algorithm (Algorithm 3) works only for tabular data. In this section, we introduce a pair of innovative techniques referred to as Growing Language and Growing Net, for generating counterfactuals to explain the predictions of any text-based ML model. These techniques are adaptations of the Growing Spheres method [82] to textual data. Growing Language and Growing Net both involve an iterative process in which words within the target text are successively replaced. The objective of this iterative process is to generate sparse counterfactual explanations, aiming to minimize the number of modified words.

Algorithm 3 outlines the iterative process employed by Growing Language and Growing Net. In the first step (lines 1 to 4), both approaches generate d sets of potential word replacements ($W = (W_1, \ldots, W_d)$) for each word in the target document ($x = (x_1, \ldots, x_d)$). The external module employed to generate the set of potential word replacements is the main distinction between these two methods. These modules are detailed in the following. Subsequently, Growing Language and Growing Net create artificial documents iteratively, detailed between lines 7 and 15 and exemplified as a tree structure in Figure 3.3. These documents are generated until every word in the original document is replaced or a valid counterfactual is discovered. During each iteration, they initialize a set of n artificial copies of the original document (x) and

Algorithm 3 Generating Counterfactual Explanations

Require: a target instance $x = (x_1, \ldots, x_d) \in X$, a black-box classifier $f : X \rightarrow Y$;
$Store\_Similar\_Words()$ → a function to store similar words for an input word
Hyper-parameters: $n = 2000$; $\theta = 0.8$ (similarity_threshold)
Ensure: one or multiple counterfactual instances
Initialize d sets $W = (W_1, \ldots, W_d)$ of candidate replacement words
for $i \in 1 \ldots d$ **do**
    $W_i \leftarrow Store\_Similar\_Words(x_i, \theta,$ part-of-speech$(x_i))$
end for
Initialize $Z = (z_1, \ldots, z_n)$ as n copy of x
Initialize C set of valid counterfactuals
while number of words modified < d **and** C is empty **do**
    for $j \leftarrow 1$ **to** n **do**
        $k = random(0, d)$ *// Ensure that $z_{j,k}$ has not been already modified*
        Replace $z_{j,k}$ with a word randomly taken from W_k
        if $f(x) \neq f(z_j)$ **then**
            $C \leftarrow z_j$
        end if
    end for
end while
return C the set of valid counterfactual instances

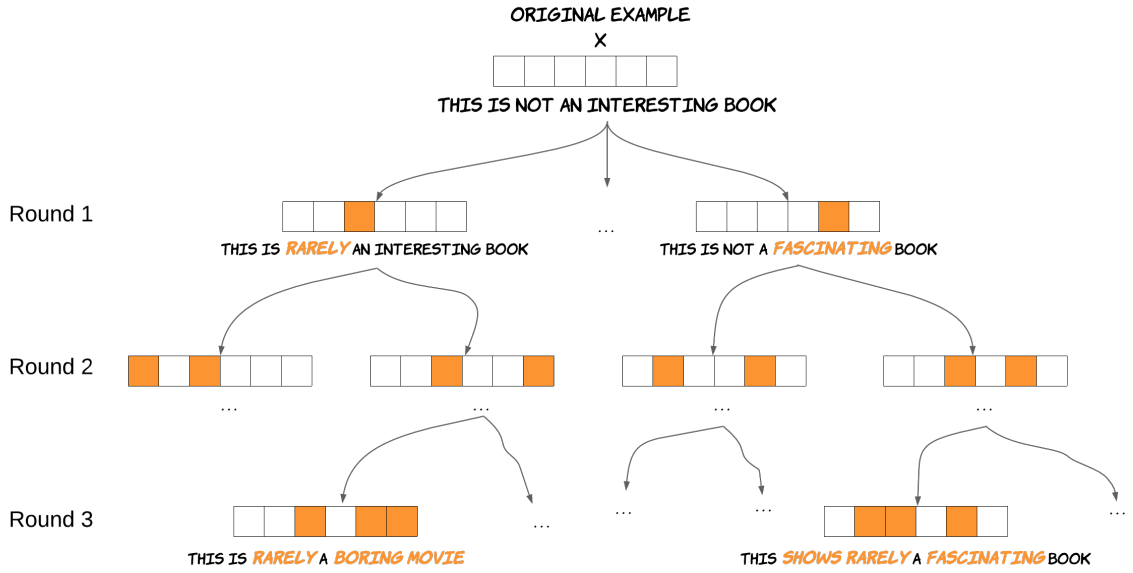


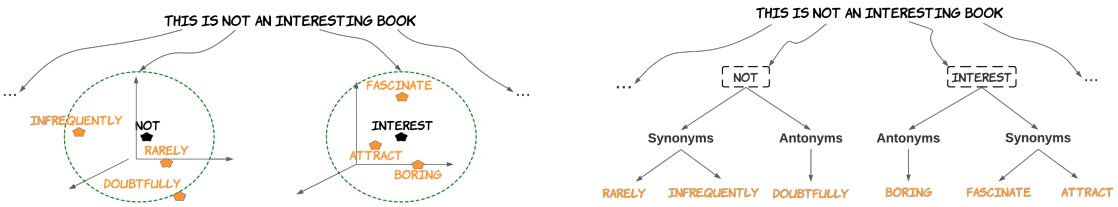
Figure 3.3 – The tree structure of the algorithm used to iteratively perturb the target document. At each round, a word from the target text is iteratively replaced by a word from its corresponding set of potential replacement words. Thus, with each successive round, the number of word replacements for generating artificial documents increases.

progressively replace individual words (x_j) with randomly selected words from their respective sets of potential replacements (W_j).

For example, consider the target review, *"This is not an interesting book"*, classified as negative by a sentiment analysis model (Figure 3.3). In the first round, both Growing Language and Growing Net generate artificial documents with only one modified word. Subsequent rounds involve the replacement of two words and so on. In this process, counterfactuals are identified, and the closest one is returned as the explanation. These methods prioritize counterfactuals closely related to the original document to provide concise and meaningful explanations.

3.3.2.1 Growing Net

Growing Net capitalizes on the rich structure of WordNet [45] to construct sets of closely related words. WordNet is a lexical database and thesaurus that organizes words and their meanings into a semantic tree of interrelated concepts. Therefore, Growing Net begins by creating sets of similar words for each term within the target document through WordNet as illustrated in Figure 3.4b. These sets are constructed based on the part-of-speech tags, such as verbs, nouns, determiners, and more, which ensures that the added words share the same



(a) Growing Language (b) Growing Net

Figure 3.4 – The mechanism employed to generate sets of potential word replacements in Growing Language involves embedding words from the target text into a latent space and capturing nearby words in this space. Conversely, Growing Net leverages the tree structure of WordNet to create sets of potential word replacements.

grammatical roles as the original terms. Additionally, Growing Net identifies words that are close in the WordNet hierarchy, which includes synonyms, antonyms, hyponyms, and hypernyms. Hyponyms are words or phrases that are more specific than the target word (x_i), while hypernyms are more general.

Subsequently, Growing Net modifies the target document by randomly substituting words in the target sentence with words from their corresponding sets. Growing Net systematically increases the gap between the artificial instance and the original document with each successive round. Consequently, at each round, Growing Net expands the number of modified words (k) until it identifies counterfactuals. Ultimately, Growing Net returns the counterfactual with the smallest Wu-Palmer Similarity (Wu-P) distance [152] as the final explanation. The Wu-Palmer Similarity is a similarity metric specifically designed for measuring the relatedness of concepts in WordNet. It considers the depth of terms in the WordNet hierarchy and the path length to their most common ancestor. In this context, the advantage of using the Wu-Palmer Similarity is that it captures not only surface-level similarity but also the depth of relatedness in the WordNet hierarchy. This is crucial because it ensures that the substituted words are not just superficially similar but also conceptually relevant to the original terms.

3.3.2.2 Growing Language

Growing Language leverages the power of large language models to restrict the space of possible perturbations. Large language models are powerful natural language processing AI systems. They are employed in this context to embed words into numerical representations, often referred to as a latent space. In simpler terms, a latent representation consists of high-dimensional vectors that capture the underlying semantic and contextual information of the

original text. This high-dimensional space facilitates the measurement of similarity between words. In the context of Growing Language, individual words from the target document are projected onto this latent space. Within this latent space, Growing Language assembles sets of candidate replacement words for each word in the original document, as illustrated in Figure 3.4a. To qualify for inclusion in these sets, words must fulfill two criteria. Firstly, they should share the same part-of-speech tags as the original words, ensuring grammatical consistency. Secondly, they must exhibit a similarity score that surpasses a predefined threshold (θ). The similarity score is computed by comparing word vectors in a high-dimensional space, where each dimension represents semantic and contextual information of the words, and the score is based on the similarity between these vectors. In our experiments, we set this threshold to 0.8 on a scale from 0 to 1. This allows Growing Language to maintain computational efficiency, as starting with a low threshold could result in longer processing times. Once these sets are established, Growing Language initiates the generation of artificial documents.

The process of generating artificial documents is iterative, with Growing Language progressively substituting more words in each successive round. This iterative process continues until an artificial document is classified differently by the black-box model, effectively finding a counterfactual. Growing Language also incorporates an adaptive mechanism to handle scenarios where replacing all the words in the original document does not produce a valid counterfactual. In such instances, Growing Language proceeds to extend each set of similar words by decreasing the similarity threshold. This threshold reduction occurs iteratively and is set at two times the initial step size, which we initialize at 0.01. Consequently, the sets of replacement words become larger, containing more words that are less similar, allowing for an expanded search for counterfactuals. Should multiple counterfactuals be found, Growing Language selects the one with the fewest modifications compared to the original document. Importantly, any language model capable of converting words and measuring word similarities could be used in this process.

3.4 Adapted Post-hoc Explanations

We now elaborate on APE, our approach to compute adapted post-hoc explanations for a target instance x and a black-box classifier f. When the decision frontier of f admits a single local linear surrogate according to our problem statement in Section 3.2, APE returns a linear-based explanation complemented with a counterfactual explanation. Otherwise, APE recommends a different explanation paradigm such as a rule-based surrogate.

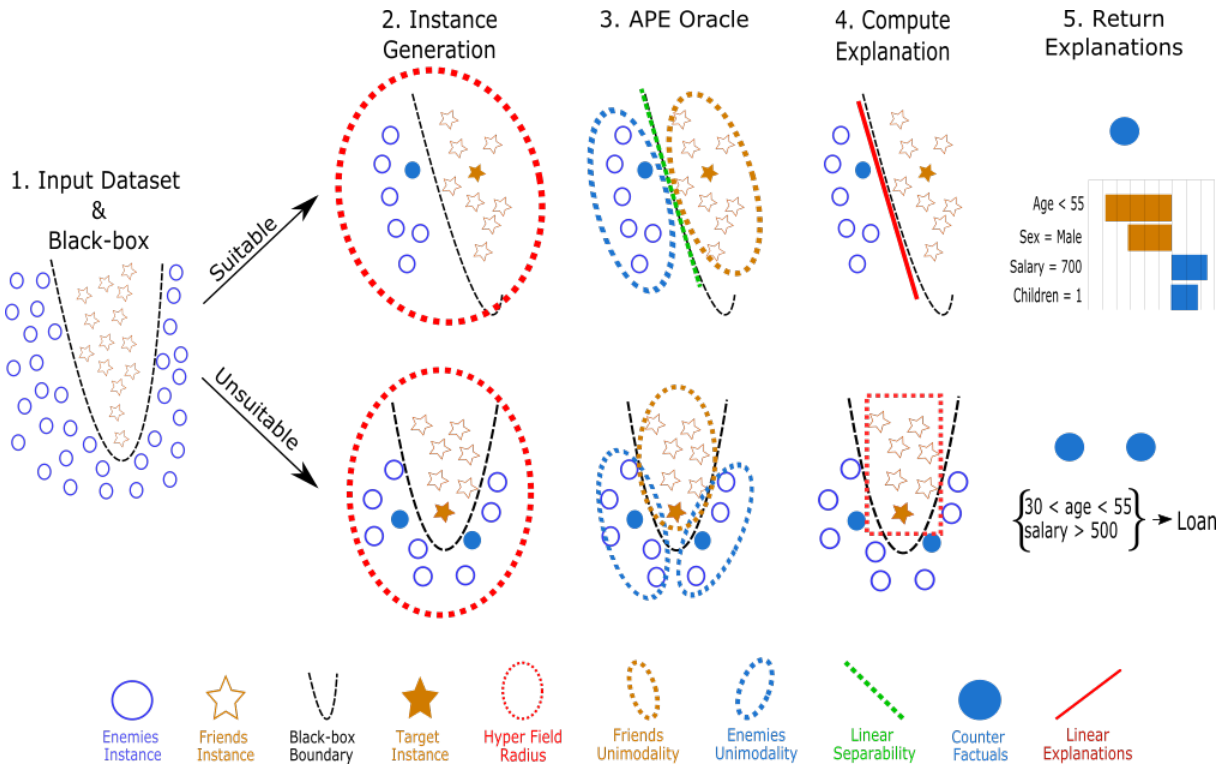


Figure 3.5 – The Adapted Post-hoc Explanation framework. In the first row, the Oracle's test indicates that a linear explanation is suitable, resulting in the return of a counterfactual and a linear explanation. Conversely, in the second row, the Oracle determines that a linear explanation is unsuitable, leading APE to generate a counterfactual for each cluster center along with a rule-based explanation.

APE is detailed in Algorithm 4 and Figure 3.5. In the first stage (line 1), APE invokes an algorithm to find the black-box decision boundary such as *Growing Fields* for tabular data or *Growing Net* and *Growing Language* for textual data. This is achieved by identifying x's closest enemy – denoted by e. Then, APE generates a set of random instances Z uniformly distributed in a locality around e (line 3). This locality constitutes a *field*, which APE samples using the $\mathcal{F}$ generation process already explained in Section 3.3.1. The size of the field depends on a radius parameter that is proportional to $dist(x, e)$, i.e., the distance between x and its closest enemy. More precisely, we set $r = {}^{1}/_{\delta} \times dist(x, e)$, where δ is the farthest distance from x to a real instance in T, i.e., f's training set. By normalizing the radius, we (a) provide users with a clear notion of distance, and (b) reduce the risk of sampling instances beyond the limits of the attribute domains. By centering the generative process at e with radius r, APE makes sure that Z covers x and contains diverse subsets E and F of friends and enemies of x – in concordance with the requirements (i) and (ii) in the problem statement in Section 3.2. The $\mathcal{F}$ generation procedure as well as the Growing Fields algorithm are detailed in Section 3.3.1.

Algorithm 4 APE

Require: a training dataset $T \subset X$, a target instance $x = (x_1, \ldots, x_d) \in X$, a black-box classifier $f : X \to Y$; number of samples n
Ensure: one or multiple counterfactual instances, a surrogate classifier g

1: $e \leftarrow$ GROWING FIELDS(T, x, f)
2: $r \leftarrow 1/\delta \times dist(x, e)$ // δ is the largest distance in T
3: $Z \sim \mathcal{F}(T, r, e)_{i \leq n}$
4: **if** APE ORACLE(Z, x, f) **then**
5: **return** e, $LS_{APE}(Z, f, x, e)$ trained on e-centered field of radius $r' \geq r$
6: **else**
7: **return** $\{e_1, \ldots, e_k\} \subset Z$, RULE-BASED SURR.$(f)$
8: **end if**

In the next step (line 4), APE characterizes the decision boundary of f. To this end, the algorithm invokes the APE Oracle (Section 3.4.1), which runs efficient unimodality and linear separability tests [139, 145] on E and F to determine whether a linear surrogate is suitable or not. The Oracle recommends a linear explanation if both sets E and F exhibit an unimodal distribution, that is, if there is only one cluster per class and we can separate those clusters with a single linear surrogate. In that case, APE returns a linear explanation and the closest enemy of x as a counterfactual explanation for $f(x)$. The linear explanation is learned via an extension of Local Surrogate [83], called LS_{APE}, applied on a superset of Z, consisting of real and artificial instances. Those instances constitute a field with a radius of at least r. We elaborate on those details in Section 3.4.2.

When the APE Oracle deems linear explanations unsuitable, APE proposes a rule-based surrogate. This happens when the instances in E or F form multiple clusters, or because Z is not linearly separable. Rule-based alternatives are Anchors [124] or shallow decision trees. In the first case, the user obtains a single rule of the form:

$$R : \mathbf{B} \Rightarrow f(\eta^{-1}(z)) = f(x) \qquad \text{with } \mathbf{B} = \bigwedge_{j \in F \subseteq \{1,\ldots,d'\}} z[j]$$

where $\eta : \mathcal{X}^d \to \{0,1\}^{d'}$ is a conversion function into a surrogate interpretable space and the left-hand side (or antecedent) of the rule is a conjunction of conditions that predicts $f(x)$. This rule R guarantees a precision $prec(R) = P(f(\eta^{-1}(z)) = f(x) \mid \mathbf{B} \wedge z \in \mathcal{Z}) \geq \tau$ for a user-defined precision threshold τ [124]. In the second case, the user gets a decision tree trained on a superset of Z, consisting of real and artificial instances. Since the decision boundary may consist of several disconnected instance clusters, APE completes its explanation

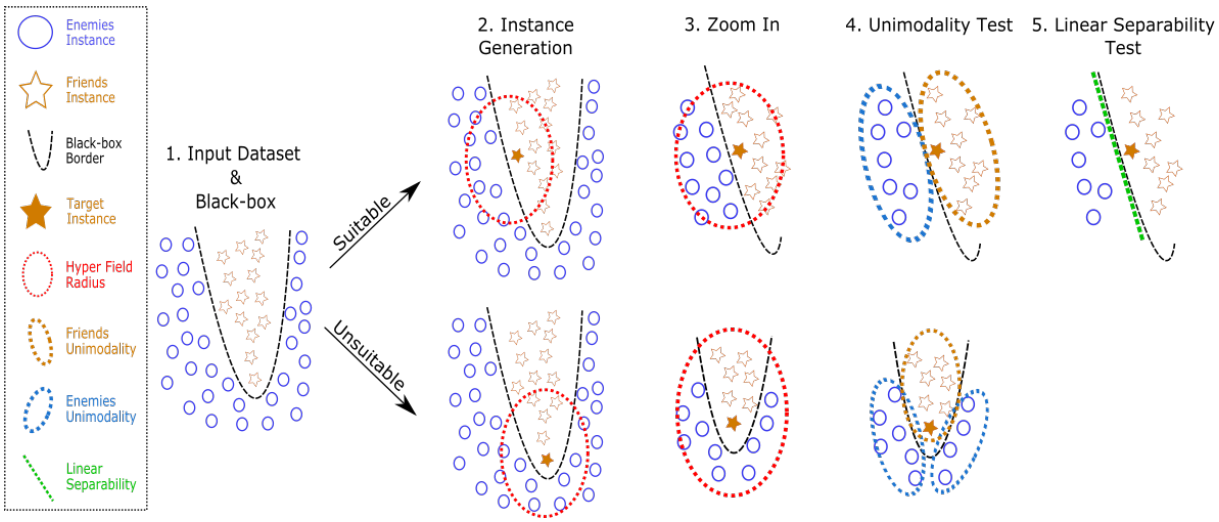


Figure 3.6 – A description of the APE Oracle that assesses the suitability of a linear surrogate for approximating a classifier locally. In the first row, a linear explanation is deemed suitable as friends and enemies of the target instance are grouped into a single cluster. In contrast, the second row shows that a linear surrogate is inadequate due to the enemies being distributed across two clusters.

with a counterfactual instance per cluster in E (see Section 3.4.3). That way users can have a comprehensive view of the different ways to change the black box's outcome $f(x)$.

Since we have already covered two of APE's foundational components, the instance generation process and the counterfactual methods, we will now elaborate on the APE Oracle and the procedures to compute the linear and rule-based surrogates.

Algorithm 5 APE Oracle

Require: instances $Z \subset X$, target instance $x = \{x_1, \ldots, x_d\} \in X$, classifier $f : X \to Y$
Ensure: Is f linearly separable in Z w.r.t. the class $f(x)$?

```
if E ⊂ Z and F ⊂ Z are unimodal then
    if Z is linearly separable w.r.t. f then
        return True
    end if
end if
return False
```

3.4.1 APE Oracle

The core of the APE algorithm is the APE Oracle described in Algorithm 5 and Figure 3.6. This Oracle determines whether the black-box decision boundary is separable by a single linear approximation. To achieve this, the Oracle applies the Libfolding unimodality test [139]

separately on the sets of friends F and enemies E of the target instance x in Z. If the test is passed, it means that F and E form each a single cluster in Z.

Various methods exist for converting textual data into numerical representations. One approach involves using vectorized representations like the 'bag of words' [63] or term-weighting [131] techniques. Another approach transforms text into latent representations, making use of methods such as AutoEncoder [9] and word2vec [100]. In our case, we adapt the Libfolding test to accommodate these two text representation strategies. The first variant transforms text into a numerical vector using a bag of words vectorizer. The second variant, on the other hand, employs a sophisticated language model [67] to embed text into a latent space. These transformed text representations are then used as inputs for the Libfolding test, enabling us to evaluate the unimodality around the target text.

This, however, does not suffice for linear separability; ergo the Oracle carries out a quick linear separability test to determine whether these clusters of friends and enemies can be told apart with a linear approximation. The test is carried out on a balanced sample $Z_b \subseteq Z$. We enforce Z_b to contain an equal number of friends and enemies of x because Z can be highly imbalanced towards the enemies of x for very small localities.

There are multiple methods to determine whether there exists a linear function that separates a two-class dataset. Such methods range from linear and quadratic programming to approaches based on computational geometry and neural networks [41]. Nevertheless, all these strategies are at least as expensive as running a linear regression on the input dataset. On those grounds, APE tabular resorts to a simple test based on the Thornton's *separability index* si [145]. If $\Gamma_{X'}(x)$ returns the closest neighbor $x^{'}$ of x in a set $X^{'} \subseteq X$, the separability index measures the ratio of instances for which that closest neighbor is a friend of x. In our setting, this can be computed according to the following formula:

$$si(X^{'}) = \frac{\sum_{x' \in X'} \mathbb{1}_{f(\Gamma_{X'}(x'))=f(x')}}{|X'|}.$$

We remark that si lies between 0 and 1 and that higher values denote higher separability. Line 2 in Algorithm 5 checks if $si((T \cap \Phi_x) \cup Z_b) = 1$. That is, the test also considers real instances that fall within the field from which Z was drawn. If the test is passed, the decision boundary is considered linearly separable enough and the Oracle returns true.

3.4.2 Linear Explanations

If the APE Oracle estimates that f's decision boundary is linearly separable around the target instance x, APE resorts to the routine LS_{APE} (described in Algorithm 6) to learn a linear surrogate g on Z and to explain $f(x)$. We could center the generative process to learn g on the target instance x as in standard LIME, or around the decision boundary as in LS. We opt for the latter alternative since LS has been shown to identify more accurately the features that influence the black box locally [83].

We recall that Z is a sample drawn from a field centered on e with radius $r = {}^{1}/_{\delta} \times \mathit{dist}(x, e)$ where e is x's closest enemy. We could therefore learn g from the instances used for the linear separability test because these are exactly what LS needs for training. We highlight, however, that nothing prevents our linear surrogate from attaining a good adherence in larger scopes. In concordance with our maximality requirement (Section 3.2), LS_{APE} carries out a posteriori expansion of the training field before reporting the linear explanation to the user. While the adherence does not decrease, that is while $m(g) \geq \tau$, LS_{APE} extends the field radius and trains a new linear explanation (line 5-10 in Algorithm 6). The threshold τ is set to the adherence of g in the initial field. We recall that the function m serves as a metric for quantifying the adherence of a surrogate relative to a model within a specific locality and can encompass metrics such as accuracy, precision, or recall. The radius is increased using the same expansion strategy of Growing Fields (lines 8-9 in Algorithm 2). Besides, we precise that we do not decrease the radius of the field because we want to guarantee the presence of the target instance within the field.

3.4.3 Rule-based Explanations

If the decision frontier in the vicinity of our target instance is too complex to be approximated with a single linear surrogate, users may apply clustering techniques on the neighborhood Z and provide different linear explanations for each of the instance clusters at the decision boundary. This would provide a complete picture of the black box behavior around the target. However, such an explanation is potentially difficult to grasp for users, because it might consist of potentially contradicting feature-attribution rankings. On those grounds, APE proposes by default a rule-based explanation when linear surrogates are considered unsuitable. Alternatives are anchors or shallow decision trees. Anchors [124] learns a single explanation rule of the form $p \Rightarrow f(x)$ such that p is a conjunction of conditions of maximal coverage and the rule has a precision of at least τ. The decision tree is learned on the set Z containing both friends F and

Algorithm 6 EXTENDED LOCAL SURROGATE (LS_{APE})

Require: instances $Z \subset X$ drawn from a field, a classifier $f : X \to Y$,
target and counterfactual instance $x, e \in X$, an adherence metric m;
Hyper-parameters: $\theta = 0.05$
Ensure: a linear surrogate classifier g

$r \leftarrow {}^{1}/_{\delta} \times dist(x, e)$
Split Z into Z_{train}, Z_{test}
$g \leftarrow$ LINEAR REGRESSION$(Z_{train}, f(Z_{train}))$
$a \leftarrow \tau \leftarrow m(g)$ on Z_{test}
while $a \geq \tau \wedge r < 1$ **do**
    $r \leftarrow \theta \times r$
    $Z \sim \mathcal{F}(T, r, e)_{i \leq n}$
    Split Z into Z_{train}, Z_{test}
    $g \leftarrow$ LINEAR REGRESSION$(Z_{train}, f(Z_{train}))$
    $a \leftarrow m(g)$ on Z_{test}
end while
return g

enemies E of x in the field centered on e, the closest enemy of x. We remark, nevertheless, that our framework could be coupled with other explanation approaches [94, 59, 31]. This is an interesting avenue for future research.

Finally, APE complements the rule-based explanation with a set of counterfactual instances $\{e_1, \ldots, e_k\} \subset E^*$. These are the centroids of the clusters defined by an extended set of enemies $E^* \supseteq E$ (generated using the $\mathcal{F}$ generation process from Algorithm 1). This set can be obtained by increasing the field ratio r while the precision of the explanation is above τ. The clusters are computed using K-means [70] and the number of clusters k is determined using the Elbow method [144].

3.4.4 Illustrative Example

We further motivate the utility of APE through an example drawn from the *Moons* dataset consisting of 2 features as shown in Figure 3.7. The application of the Libfolding unimodality test on the set of closest enemies E, represented by red circles surrounding the target instance $x = [0.05, 0.29]$ reveals a multimodal distribution. The k-elbow method identifies three enemy clusters whose centers are: $z_1 = [-0.91, 0.25]; z_2 = [0.01, 0.76]$ and $z_3 = [0.96, 0.19]$. Subsequently, we apply LS_{APE} on those counterfactual instances, serving as centers for the generative process to learn linear explanations. Our observations reveal contradictory explanations since the attribution of the first feature for z_1 is 0.79 whereas for z_3 it is -0.53.

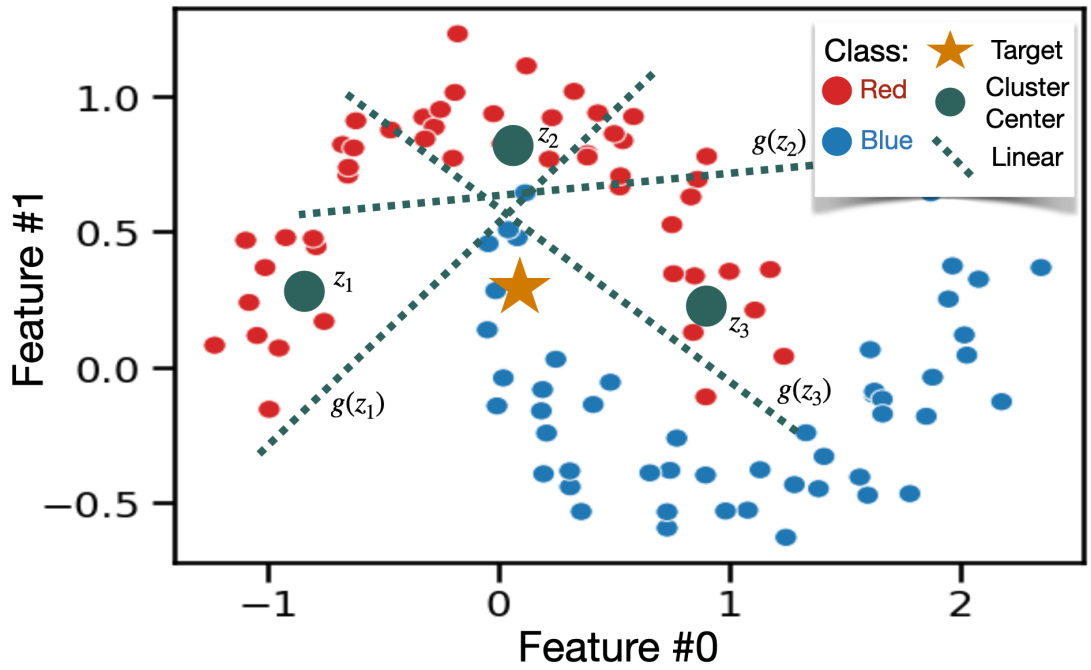


Figure 3.7 – An illustrative example of the moon dataset. The yellow star depicts the target instance while the green circles z_1, z_2, and z_3 represent the cluster centers of the counterfactuals. Corresponding linear explanations are indicated by the dashed lines of the same color.

We also demonstrate examples of the explanations provided by APE using the *Titanic* dataset. In this dataset, the task is to predict whether a passenger of the Titanic survived or not. We consider two instances, x_1 and x_2 with the following attributes values:

$$x_1 = \{age = 42, class = 1, sex = F, siblings = 0, children = 3\},$$
$$x_2 = \{age = 25, class = 2, sex = M, siblings = 1, children = 0\}.$$

A multi-layer perceptron classifier predicts $f(x_1) = Survived$ and $f(x_2) = Died$. APE Oracle determines that a linear explanation is suitable to approximate the model in the locality of x_1. Thus, it provides the following explanations g_1:

$$g_1 = \{\{e\},\ g(x)_{survived} = 0.5 \times sex + 0.2 \times class + 0.2 \times children\},$$

with the counterfactual explanation (only perturbed attributes shown) $e = \{\ldots, sex = M, \ldots\}$. This explanation suggests that changing the individual's gender is sufficient to predict survival, highlighting gender as the most important factor. The linear explanation also indicates that the number of brothers and sisters does not impact the model prediction.

Conversely, APE Oracle determines that the region around x_2 is not linearly separable, and a rule-based along with counterfactual explanations would be more appropriate. Therefore, APE produces the explanation g_2:

$$g_2 = \{\{e_1, e_2, e_3\},\ sex = M \wedge children = 0 \Rightarrow Died\},$$

with counterfactuals $e_1 = \{\dots, sex = F, \dots\}$, $e_2 = \{\dots, children = 2, \dots\}$, and $e_3 = \{\dots sex = F,\ children = 1, \dots\}$. The rule-based explanation specifies the conditions under which the model confidently predicts that an individual died on the Titanic. In contrast, the counterfactual explanations reveal that, for instance, if the individual were a female, as depicted by e_1, or had two children, as suggested by e_2, the model would have predicted survival.

3.5 Experiments

We executed three series of experiments to evaluate the performance of APE on tabular and textual data:

— In the first set of experiments (Section 3.5.2), we focused on the APE's Oracle, specifically on its ability to identify when a linear explanation provides an accurate approximation for a given black-box classifier and target instance.

— In the second set of experiments, we compared the explanations generated by APE with those produced by LIME [123] and LS [83] in terms of adherence (Section 3.5.3).

— The third set of experiments (Section 3.5.4) involved an ablation study of the APE Oracle's two components to assess their impact on the adherence of APE.

Then, we compare in Section 3.5.6, the quality of the counterfactual generated by Growing Fields with those output by Growing Spheres [82]. The source code of Growing Fields, APE, as well as the experimental datasets and additional results, are available on Github [2].

2. `https://github.com/j2launay/APE`

3.5.1 Experimental Setup

3.5.1.1 Datasets

Table 3.1 describes our tabular experimental datasets. The list comprises 6 real and 6 synthetic datasets, the latter generated with scikit-learn [3]. Five of those synthetic datasets contain only numerical features. The real datasets were chosen to provide a mix of numerical and categorical features. For the experiments on textual data, we used four datasets described in Table 3.2. This includes three real-world and one dataset created by combining two distinct sets of articles, one with real articles [4], and one with fake articles [5]. Each of these datasets was designed for binary classification tasks, except for the Ag News dataset, which features four target classes. Moreover, it is worth noting that a multi-class classification problem can always be redefined as a collection of binary classification problems, with one binary problem per class.

Name	Features		Instances	Models				
	Numerical	Categorical		GB	MLP	RF	VC	SVM
Adult	2	10	48842	86%	84%	84%	79%	71%
Blob †	2	0	1000	94%	93%	94%	94%	94%
Blobs †	12	0	5000	97%	98%	97%	98%	99%
Blood	4	0	748	74%	62%	75%	50%	44%
Cat Blobs †	4	4	5000	97%	98%	98%	99%	97%
Cancer	10	20	569	97%	94%	98%	98%	92%
Circles †	2	0	1000	94%	95%	94%	96%	97%
Diabetes	8	0	768	70%	66%	74%	74%	71%
M Blobs †	20	0	7500	99%	99%	99%	99%	100%
Moons †	2	0	1000	97%	88%	97%	94%	97%
Mortality	15	52	1614	64%	69%	67%	66%	69%
Titanic	1	5	1046	83%	87%	84%	87%	61%

Table 3.1 – Number of instances, numerical and categorical features for each dataset. The † indicates that datasets are synthetic. The last five columns represent the black-box model's test set accuracy.

3.5.1.2 Black-box Classifiers

We assessed APE's performance on a variety of classifiers with different architectures – i.e., ensemble methods, piecewise-constant functions, smooth functions – implemented in scikit-learn [116] with default values for the hyperparameters unless stated otherwise. For tabular

3. `http://scikit-learn.org`
4. `https://www.kaggle.com/datasets/rmisra/news-category-dataset`
5. `https://www.kaggle.com/competitions/fake-news/overview`

Name	Nb Words			Instances	Model			
	Total	Average	STD		NN	RF	NB	BERT
Polarity	11646	20.8	9.3	10660	72%	67%	72%	83%
Fake †	19419	11.8	3.2	4025	84%	84%	87%	91%
Ag News	92806	38.8	11	12000	86%	87%	90%	97%

Table 3.2 – Information about the experimental datasets with textual data. † indicates generated datasets. The three columns under "Nb Words" represent respectively (a) the total number of distinct words in the whole dataset, (b) the average number of words per sentence, and (c) the standard deviation. The fourth column indicates the number of text documents per dataset. The final columns show the average accuracy of the different complex models.

data, the classifiers included: (i) Gradient Boosting (GB) with 20 tree estimators, (ii) Multi-layer Perceptron (MLP) with a logistic activation function, (iii) Random Forest (RF) with 20 tree estimators, (iv) Gaussian Naive Bayes (NB), (v) Support Vector Machine (SVM) with a balanced class weight, (vi) Decision Tree (DT), (vii) Logistic Regression (LR) and (viii) a Voting ensemble (VC) composed of LR, SVM, and NB classifiers. For textual data, the classifiers comprised: (i) Multi-Layer Perceptron (MLP) with 100 neurons and four layers, (ii) Random Forest with 500 trees, (iii) Gaussian Naive Bayes (NB), and (iv) a BERT base model (BERT) implemented in the python package "transformers" [132], fine-tuned on the dataset. In addition to the class of an instance, the classifiers can provide class probabilities. The classifiers were trained on 70% of the data points and their accuracy was tested on the remaining 30%. Table 3.1 presents the accuracy of the tabular models while Table 3.2 exhibits the accuracy of the textual models.

3.5.1.3 Explanation modules

APE and the competitors were tested on a random sample of 100 target instances drawn from the test datasets. All the explanation modules had access to the training set used for training the classifiers (referred to as T in Algorithm 4). APE's evaluation considered two variants: one using Anchors and the other employing shallow decision trees (with a maximum depth of 3) as explanation solutions when linear explanations were considered unsuitable. These variants are denoted as APE_a and APE_t, respectively. Anchors requires a precision goal parameter τ for rules, which we set to 0.95. Nevertheless, the semantics of τ are purely indicative, as the algorithm will always generate an explanation even if the specified precision goal is unattainable in the surrogate's training set. In line with LIME and LS, the training

instances for learning the linear surrogate were labeled with the class probabilities of the target class $f(x)$ as provided by the black-box classifiers.

3.5.1.4 Metrics

To assess the quality of explanations, we employed various metrics. Adherence, which measures how well explanation surrogates match the behavior of the original model, was evaluated by calculating the accuracy score of the surrogate models within the region (e.g., field) on which they were trained. This evaluation used 70% of the generated artificial instances (lines 5 and 7 in Algorithm 4) for training the surrogates, while the remaining 30% were reserved for accuracy evaluation. Explanation fidelity was assessed differently for tabular and textual data. In cases where the actual features used by the input classifier were known, fidelity was measured using precision and the Kendall rank correlation coefficient on the sets of features reported by the explanations. The precision score indicates the proportion of features in the explanation that were effectively used by the black-box classifier. The Kendall coefficient quantifies the agreement between the feature attribution rankings of the explanation and the actual contribution ranking in the black box. For textual methods, explanation fidelity was evaluated by removing words identified as important by the explanation and measuring the resulting change in the black box prediction.

3.5.2 APE Oracle Evaluation

In this section, we evaluate the APE Oracle's ability to determine whether a linear surrogate is appropriate for a given scenario through two key evaluations. Both evaluations involve comparing the adherence and fidelity of linear surrogates learned using LS_{APE} when the Oracle predicts suitability and when it does not. We expect higher adherence and fidelity in the first case.

3.5.2.1 Adherence Evaluation

We computed the adherence (accuracy) of the linear surrogate for each black-box classifier across 100 test instances on our experimental datasets. The surrogates were computed using LS_{APE}. For each target instance, the APE Oracle determines whether or not the decision boundary admits a single accurate linear approximation (Yes or No).

	Is a Linear Explanation Suitable?													
	GB			MLP			RF			VC			SVM	
	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$
Adult	**0.555**	0.486	0.65	**0.507**	0.397	0.60	**0.659**	0.483	0.47	**0.334**	0.304	0.25	**0.679**	0.643	0.35
Blob	**0.891**	0.782	0.57	**0.890**	0.760	0.49	**0.874**	0.730	0.56	**0.899**	0.748	0.46	**0.894**	0.744	0.43
Blobs	**0.855**	0.636	0.78	**0.723**	0.606	0.86	**0.783**	0.655	0.82	**0.745**	0.610	0.68	**0.717**	0.599	0.80
Blood	\	0.437	0.99	\	0.497	1.00	\	0.283	1.00	\	0.223	1.00	\	0.622	1.00
Cancer	**0.502**	0.381	0.20	**0.501**	0.499	0.12	0.510	\	0.00	**0.411**	0.382	0.21	0.499	\	0.02
Cat Blobs	**0.910**	0.898	0.70	**0.958**	0.900	0.86	0.874	**0.958**	0.50	**0.967**	0.936	0.72	**0.883**	0.794	0.48
Circles	**0.945**	0.723	0.09	0.958	\	0.00	**0.950**	0.708	0.04	0.948	\	0.00	0.949	\	0.00
Diabetes	**0.630**	0.399	0.92	**0.802**	0.585	0.96	\	0.453	0.98	**0.673**	0.258	0.96	**0.717**	0.518	0.88
M Blobs	\	0.833	0.97	\	0.967	1.00	**0.863**	0.845	0.82	\	0.947	0.99	**0.944**	0.942	0.71
Moons	**0.923**	0.708	0.55	**0.917**	0.802	0.59	**0.918**	0.727	0.42	**0.916**	0.881	0.85	**0.920**	0.750	0.50
Mortality	\	0.826	1.00	\	1.000	1.00	\	0.839	1.00	\	0.518	1.00	\	0.420	1.00
Titanic	**0.761**	0.667	0.06	0.919	\	0.00	0.973	**1.000**	0.04	**0.999**	0.997	0.16	0.715	\	0.00

Table 3.3 – Average accuracy calculated on 100 instances per black-box model and tabular dataset for LS_{APE} in relation to both Oracle outcomes. The columns labeled "Yes" and "No" represent the average accuracy of LS_{APE} when the Oracle indicates that a linear explanation is suitable or unsuitable. "\" denotes a non-meaningful accuracy score, i.e., there were fewer than 3 instances in that case. The columns labeled $Prop_{no}$ denote the proportion of cases where the Oracle does not predict linear suitability. The colors blue, orange, and red correspond to $Prop_{no} \leq 33\%$, $33\% > Prop_{no} \geq 66\%$, and $Prop_{no} \geq 66\%$ respectively. Each row reports results for a specific dataset, such as "Adult" in the first row.

Tabular data. Table 3.3 presents the mean adherence (accuracy) of the linear surrogates. The results clearly confirm that (i) the modality of the instances at the decision boundary plays a pivotal role in the quality of a linear surrogate, and (ii) APE's linear suitability test is pertinent. Specifically, when the APE Oracle predicts a linearly separable decision boundary, the surrogate's accuracy is on average 0.124 points higher compared to cases where the Oracle predicts the opposite.

Textual data. Table 3.4 reports the average accuracy of LS_{APE} for both possible outcomes of the APE Oracle. The results we obtained demonstrate that the APE Oracle's suitability test also works for textual data. Therefore, we observe that learning linear explanations in line with the recommendations of the APE Oracle test guarantees explanations with higher adherence. This underscores the importance of inspecting the decision boundary before generating an explanation. The success of the APE Oracle in improving explanation quality for both tabular and textual data, suggests its potential practical application in real-world scenarios.

	Is a Linear Explanation Suitable?											
	RF			MLP			NB			BERT		
	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$
Ag News	**0.994**	0.988	0.63	**0.992**	0.987	0.58	**0.992**	0.988	0.65	0.925	**0.965**	0.41
Fake	**0.941**	0.891	0.14	**0.822**	0.819	0.11	0.833	**0.864**	0.11	**0.952**	0.938	0.08
Polarity	**0.808**	0.770	0.36	**0.87**	0.851	0.27	**0.857**	0.851	0.27	**0.965**	0.956	0.41

Table 3.4 – Average accuracy computed on 100 instances per black-box model and textual dataset for LS_{APE} with respect to both oracle outcomes. The columns labeled "Yes" and "No" represent the average accuracy of LS_{APE} when the oracle indicates the suitability or unsuitability of a linear explanation. "\" denotes a non-meaningful accuracy score, i.e., there were less than 3 instances in that case. The columns labeled $Prop_{no}$ indicate the ratio of cases where the oracle does not predict linear suitability. The colors blue, orange, and red correspond to $Prop_{no} \leq 33\%$, $33\% > Prop_{no} \geq 66\%$, and $Prop_{no} \geq 66\%$ respectively.

General Insights. Our observations also reveal that the proportion of linearly separable cases is primarily influenced by the characteristics of the dataset. For example, datasets such as Blood and Mortality exhibit little suitability for linear explanations, while datasets such as Circles, Fake, and Titanic tend to have a higher proportion of cases that are adapted for linear surrogates. However, it is worth noting that the choice of the black-box classifier's architecture can also impact this proportion. For instance, in the Adult dataset, only 25% of the target instances within the Voting Ensemble (VC) are considered unsuitable for a linear explanation, which contrasts with other datasets where this proportion is higher. Interestingly, this pattern differs when we consider the Gradient Boosting (GB) classifier, where 65% of the target instances are deemed unsuitable for linear explanations according to the Oracle.

Moreover, we observe that even in cases where the Oracle rejects linear suitability, the adherence of the linear surrogate can still be high. For instance, the Cat Blobs dataset with the GB black box exemplifies this phenomenon. This can be attributed to the fact that multimodal scenarios characterized by clustered data, may still exhibit a degree of linear separability if the individual clusters mostly consist of instances from the same class. In such cases, the APE framework tends to favor rule-based explanations accompanied by multiple counterfactual instances. This preference for rule-based explanations, along with the production of several counterfactuals, demonstrates various ways to alter the classifier's prediction. This highlights the complex nature of the decision boundary in the locality of the target instance. Consequently, APE adopts a two-step process, first testing for unimodality and then assessing linear separability, to effectively address these diverse scenarios.

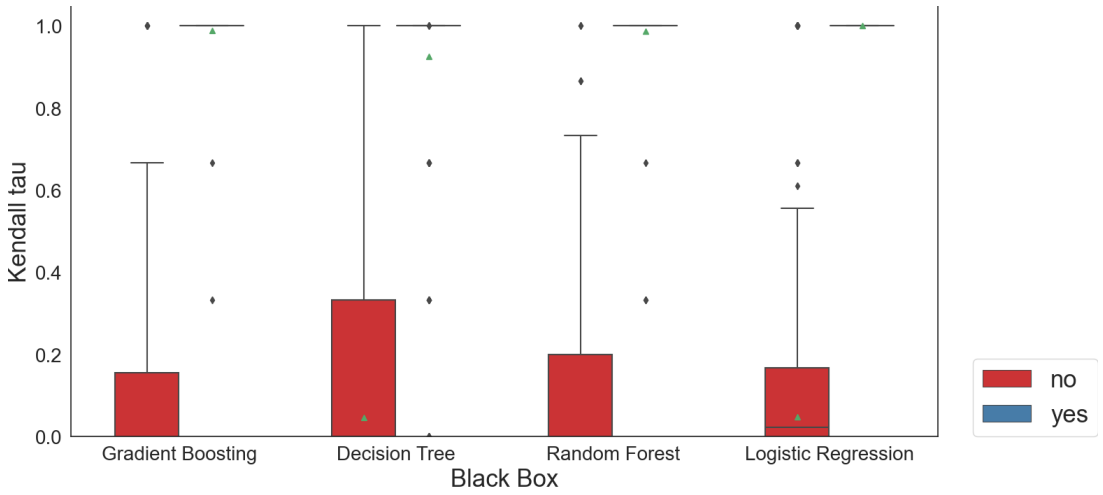


Figure 3.8 – Average Kendall's rank correlation coefficient of the LS_{APE}'s explanations computed on 100 instances for 7 tabular datasets and 4 "glass" black-box models across the Oracle's outcomes.

3.5.2.2 Fidelity Evaluation

This evaluation aims to show that when the APE Oracle indicates the suitability of a linear explanation, then this explanation faithfully reflects the inner workings of the black box. A good explanation is one that relies solely on features genuinely used by the complex model. To gauge the fidelity of linear surrogates under both potential outcomes of the APE Oracle, we employ two distinct metrics. For tabular models, we resort to a set of "glass" black-box classifiers, essentially white-box classifiers treated as black boxes, where we have control over the features used for predicting the class of any instance. Conversely, since it is more difficult to control the words used by a prediction model on textual data, we assess explanation fidelity through the insertion and deletion score. This score measures the difference between the black box model's prediction of the original text and the same text where the words considered as important by the explanation are removed.

Tabular data. Our evaluation on tabular data focuses on datasets with a minimum of 8 features. We train classifiers on modified datasets where half of the features were set to 0 for all instances. Those features were randomly chosen. For example, in the context of the Blobs dataset with 12 features, our Decision Tree model considers only 6 of these features, as the others do not contribute to the classification. When evaluating a linear explanation, we consider it a satisfactory explanation if it employs the same features as the "glass" black box,

	Is a Linear Explanation Suitable?											
	MLP			NB			RF			BERT		
	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$
Ag News	**0.505**	0.342	0.63	**0.3**	0.286	0.58	0.012	**0.061**	0.65	**0.563**	0.206	0.41
Fake	**0.398**	0.361	0.14	**0.268**	0.225	0.11	**0.125**	0.123	0.11	**0.244**	0.233	0.08
Polarity	**0.323**	0.283	0.36	**0.199**	0.184	0.27	**0.076**	0.071	0.27	**0.244**	0.223	0.41

Table 3.5 – The average difference in the model prediction's probability between the original text and the same text without the words identified as important by LS_{APE}. This score is computed based on 100 instances per black-box model and dataset, taking both Oracle outcomes. The columns labeled "Yes" and "No" represent the average change in the prediction probability of the classifier when the Oracle indicates the suitability or unsuitability of a linear explanation. The columns labeled $Prop_{no}$ denote the ratio of cases where the Oracle does not predict linear suitability. The colors blue, orange, and red indicate $Prop_{no} \leq 33\%$, $33\% > Prop_{no} \geq 66\%$, and $Prop_{no} \geq 66\%$ respectively.

and it is an unsatisfactory explanation if it relies mostly on features not used by our decision tree model. To perform this assessment, we use the upper half of the features according to the ranking provided by LS_{APE}, based on the absolute value of the attribution coefficient, from the linear surrogates.

Figure 3.8 illustrates the Kendall rank correlation coefficient for various classifiers, including Gradient Boosting (GB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). For LR, we use the feature coefficients of the logistic function as the ground truth. Similarly, we extract the ground truth for DT by collecting the features encountered along the classification path of the instances. For the GB and RF classifiers, we construct feature rankings using the Gini importance score [17] provided by scikit-learn.

Our observations indicate that when the APE Oracle indicates linear suitability, the rank correlation is consistently very close to 1. This suggests that LS_{APE} accurately captures the actual feature importance ranking within the complex model. However, when the Oracle discourages linear explanations, LS_{APE} faces challenges in identifying the features employed by the "glass" black-box classifier. These findings underscore the effectiveness of APE's linear suitability test as a reliable indicator of the expected quality of a linear surrogate, which translates into faithful explanations for black-box classifiers. While precision and Kendall's tau metrics yield similar results, we opt for Kendall's tau in Figure 3.8 since it accounts for the ranking order of each feature.

Textual data. We assess explanation fidelity by examining the average changes in the black box model's prediction when the words identified as important by the explanation are removed [117]. We report this mean probability difference in Table 3.5. We consider that an explanation is good if removing a word considered important from the original text decreases the classifier's confidence in the same class. Conversely, an explanation is bad if removing the top words does not decrease confidence or, worse, increases confidence in the classifier's prediction.

Our results consistently demonstrate that judiciously selecting when to employ a linear explanation ensures that the elements highlighted as important by the explanation genuinely influence the model's classification. Notably, for cases where the APE Oracle test indicates the suitability of a linear explanation (column 'Yes'), the average difference between the model's prediction before and after removing the most important words consistently increases. This implies that the words identified as important by the adapted linear explanations have a more significant impact on the model's prediction compared to the top words from non-adapted explanations. This observation holds true for most scenarios, except for the Random Forest classifier on the Ag News dataset. We note that the Ag News dataset exhibits a mean change exceeding 0.5 for both MLP and BERT models when linear explanations are suitable, which means that such explanations effectively impact the model's prediction. These findings underscore the benefit of studying the decision boundary before approximating it with a linear surrogate, as it serves as a reliable indicator of the potential fidelity of the computed explanations for black-box classifiers.

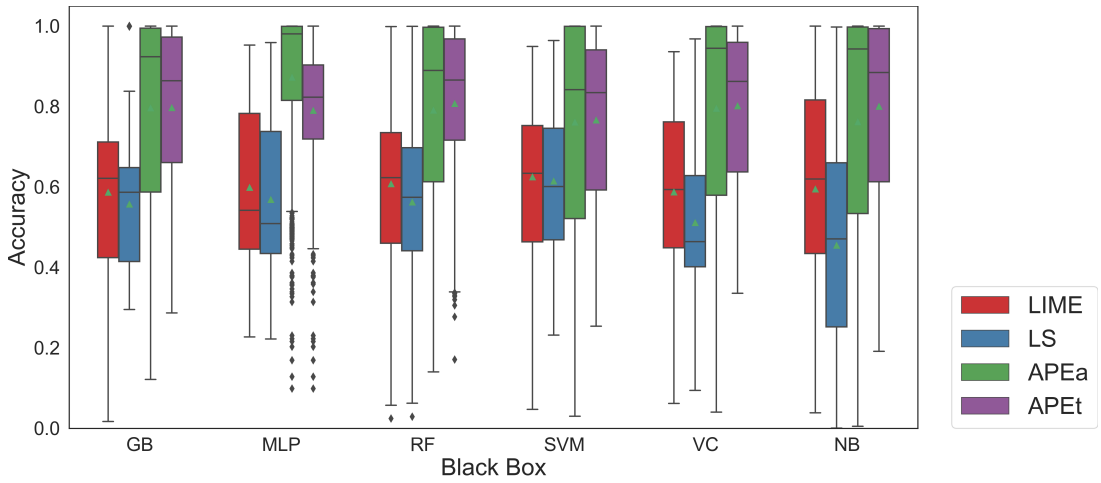


Figure 3.9 – Average accuracy per black-box model on 100 instances of the experimental tabular datasets for APE_a and APE_t.

3.5.3 Comparison with other Explanation Methods

We compare the average accuracy of linear surrogates generated by LIME, LS, and the APE's variants APE_a and APE_t across 100 target instances. We exclude SHAP [94] from this evaluation because it does not compute a linear approximation of the black box such as LIME and LS [49]. Even though its variant Kernel SHAP uses linear regression, this method approximates the Shapley values. APE's variants produce either an anchor or a shallow decision tree when the APE Oracle does not predict linear suitability, otherwise, they both invoke LS_{APE}.

Tabular data. The results are depicted in Figure 3.9. For LS, we exclude the datasets with categorical attributes since they are not supported by this method. The findings suggest that regardless of the rule-based surrogate, APE achieves the highest accuracy. However, we observe that the performance of its two variants depends on the specific characteristics of the black-box model. On average, APE_a offers higher adherence but also exhibits greater variability. All in all, this evaluation demonstrates that the judicious selection between linear and rule-based explanations on a per-instance basis brings an average adherence gain of 0.21 points when compared to always choosing LIME or LS.

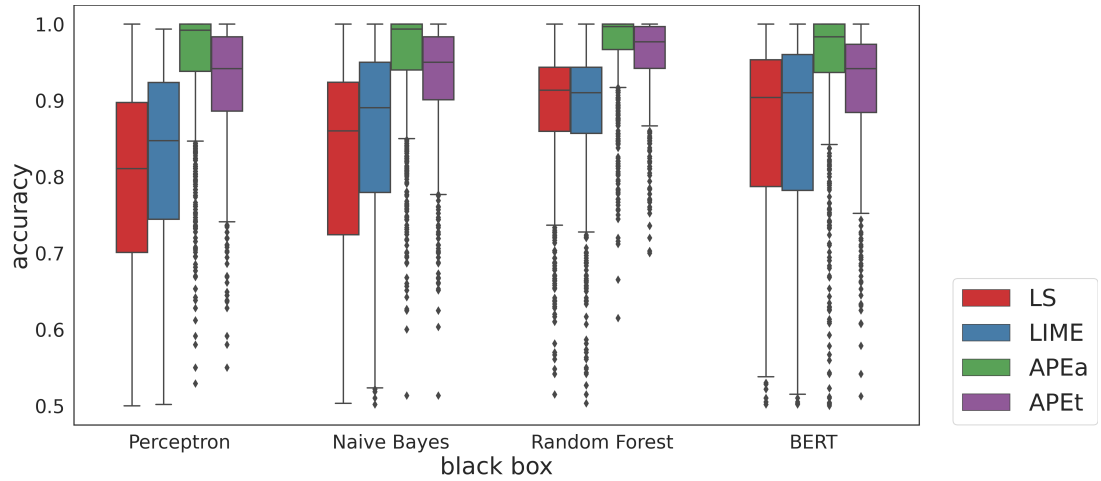


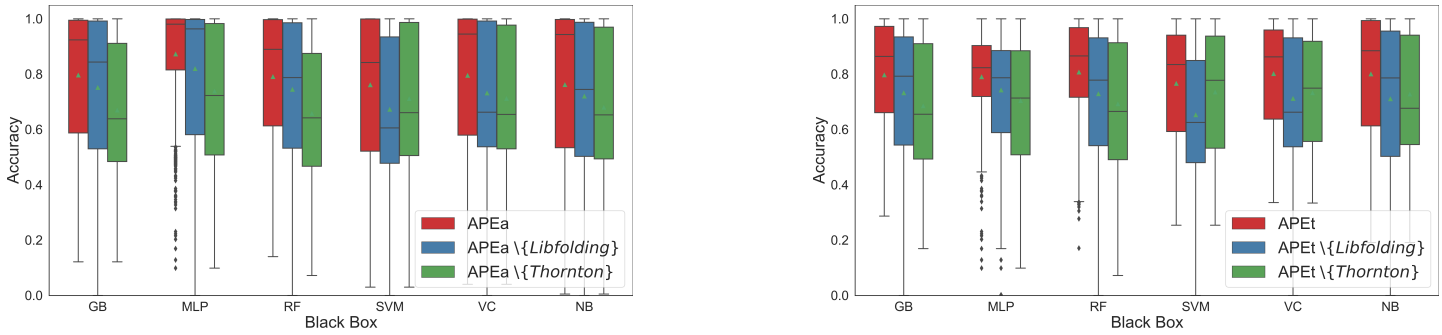
Figure 3.10 – Average accuracy achieved by APE_a and APE_t on each black-box model based on 100 instances drawn from the textual datasets.

Textual data. Figure 3.10 displays the average accuracy of LIME, LS, APE_a, and APE_t. Notably, both methods employing the APE framework consistently outperform LIME and LS

across all complex models. Additionally, we observe that the variance of the techniques restricted to linear surrogates is higher than those of APE_a and APE_t.

General Insights. These findings provide valuable insights, suggesting that when APE chooses to report a linear explanation, the decision frontier is indeed linearly separable. This is supported by the superior performance of both APE_a and APE_t compared to using linear surrogates alone. This observation holds true even for black box models with a relatively high proportion of linearly suitable frontiers, e.g., SVM and tabular RF (see Table 3.3).

While these results may suggest that APE's performance primarily relies on Anchors or decision trees and that we should consistently pick a rule-based surrogate, there are nuances to consider. We provide, in Tables 3.3 and 3.4, the proportion of times APE selects a rule-based explanation over a linear surrogate for tabular and textual data, respectively. We remind the reader that APE Oracle tests the suitability of a linear explanation. Therefore, it cannot predict the performance of an anchor or a rule-based explanation. Thus, in Figure 3.9 and 3.10, we compare the accuracy of both APE_t and APE_a, our two methods that differ only in the choice of the rule-based surrogate model. We observe that APE_a with Anchors enhances the fidelity of the surrogate while using a shallow decision tree reduces standard deviation error. Therefore, the advantages of one rule-based method over another warrant further investigation.



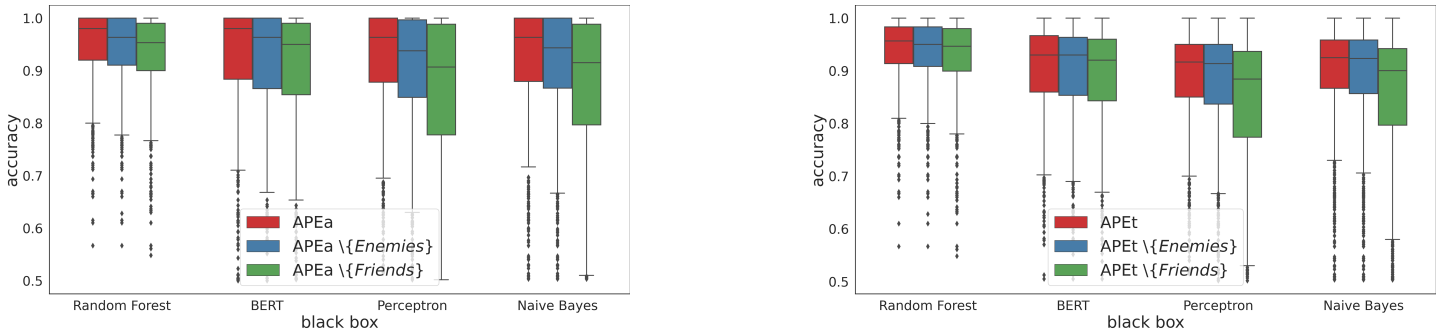
(a) APE_a (b) APE_t

Figure 3.11 – Average accuracy per black box computed on 100 instances of the tabular datasets for APE_a in (a) and for APE_t in (b) when we remove the Libfolding unimodality and linear separability tests from the APE Oracle.

3.5.4 Ablation Study

We now carry out an ablation study to assess the individual contributions of the components of the APE Oracle, in particular the unimodality and separability tests. This study is conducted through the adherence metric.

Tabular data. We report the average accuracy of APE_a and APE_t in Figures 3.11a and 3.11b, compared to the same variant of APE that excludes either the Libfolding unimodality test ($APE \setminus \{Libfolding\}$) or Thornton's separability test ($APE \setminus \{Thornton\}$). The results show that the unimodality and separability tests complement each other, and solely assessing linear separability around the decision boundary, as suggested by the accuracy of $APE \setminus \{Libfolding\}$, is insufficient to predict linear suitability.



(a) APE_a (b) APE_t

Figure 3.12 – Average accuracy attained by APE_a in (a) and for APE_t in (b) across different black-box models and textual datasets. The computation is based on 100 instances sampled from the experimental datasets. The results are displayed for both methods using the Libfolding unimodality test from the APE Oracle, on the friends or enemies instances.

Textual data. In the context of textual data, we conduct two separate ablation studies. The first study is designed to examine the benefits resulting from the implementation of the unimodality test on both the friends and enemies of the target instance. The second study compared how the choice of the data space on which the unimodality test is applied,(either bag of words or latent space) influences the test's effectiveness.

Therefore, we report in Figure 3.12a, the average accuracy of APE_a, compared to its variants that apply the Libfolding unimodality test to specific subsets of instances: either the friends

(APEa\{$Enemies$}) or the enemies (APEa\{$Friends$}). Similarly, we examine the results for APE_t and its respective variants in Figure 3.12b. These findings highlight a critical insight: characterizing the distribution of either the friends or the enemies alone is not enough to detect multimodality. The limitation arises from the fact that analyzing the distribution of one group (friends or enemies) alone does not provide a comprehensive understanding of the decision boundary. As a result, it is crucial to examine the distribution of both friends and enemies before selecting an appropriate explanation technique.

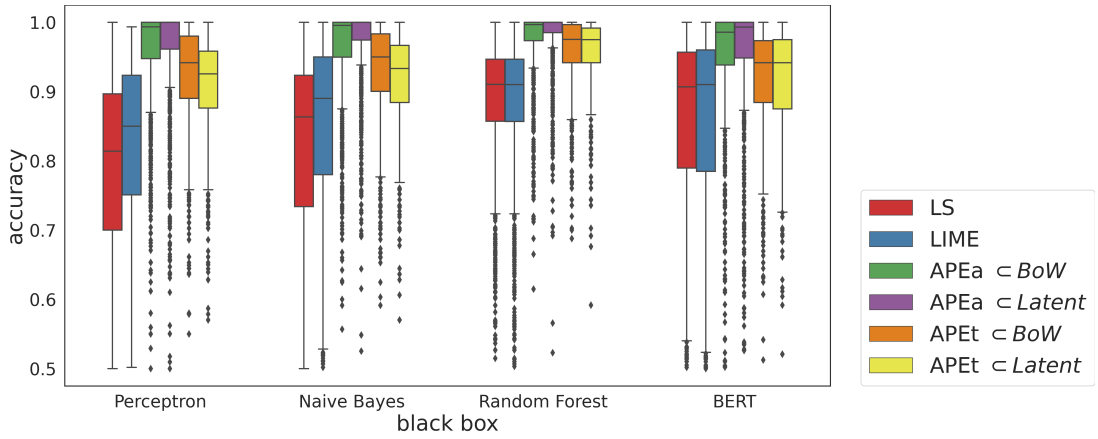


Figure 3.13 – Average accuracy calculated for 100 instances per textual dataset and complex model. We present a comparison between the outcomes of APE_a and APE_t based on the space used for the APE Oracle's test ('BoW' or 'Latent').

In Figure 3.13, we examine the impact of employing different data spaces for the Libfolding test, specifically the bag of words representation (BoW) and the latent space (Latent). Our focus is to understand how this choice influences the performance of two APE variants, APE_a and APE_t, in comparison to traditional linear explanation methods such as LIME and LS. The findings from this analysis reveal that regardless of the space chosen for testing the instance distribution, both APE_a and APE_t consistently outperform linear explanation methods. Nevertheless, the choice of the data space does not have the same effect on APE_a and APE_t, and this effect also depends on the target black box model. To illustrate this point, consider MLP and NB, where the bag of words representation (BoW) enhances the average accuracy of APE_t, while the latent representation increases the average accuracy and reduces the variance of APE_a. With these results, we gain valuable insights into the role of space selection in determining the effectiveness of the APE framework. They underscore the importance of employing adaptable strategies when interpreting textual data.

3.5.5 Summary and Discussion of Linear Suitability Results

The extensive evaluation presented in this section focuses on the concept of "linear suitability" within the APE framework. These evaluations aim to determine whether the APE framework accurately identifies scenarios where linear explanations are most appropriate. Several key findings emerge from the results.

Firstly, the APE Oracle's ability to assess linear suitability significantly impacts the quality of explanations. When the APE Oracle predicts that a linear explanation is suitable for a particular instance, the resulting explanation exhibits higher adherence and fidelity. In essence, these explanations not only have a good adherence but also accurately represent the inner workings of the complex black-box models, leading to more accurate and interpretable results.

Secondly, the evaluation underscores the nuanced nature of linear suitability. It reveals that the APE Oracle's test for linear separability in the decision boundary is a strong indicator of the potential success of linear explanations. In instances where the decision boundary is indeed linearly separable, APE's linear explanations perform exceptionally well, offering accurate and insightful results. These findings hold true even for complex black-box models, such as Support Vector Machines and Random Forests, which can often produce linearly suitable decision boundaries.

Furthermore, the results suggest that the proportion of linearly separable cases can be influenced by dataset characteristics. For instance, datasets like Circles, Fake, and Titanic exhibit a higher prevalence of linearly suitable boundaries, likely due to the nature of the data distribution. However, the choice of the black-box classifier architecture also plays a role, with different classifiers yielding varying proportions of linearly suitable cases.

In conclusion, the APE framework's linear suitability assessment offers a practical and effective method for distinguishing when linear explanations should be applied. This two-step process, involving unimodality and separability tests, allows APE to adapt to a wide range of scenarios, providing a flexible approach to generating interpretable explanations for complex black-box models. By leveraging this understanding of linear suitability, APE demonstrates its versatility and robustness in providing explanations that are both faithful to the model's behavior and accurate in representing the decision boundaries of complex classifiers.

3.5.6 Counterfactuals Evaluation

We now evaluate the quality of the counterfactuals generated by Growing Spheres and Growing Fields for our 100 test instances. Our evaluation considers two key aspects: (a)

the resemblance of these counterfactual instances to real-world examples, and (b) the runtime performance of each method. Note that we detail the experimental results for Growing Language and Growing Net in Section 4.4.

3.5.6.1 Quality of the Counterfactuals

In line with the literature in counterfactual explanations [147], we assess the quality of counterfactual instances generated by Growing Fields and Growing Spheres by measuring their resemblance to actual instances. We resort to the Mahalanobis and Euclidean distances, computed between the generated counterfactuals and the entire set of enemies in the test instances. The Mahalanobis distance quantifies to which extent our counterfactual explanations are outliers w.r.t. the distribution of non-synthetic enemies. For the sake of brevity, we report results based on the Mahalanobis distance since the results for Euclidean and Manhattan distances show the same behavior when comparing Growing Fields and Growing Spheres [82].

Figure 3.14 shows the average distance between the counterfactual generated and the original instance. It is important to note that these distances are normalized with respect to the farthest instance from the input dataset T. As a result, we can consider a counterfactual to be realistic if the distance to the target instance is lower than one, which means it is lower than the farthest observed distance. However, the results show that this is not always the case. Specifically, our findings reveal that, on average, APE (Growing Fields) finds more realistic counterfactual instances than Growing Spheres. These observations hold especially true when examining the multi-layer perceptron classifier (MLP). On this classifier, Growing Spheres finds counterfactual instances that are very unrealistic, as the normalized distance to their closest real counterfactual is far higher than one. The counterfactuals generated by APE tend to be, on average, 0.356 points closer to actual instances. This improved performance of Growing Fields can be attributed to its ability to account for the variance and amplitude of the attributes when generating synthetic instances. Although Growing Spheres occasionally generates counterfactuals that are as similar to the target as Growing Fields, there are cases for which the counterfactuals produced by Growing Spheres are significantly distant from real-world examples, a shortcoming not observed in the case of Growing Fields.

3.5.6.2 Runtime Performance

We present in Table 3.6 the runtime performances of both methods, Growing Spheres and Growing Fields. By leveraging the dataset distribution, Growing Fields efficiently samples

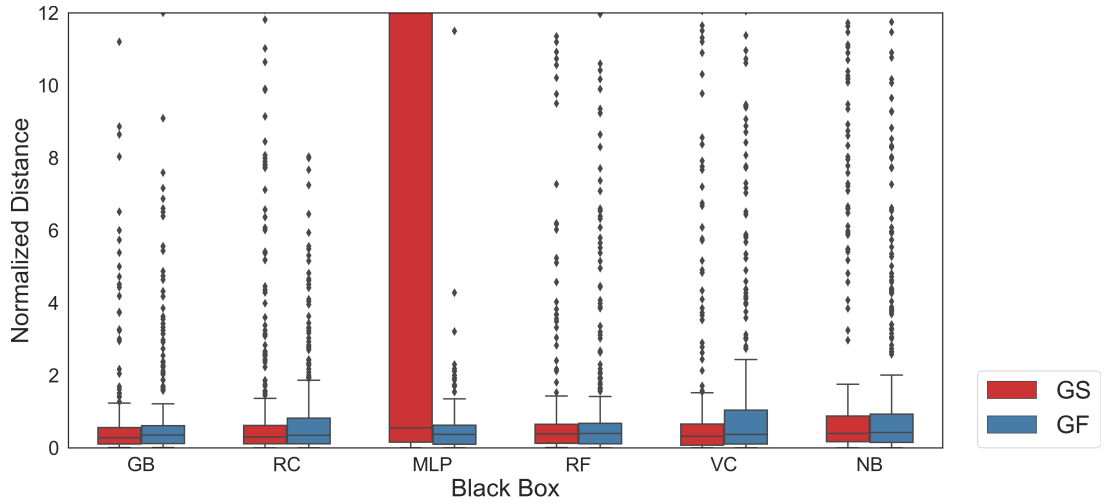


Figure 3.14 – Average Mahalanobis distance between the Growing Spheres (GS) and Growing Fields (GF) counterfactual instances and their closest real enemy.

artificial instances and generates the closest artificial counterfactual, outperforming Growing Spheres in terms of speed and realism. This significant improvement arises from Growing Fields' capacity to adjust the expansion speed of the field radius based on data distribution. Specifically, it ensures that features experience an expansion speed adapted to the magnitude of their values. We notice that on average over 100 instances, seven datasets and five black box models, Growing Fields achieves an average speed-up of two orders of magnitude compared to Growing Spheres to discover the closest counterfactual.

Furthermore, the runtime of generating a counterfactual with Growing Fields ranges from 0.08 to 3.03 seconds. In contrast, Growing Sphere exhibits considerable runtime varying between 1.53 and 725 seconds. Consequently, Growing Fields offers a superior balance of speed and reliability for generating counterfactual explanations compared to Growing Spheres.

3.6 Discussion and Conclusion

The fundamental question of what makes an explanation suitable for a particular use case lies at the junction of XAI and cognitive sciences. For this reason, this research question has not been addressed from a holistic perspective but rather from different, still complementary, angles.

On the one hand, the XAI community has put emphasis on the development of post-hoc explanation paradigms that identify the features that play a role in the predictions of

	GB		MLP		RF		RC		VC	
	GS	GF	GS	GF	GS	GF	GS	GF	GS	GF
Blood	31.8	**0.14**	212	**0.26**	96.4	**0.70**	371	**0.11**	725	**2.09**
Blob	6.71	**0.12**	15.6	**0.18**	12.7	**0.47**	14.6	**0.08**	38.2	**1.50**
Blobs	82.9	**0.42**	158	**0.53**	153	**0.86**	162	**0.35**	408	**2.47**
Circles	1.53	**0.19**	2.05	**0.30**	2.21	**0.65**	10.2	**0.12**	26.8	**3.03**
Diabetes	4.36	**0.21**	101	**0.28**	45.7	**0.58**	23.4	**0.14**	42.8	**2.14**
M Blobs	237	**1.00**	260	**0.84**	315	**1.69**	241	**0.62**	573	**2.82**
Moons	1.69	**0.15**	4.24	**0.20**	3.41	**0.47**	5.24	**0.09**	14.0	**1.56**

Table 3.6 – Average runtime (in seconds) of Growing Spheres (GS) and Growing Fields (GF) over 100 instances per black box and dataset.

an AI model. Among those, feature attribution rankings based on linear surrogates such as LIME [123] or LS [83], enjoy notable popularity, because they can provide accurate per-instance explanations [68]. Besides, practitioners from most disciplines are familiar with linear models. These surrogate models are learned so that they optimize for user-agnostic criteria such as the *fidelity*, i.e., the degree to which the surrogate mimics the black-box model it aims to uncover. Most of the literature in classical XAI has pushed the state of the art towards novel approaches – or improvements of existing ones. As a result, none of these works tackles the question of when a linear surrogate is objectively a reliable explanation.

On the other side of the spectrum, cognitive and social sciences study the subjective and human aspects of explaining AI models. In that spirit, the suitability of an explanation is characterized by its comprehensibility and plausibility [47]. Comprehensibility captures the extent to which a user grasps an explanation and can use it to accomplish well-defined tasks [11], e.g., determine the features used by the black-box system, predict the black box's answer, etc. On the other hand, the plausibility dimension models the cognitive preferences and background of the users. As pointed out by several studies [47, 79], users can reject an explanation if it contradicts common sense, for instance, if the explanation is too simplistic given that the underlying problem is deemed complex. The consensus seems to indicate that showing plausible and sound explanations increases trust in AI systems [79, 150], whereas the effects on comprehensibility and task efficiency are mixed.

While the XAI and cognitive science communities may appear somehow unreconciled, the relevance of the quality dimensions targeted by classical XAI methods has been justified by user studies. It has been suggested [79] that in the context of recommender systems, low fidelity harms trust in explanations. In this line of thought, we introduced in this chapter, a novel approach to determine *a priori* the pertinence of local linear explanations for a given use case.

Our decision is driven by standard user-agnostic desideratum, namely the fidelity of explanation. The results of our experiments provide evidence that characterizing the decision boundary of a black-box classifier around a target instance and making informed choices between linear and rule-based explanations are indeed feasible. In that spirit, the answers generated by APE, are promising for users of AI systems and linear surrogate explanations. When APE favors a rule-based explanation, it effectively communicates that the classification boundary is likely complex, and relying on a unique linear attribution explanation would result in incompleteness or inaccuracy.

Furthermore, evidence also suggests that multi-paradigm explanations can have a positive impact on comprehensibility [102, 151]. In particular, counterfactual explanations can be a complement to attribution-based or rule-based explanations and enrich the user's experience. Indeed, when the classifier's decision boundary has a multimodal distribution, APE presents a diverse and representative set of scenarios. These scenarios effectively illustrate how changes in the input features impact the classifier's output. This perspective extends to the combination of linear attribution and rule-based explanations as well. Thus, our work does not discourage the exploration of alternative combinations of explanation paradigms; rather, it provides valuable insights into the nature of the classifier's decision boundary. Such insights can be particularly useful in scenarios where the objective is to replace the black-box model, such as reverse engineering, or when a single, comprehensive, and unambiguous explanation is required.

Looking ahead, we envision to explore several avenues for future works. We have shown that APE is adapted to diverse data types such as textual and tabular data, thus, one possible avenue may be to extend APE to accommodate various distance metrics, and a wider spectrum of machine learning tasks, such as regression. Furthermore, we envisage enhancing our framework to support other explanation paradigms. For instance, future versions of APE may determine when a rule-based explanation is suitable for a specific instance and which rule-based model should be employed, such as Anchors or a decision tree. Additionally, we see an opportunity for research that integrates concepts of coverage, complexity, and plausibility when deciding on the most suitable explanations for a given use case. Indeed, methods that produce global explanations by combining local explanations may prefer simple explanations of high coverage over faithful and very specific explanations.

In the present chapter, we explored the importance of selecting an appropriate explanation paradigm to enhance the quality of generated explanations. However, as we adapted the unimodality test to textual data, a new challenge emerged: select the correct conversion space used by the surrogate. When working with textual data, there are two primary approaches to

convert text into numerical representations suitable for ML models. The first approach involves using an interpretable space that converts words into vectorized representations, such as a bag of words. The second approach employs more complex methods to embed text into a latent space. In the next chapter, our objective is to assess how the choice between these two conversion methods impacts the quality of the generated explanations. Specifically, we will focus on counterfactual explanation techniques and compare a handful of state-of-the-art counterfactual techniques, including Growing Language and Growing Net. The goal is to provide a detailed analysis of the spectrum of existing counterfactual methods. This analysis focuses on the techniques employed to perturb input text and offers valuable insights into the question of employing a black box (in the form of a latent space) to explain another black box.

CHAPTER 4

EXPLAINING A BLACK BOX WITHOUT ANOTHER BLACK BOX

Contents

4.1 Context **93**

4.2 Counterfactual Techniques for Textual Data **96**

4.3 Comparative Study **98**

4.3.1 Complexity Spectrum 98

4.3.2 Experimental Information 100

4.4 Results **102**

4.4.1 Counterfactual Quality 103

4.4.2 Method Quality 105

4.5 Conclusion **108**

4.1 Context

The latest advances in Artificial Intelligence, and more particularly in machine learning, have significantly transformed various natural language processing (NLP) tasks [34, 92, 132], including text generation, fake news detection, sentiment analysis, and spam detection. These

notable improvements can be attributed, in part, to the adoption of methods that encode and manipulate text data within latent representations. In this context, latent representations refer to abstract, high-dimensional vector spaces where text is transformed into numerical form, thus capturing the underlying semantics, structure, and patterns of language. These latent representations offer a bridge between raw text data and machine learning models, allowing algorithms to operate on a more structured and semantically rich representation of language.

However, the impressive gains in accuracy achieved by modern algorithms, such as Transformers models [34], can be diminished by the lack of interpretability of those algorithms [137]. Indeed, a model could make correct predictions for the wrong reasons [60, 99]. Unless the machine learning model is a white box, explaining the results of such an agent requires an explanation layer that interprets the internal workings of the black box in a post-hoc manner.

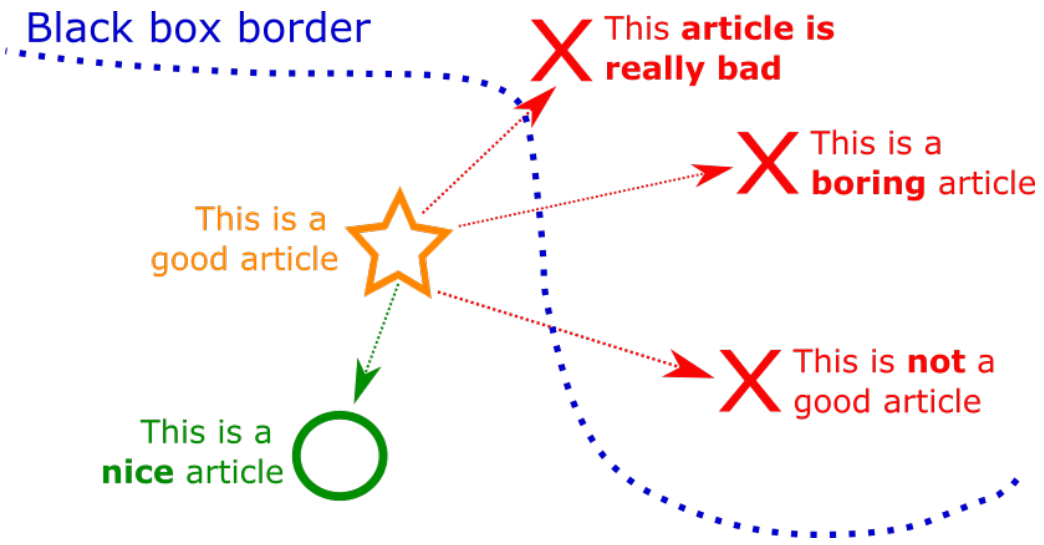
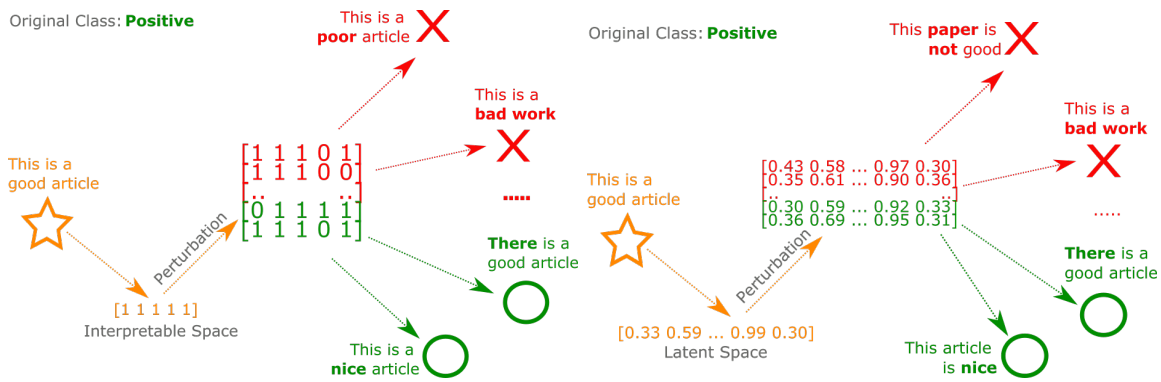


Figure 4.1 – Target instance is represented by the sentence "This is a good article" while other texts are artificial textual documents. The blue dashed line represents the classifier's decision boundary for predicting review polarity. Texts shown in red are classified as negative, and those in green are classified as positive.

While there are several ways to explain the outcomes of an ML model a posteriori, our previous chapter concentrated on feature attribution techniques. We now shift our attention to counterfactual explanations, a domain that experienced notable popularity over the last five years [55, 102]. A counterfactual explanation is a counter-example that is similar to the original text, but that elicits a different outcome in the black box [151]. Consider the classifier depicted in Figure 4.1, for sentiment analysis applied to the review "This is a good article" – classified as positive. In this toy example, a counterfactual could be the phrase "This is a boring article".

Through this explanation, the counterfactual technique conveys that the adjective "good" was a possible reason for this sentence to be classified as positive, and changing the polarity of that adjective may change the classifier's response.

In the literature, counterfactual explanation methods operate by increasingly perturbing the target text until the classifier's answer changes. These methods lie in a spectrum spanning from fully transparent to fully opaque techniques. On one side of the spectrum, *transparent* methods perturb the target text by adding, removing, or changing words and syntactic groups [98, 127, 159] in the original target text. To illustrate this concept, refer to Figure 4.2a, which provides a visual representation. Here, the yellow star represents the original target text, which is transformed into a binary vector. In this vector, a '1' indicates the preservation of the original word, while a '0' means the word will be replaced by another word. On the opposite side, more recent methods [64, 111, 126] embed the target text in a latent space on which perturbations are carried out subsequently. The embeddings are a compressed representation of the classifier's training data, which filters noise and retains the essential information for classification. An example of a latent perturbation is presented in Figure4.2b, where the input text is encoded into a high-dimensional numerical vector. Small perturbations are introduced to these vectors within the latent space before bringing them back to the original data space. These methods are classified as *opaque* due to the inherent lack of interpretability of the latent space [90].



(a) Transparent perturbation (b) Opaque perturbation

Figure 4.2 – The mechanism employed to perturb the target documents by the transparent and opaque methods. The transparent techniques convert the input text to a vector representation, where 1 indicates the presence of the input word and 0 denotes a replacement. The opaque methods embed words from the target text into a latent space and perturb the text in this high-dimensional space.

In this chapter, our primary objective is to determine whether the use of complex latent spaces is an interesting approach for generating counterfactual explanations. To this end,

we conduct a comparative examination of these two families of counterfactual approaches, in order to clarify the advantages of one over the other. Based on our empirical analysis, we discovered that for some downstream NLP tasks such as spam detection, fake news or sentiment analysis, learning a compressed representation can be overkill. To illustrate this, and as a proof of concept, we employed two transparent counterfactual explanation techniques, namely, Growing Language and Growing Net, introduced in Section 3.3.2. Intriguingly, these transparent techniques outperform opaque methods, mainly because, opaque approaches often produce non-intuitive counterfactual explanations, i.e., counter-example texts that bear no resemblance to the target text. This not only contradicts the essence of counterfactual explanations but also raises questions about the actual level of transparency achieved when explaining a black box with another black box.

In the previous chapter, we introduced a framework designed to enhance explanation fidelity. Our focus was primarily on enhancing fidelity from a data-oriented perspective, as prior research has established the profound impact of fidelity on user interaction with explanations [47, 79]. On this basis, our current research aims to determine whether the complexity of opaque methods translates into significant performance improvements over transparent methods. Consequently, we evaluate the performance of transparent counterfactual explanation methods compared to opaque methods on three classic NLP tasks: fake news detection, sentiment analysis, and spam detection. Before delving into our experimental setup and results in Section 4.3, we will begin by reviewing the various counterfactual explanation methods found in the literature in Section 4.2.

The work covered in this chapter served as the foundation for the paper titled: Explaining a Black Bow without a Black Box, which is planned for submission to the NAACL 2024 conference.

4.2 Counterfactual Techniques for Textual Data

Counterfactual explanation methods compute contrastive explanations for ML black-box algorithms by providing examples that resemble a target instance but that lead to a different answer in the black box. These counterfactual explanations convey the minimum changes in the input that would change a classifier's outcome. Social sciences [102] have shown that human explanations are contrastive and Wachter et al. [151] have illustrated the utility of counterfactual instances in computational law. When it comes to NLP tasks, a good counterfactual explanation

should be sparse [115], i.e., look like the target instance, and be fluent [105], i.e., read like something someone would say.

Counterfactual approaches have gained popularity in the last few years. As illustrated by the surveys, first by Bodria et al. [15] and later by Guidotti [55], around 50 additional counterfactual methods appeared in a one-year time span. Despite this surge of interest in counterfactual explanations, their study for NLP applications remains underdeveloped [127]. In the following, we elaborate on the existing counterfactual explanation methods for textual data along a spectrum that spans from transparent to opaque approaches.

Transparent Approaches. Given an ML classifier and a target text (also called a document), transparent techniques compute counterfactual explanations in a binary space. Each dimension represents the presence (1) or absence (0) of a word from a given vocabulary. Hence, to perturb a text, these methods toggle on and off 0s and 1s, where 0s are tantamount to adding, removing, or replacing words until the classifier yields a different answer. This was first proposed by Martens and Provost [98] who introduced Search for Explanations for Document Classification (SEDC), a method that removes the words for which the classifier exhibits the highest *sensitivity*. These are words that impact the classifier's prediction the most. Similarly, feature-attribution explanation methods such as LIME [123] and SHAP [94], mask words randomly from the target text. More recently, Ross et al. [127] developed Minimal Contrastive Editing (MICE), a method that employs a Text-To-Text Transfer Transformer to fill masked sentences. Yang et al. [159] presented Plausible Counterfactual Instances Generation (PCIG), which generates grammatically plausible counterfactuals through edits of single words with lexicons manually selected from the economic domain. Since these methods are tailored for specific tasks or require manual selection, we removed these methods from our experiments.

Opaque Methods. We define opaque approaches as those perturbing the input text in a latent space in $\mathbb{R}^n$. Methods such as Decision Boundary [64], xSPELLS [111] or counterfactualGAN [126] operate in three phases. First, they embed the target instance onto a latent space. This is accomplished by employing specific techniques such as Variational AutoEncoder (VAE) in the case of xSPELLS, or a pre-trained language model (LM) for counterfactualGAN. Second, while the classifier's decision boundary is not traversed, these methods perturb the latent representation of the target phrase. This is done by adding Gaussian noise in the case of xSPELLS, whereas counterfactualGAN resorts to a Conditional Generative Adversarial Network. Finally, a decoding stage generates sentences from the latent representation of the perturbed documents. There also exist methods such as Polyjuice [158], Generate Your Counterfactuals (GYC) [96] and Tailor [128] that perturb text documents in a latent space, such as LM and Transformers,

but can be instructed to change particular linguistic aspects of the target text, such as locality or grammar tense. Such methods are not particularly designed to compute counterfactual explanations but are rather conceived for other applications such as data augmentation.

Unlike pure word-based perturbation methods, latent representations are good at preserving *semantic closeness* for small perturbations. That said, these methods are not free of pitfalls. First, methods such as xSPELLS and CounterfactualGAN are deemed opaque since a latent space is not human-understandable [137]. Therefore, an intriguing paradox emerges between developing explanation techniques through latent space and methods explaining latent spaces [90]. It invites us to consider a fundamental question: Is it prudent to employ non-interpretable mechanisms to gain insights into complex classifiers? Moreover, existing latent-based approaches do not seem optimized for sparse counterfactual explanations. We prove this through experimental results that reveal a minor alteration in a latent space can cause a significant alteration in the original space.

4.3 Comparative Study

Until now, we presented counterfactual explanation techniques as either opaque or transparent. However, the landscape is more nuanced. In this section, we provide an exploration of the complexity spectrum, which is depicted in Figure 4.3. Subsequently, we delve into the details of the experiment setup in Section 4.3.2, which is essential for replicating our experiments and understanding the methodology.

4.3.1 Complexity Spectrum

We introduce a complexity spectrum that serves as a visual guide to discern the transparency of various counterfactual explanation methods. The spectrum spans from the most transparent methods on the left to the most opaque ones on the right.

Fully Transparent. At the leftmost end of the spectrum, we find the fully transparent approaches. This starts with the entirely transparent method SEDC [98], which perturbs text instances by hiding only highly sensitive words within the text.

Transparent Counterfactual Methods. We positioned Growing Net as the second most transparent method as it goes beyond simple word masking by replacing words. Growing Net leverages the knowledge and tree structure of WordNet to select word substitutions more judiciously. This enables Growing Net to provide the user with information such as synonyms,

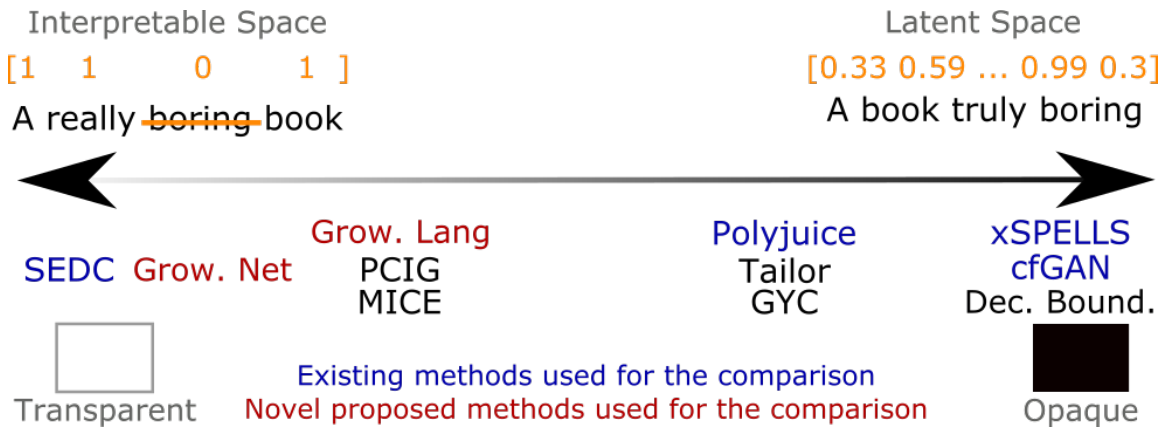


Figure 4.3 – Spectrum for counterfactual explanation techniques that goes from the most transparent methods on the left (e.g., SEDC) to the most opaque methods such as CounterfactualGAN, and xSPELLS, passing by our methods in red Growing Net and Growing Language. Transparent methods perturb documents in a binary space; opaque methods do it in a latent space.

antonyms, and hypernyms. Methods like PCIG, MICE, and Growing Language are classified as relatively more opaque among the transparent counterfactual techniques. They employ the latent space to identify semantically close word substitutions. However, despite their reliance on black-box techniques, these methods are considered transparent because their generated explanations preserve the document's structure while revealing which words should be replaced and by which other words.

Partially Opaque. Polyjuice, Tailor, and GYC fall in the category of partially opaque methods, as they leverage control codes to perturb the target document. Control codes are specific instructions that guide the model to perform a certain task, such as translating, summarizing, or changing the tense of a text. While these modifications occur in a latent space, the inclusion of control codes enhances the clarity regarding why a modification influences the model's prediction.

Fully Opaque. On the far right of the complexity spectrum, we encounter fully opaque approaches such as Decision Boundary, xSPELLS and counterfactualGAN. These methods operate in a completely opaque manner, making it challenging for users to discern the underlying process of counterfactual generation.

This complexity spectrum provides valuable insights into the transparency and opacity of counterfactual explanation methods, enabling a more nuanced understanding of their capabilities.

4.3.2 Experimental Information

In this section, we provide a detailed account of the counterfactual generation process, the datasets employed, the classifiers to be explained, and the metrics applied in our experiments.

4.3.2.1 Counterfactual Generation

We start by outlining the six distinct counterfactual methods taken for generating counterfactuals. We excluded PCIG and MICE from our subsequent analysis for various reasons. Regarding PCIG, we highlight that this method relies on domain-specific rules from the field of economics, which limits its applicability to our diverse datasets. In the case of MICE, we observe this approach relies on transformer models to identify semantically relevant word replacements, which deviates from our focus on less complex methods. Furthermore, we have excluded methods such as Text Attack, introduced by Morris et al. [105], from our analysis. These adversarial methods are not designed for explanatory purposes but rather to fool the model. Similarly, we have removed Linguistically-Informed Transformation (LIT), introduced by Li et al. [88], which is a method aimed at automatically generating sets of contrasts. These methods aim to generate documents that are outside the data distribution and are therefore unrealistic. This goes against the desired attributes for effective counterfactual explanations

We fill the middle ground with two methods, Growing Net and Growing Language, which implement a similar strategy to existing transparent methods. However, they do so with fewer methodological complexities. We adapted the code used to generate counterfactuals for the three transparent methods (SEDC, Growing Net, and Growing Language) and made it available on GitHub [1]. We employed the original code for the opaque methods, as described below.

SEDC: We modified the code used for word masking to ensure its compatibility with classification models that do not output class probabilities. This modified code version is accessible in Python on our GitHub as a variant of the counterfactual method class. This class proposes to choose from SEDC, Growing Net, or Growing Language, all specialized in generating transparent explanations.

Polyjuice: To generate counterfactuals, we utilized the code available in the repository `https://github.com/tongshuangwu/polyjuice`. Default hyperparameters were used to perturb texts from each test set until we found instances classified differently by the model.

xSPELLS: We employed the V2 version of xSPELLS, found in the repository `https://github.com/lstate/X-SPELLS-V2`, with default hyperparameters.

1. `https://github.com/j2launay/ebbwbb`

counterfactualGAN: We used the code provided in the official release page of the paper, which can be accessed at `https://aclanthology.org/2021.findings-emnlp.306/`. We executed counterfactualGAN (cfGAN) with the default hyperparameters.

This comprehensive approach to counterfactual generation ensures a diverse set of methods to evaluate and compare in our experiments.

4.3.2.2 Datasets

For our experiments, we used three datasets designed for three different applications: (a) spam detection in messages, (b) sentiment analysis, and (c) detection of fake news from newspaper headlines. Each of these datasets comprises two target classes and contains between 4000 and 10660 textual documents. The average number of words in each document is between 11.8 and 20.8 as reported in Table 4.1.

Concerning the fake news detection dataset, we constructed it by taking real newspaper titles from a dataset [2] and fabricated titles from a fake news dataset [3]. This combined dataset is publicly available on our GitHub [4]. As for the polarity [5] and spam [6] detection datasets, we took them from Kaggle. We divided each dataset into training and testing sets using the scikit-learn library's function: [7]: $train\_test\_split$ with a test size of 30% and a random seed of 1.

Name	Nb Words			Instances	Models' Accuracy	
	Total	Average	STD		Neural Network	Random Forest
Fake	19419	11.8	3.2	4025	84%	84%
Polarity	11646	20.8	9.3	10660	72%	67%
Spam	15587	18.5	10.6	8559	100%	100%

Table 4.1 – Information about the experimental datasets. The three columns under "Nb Words" represent respectively (a) the total number of distinct words in the whole dataset, (b) the average number of words per sentence, and (c) the standard deviation. The "Instances" column indicates the number of text documents per dataset. The last columns show the average accuracy of the two classifiers for each dataset.

2. `https://www.kaggle.com/datasets/rmisra/news-category-dataset`
3. `https://www.kaggle.com/competitions/fake-news/overview`
4. `https://github.com/j2launay/ebbwbb`
5. `https://www.kaggle.com/datasets/nltkdata/sentence-polarity`
6. `https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset`
7. `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

4.3.2.3 Black-box Classifiers

Our evaluation uses two distinct black-box classifiers implemented using the scikit-learn library and already employed in [111]. These black boxes are (i) a Random Forest (RF) consisting of 500 tree estimators, and (ii) a straightforward neural network (DNN) with the same amount of neurons as there are words in the dataset.

In addition to predicting the class associated with a textual document, these classifiers provide class probabilities. We trained both classifiers on 70% of the dataset and their accuracy was tested on the remaining 30%. We also selected the target instance to explain within this test set. Across all datasets, the average accuracy of these two classifiers ranges from 67% to 100%. Detailed results are presented in Table 4.1.

We employed a count vectorizer [8] to convert the texts from the dataset into a matrix of tokens as input for the model.

4.4 Results

In this section, we present the outcomes of our comprehensive evaluation of six counterfactual explanation techniques, each positioned along the transparency spectrum illustrated in Figure 4.3. We conducted six rounds of experiments categorized into two main aspects. First, we assess the quality of the generated counterfactual explanations based on three essential criteria: (i) minimality, (ii) outlierness, and (iii) plausibility. Second, we evaluate the methods themselves in terms of (iv) flip change, (v) stability, and (vi) runtime. Our evaluation is based on three well-known NLP classification tasks: spam detection, polarity review, and fake news detection. For each task, we trained both a neural network, specifically a multi-layered perceptron, and a random forest classifier as described in Section 4.3.2.3. In total, we generated counterfactual explanations for 100 target texts per dataset, black-box classifier, and counterfactual methods, serving as the inputs for our comprehensive evaluation. The explanation methods are ordered following their positions on the complexity spectrum from Figure 4.3, ranging from the most transparent to the most opaque methods. Additionally, we have made our code and datasets available on GitHub [9].

8. `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`

9. `https://anonymous.4open.science/r/ebbwbb-4B55/README.md`

4.4.1 Counterfactual Quality

A high-quality textual counterfactual explanation adheres to several essential criteria [55], including (i) minimal changes that make the counterfactual look closely to the target text, (ii) non-outlierness, meaning the counterfactual instance should resemble other phrases in the classifier's training/testing set, and (iii) linguistic plausibility, which entails that the counterfactual should sound like something a person would naturally write or say. The minimality criterion is quantified by measuring the distance between the counterfactual and the target sentence. Outlierness is operationalized as the distance of the counterfactual from the "data manifold". We thus measure outlierness through the distance between the counterfactual and the closest real instance in the dataset [32]. Linguistic plausibility, though typically evaluated through user studies [96, 158], is approximated here following Ross et al. [127, 128] and using perplexity scores based on a GPT language model [19], where lower scores indicate higher plausibility.

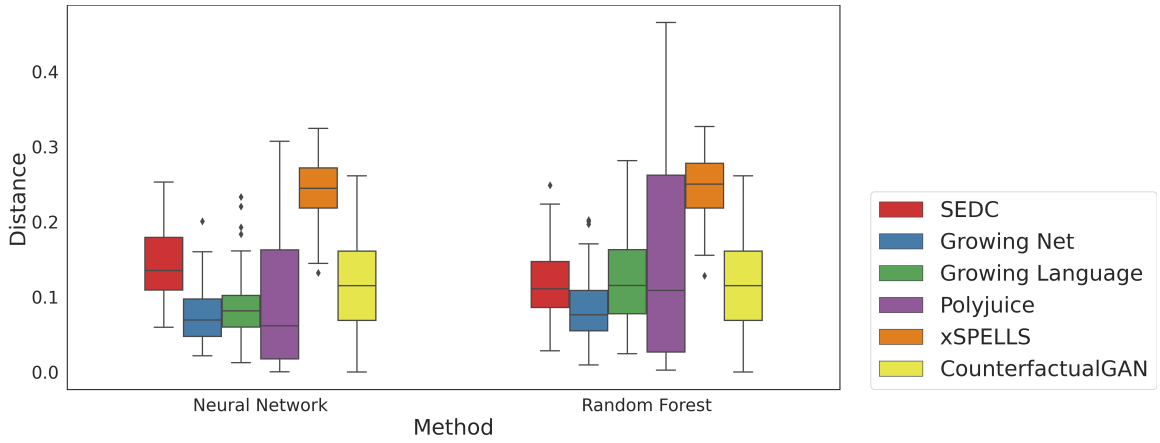


Figure 4.4 – Minimality as the distance between the closest counterfactual and the target document (the lower the better).

Figure 4.4 displays the results for minimality, which is the distance between the generated counterfactuals and their corresponding target texts. To assess minimality, we embedded both the counterfactual and the original text using a sentence model derived from GPT [108]. Subsequently, we computed the cosine distance between these embeddings. This ensures that our measurement considers a balance between lexical similarity and latent features, such as "style". Notably, our findings reveal that methods positioned in the middle-ground, particularly Growing Net, performed favorably compared to opaque approaches. It is worth noting that xSPELLS introduced the most significant changes to the original text. Similarly, we observe a high variance in the minimality of the counterfactuals generated by Polyjuice, indicating that

some counterfactuals were notably distant from their corresponding target instances. While these methods introduced minor perturbations to the original text, these modifications occurred within a latent space. Nothing guarantees, however, that these minor adjustments translate into visually subtle modifications of the target phrase when the resulting phrase is brought back to the original space.

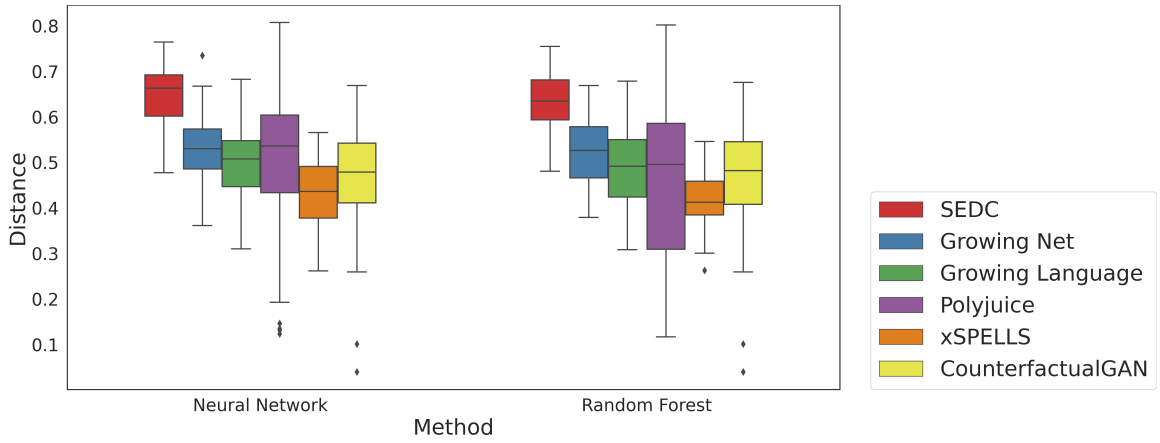


Figure 4.5 – Outlierness, quantified as the measure of distance between the counterfactuals and the nearest instance within the test set.

Figure 4.5 illustrates the results for outlierness, which are computed as the distance between the generated counterfactuals and the closest test instances within our experimental datasets – hence the lower the better. We observe that xSPELLS excelled in this criterion. This success can be attributed to its reliance on VAEs that are fine-tuned on the dataset and designed to create a compressed representation of it. At the other extreme, SEDC performed poorly, because removing words from the target text may lead to incomplete sentences that are identified as outliers. Notably, both Growing Net and Growing Language outperformed counterfactualGAN and Polyjuice in terms of outlierness, despite not relying extensively on heavy neural-network-based machinery.

Figure 4.6 presents the plausibility of the counterfactuals, measured through the perplexity scores obtained from a language model. This score is computed by calculating the mean squared error (MSE) loss of a GPT model when predicting the next token in the counterfactual. We normalized the perplexity scores where lower scores indicate higher plausibility. SEDC and Polyjuice generated texts with the lowest plausibility, which is expected since SEDC masks words, leading to nonsensical sentences. In contrast, both Growing Net and Language achieved perplexity loss similar to those of xSPELLS and counterfactualGAN.

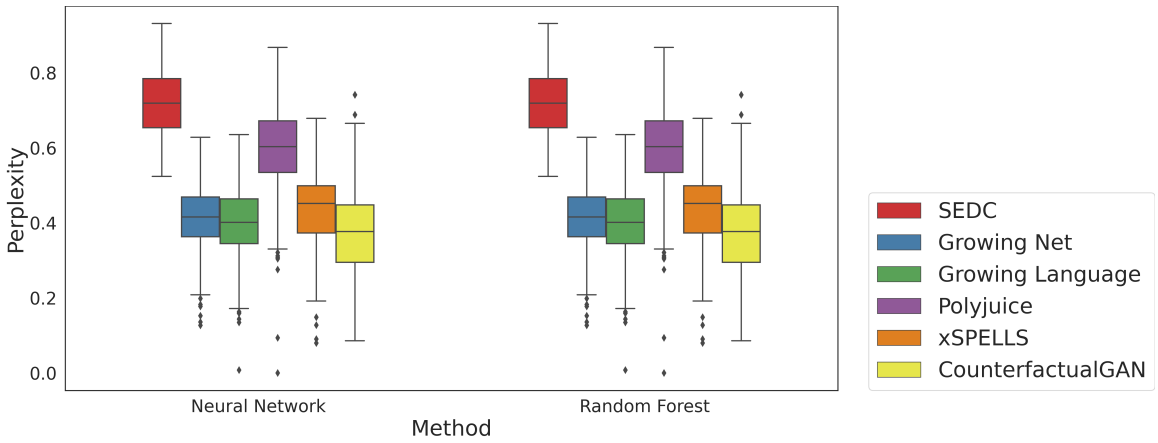


Figure 4.6 – Perplexity as the MSE loss of a GPT model on the generated counterfactuals.

Method	SEDC		Grow. Net		Grow. Lang.		cfGAN		xSPELLS		Polyjuice	
Black box	DNN	RF	DNN	RF	DNN	RF	DNN	RF	DNN	RF	DNN	RF
spam	0.57	0.55	0.36	0.35	**0.65**	0.49	0.57	0.57	0.34	**0.61**	0.14	0.21
fake	0.90	0.98	0.79	0.59	0.86	0.86	**1**	**1**	**1**	**1**	0.38	0.28
polarity	0.95	0.92	**1**	0.85	0.93	0.93	0.65	0.65	**1**	**1**	0.46	0.46

Table 4.2 – Average label flip per dataset and black box of the six counterfactual methods.

4.4.2 Method Quality

In this section, we compare the quality of the counterfactual explanation methods based on three key criteria: (iv) label flip rate, which measures how frequently these methods successfully produce a counterfactual, that is, an instance classified differently by the model, (v) stability, the average similarity between five generated counterfactuals for the same document, and (vi) runtime, the time it takes for each method to generate a counterfactual explanation.

Table 4.2 provides an overview of the label flip results, which indicates the methods' ability to find a counterfactual for a given text document. It is noteworthy that xSPELLS achieves the highest label flip rate, except for the spam detection dataset using neural networks. Additionally, our transparent methods consistently outperformed Polyjuice on every task and counterfactualGAN for polarity review. Notably, Growing Net exhibits strong performance for the polarity dataset. This highlights the effectiveness of replacing words with antonyms as a means to discover counterfactuals. We also emphasize that both Growing Net and Growing Language can be fine-tuned for a more exhaustive search by adjusting its parameters, for

example by lowering the similarity threshold or incorporating additional terms from WordNet's tree structure. While this can enhance the label flip rate, it may result in longer runtimes.

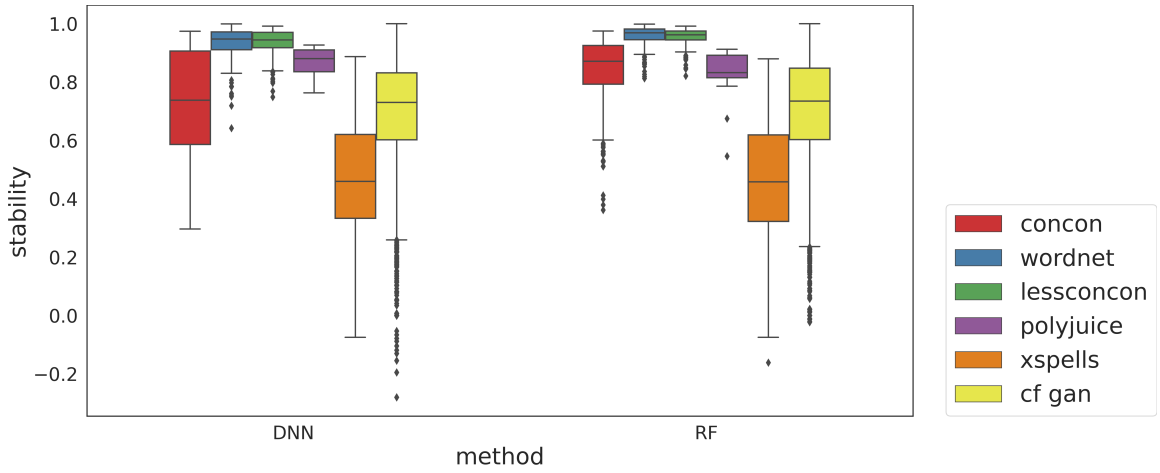


Figure 4.7 – Stability as the average similarity between the generated counterfactuals for five successive runs on a single text document. A similarity close to 1 indicates that the five counterfactuals generated are similar.

We present in Figure 4.7, the stability results that measure the average similarity between five counterfactuals generated for the same target document. For each of the six approaches, we iterated five times over the same target document and measured the average similarity between each of the five generated counterfactuals. A method exhibits good stability if it systematically generates similar counterfactuals for the same document.

The results highlight that both Growing Language and Growing Net are the most stable. These methods obtain almost perfect stability (stability index close to one) while opaque methods and especially xSPELLS have a wide variability across different runs. We believe that these observations are due to variations in the set of similar words. Opaque methods, in particular, require re-training of the latent module, such as the VAE or the GAN, for each run. In contrast, the procedures employed by Growing Language and Growing Net are deterministic, and the sets of similar words remain consistent across successive generations. Our results also show similar stability for SEDC and Polyjuice, however, SEDC exhibits higher variance on the Neural Network classifier. We note that the high stability of Growing Language, Growing Net, and SEDC is because they have fewer steps that rely on randomness. Our results on stability suggest that transparent methods are less sensitive to random seeding than methods based on NN learning.

dataset	method	DNN	RF
spam	SEDC	21.45 (12.57)	15.8 (9.36)
	Growing Net	0.97 (1.0)	0.74 (0.8)
	Growing Language	60.21 (15.55)	56.9 (13.84)
	Polyjuice	31.53 (7.07)	62.23 (183.56)
	counterfactualGAN	1.4 (0.02)	1.43 (0.03)
	xSPELLS	219.45 (16.78)	197.63 (16.26)
fake	SEDC	30.7 (14.34)	13.02 (6.15)
	Growing Net	1.54 (1.4)	1.03 (0.76)
	Growing Language	54.81 (28.11)	54,66 (12,16)
	Polyjuice	38.23 (7.75)	70.04 (185.15)
	counterfactualGAN	1.03 (0.18)	1.0 (0.13)
	xSPELLS	84.11 (6.47)	85.63 (7.0)
polarity	SEDC	12.91 (9.77)	12.39 (9.41)
	Growing Net	0.69 (0.91)	0.63 (0.83)
	Growing Language	74.7 (33.3)	73.56 (32.44)
	Polyjuice	81.07 (30.34)	82.49 (47.66)
	counterfactualGAN	1.27 (0.23)	1.29 (0.25)
	xSPELLS	135.69 (19.32)	115.94 (10.9)

Table 4.3 – Average runtime in seconds for an instance (and standard deviation) per dataset and black box of the six counterfactual methods. Note that the time for CounterfactualGAN (cfGAN) does not take into account the time to fine-tune the GAN on the target dataset. The training takes 6755, 4300, and 5770 seconds for the spam, fake, and polarity datasets, respectively.

Finally, our runtime results are in Table 4.3. The table details the average and standard deviation of the runtime for each counterfactual explanation method across datasets and classifiers. Notably, CounterfactualGAN and Growing Net emerged as the fastest methods for generating counterfactuals. However, it is important to note that CounterfactualGAN requires the training of the Variational AutoEncoder (VAE) on each specific dataset, a process that incurs long training times. The time needed for fine-tuning varies, ranging from 4300 seconds for fake news title detection to 6755 seconds for spam detection.

Furthermore, we observe that xSPELLS and Growing Language exhibit the slowest runtime performance. Growing Language, for instance, requires approximately 60 seconds to generate a single counterfactual, while xSPELLS exhibits varying runtimes, ranging from 85 seconds for fake news detection to 219 seconds for spam detection. In summary, the two opaque methods necessitate a training step to adapt to specific datasets, whereas our two novel methods, along with SEDC, offer faster and less effortful development.

To conclude, our runtime experiments reveal that methods like Growing Net are fast enough to be employed for the real-time generation of counterfactual explanations. This stands in sharp contrast to xSPELLS, which is two orders of magnitude slower due to its decoding phase from the VAE latent space to the original space.

4.5 Conclusion

Our evaluation provides valuable insights into the landscape of generating counterfactual explanations for downstream NLP tasks. One of the most striking findings is that complexity, often associated with the use of neural networks and latent spaces, does not necessarily equate to superior performance in this context. Surprisingly, our results demonstrate that simpler approaches, characterized by a systematic and judicious strategy for word replacement within the target sentence, consistently yield satisfactory outcomes across a range of quality dimensions.

The results of our study prompt a deeper reflection on the optimal strategies for generating counterfactual explanations in the field of NLP. It invites readers to contemplate the broader implications of our findings and their implications for the development of transparent approaches versus improving opaque methods. The choice between these approaches should be made judiciously, considering the specific requirements and constraints of the application at hand.

Furthermore, our findings underscore the critical importance of transparency and interpretability in AI and machine learning. As we navigate in the complex landscape of increasingly sophisticated AI models, the need for transparency, accountability, and trust becomes paramount, especially in applications where human decisions are influenced by AI recommendations. The paradox of explaining one black box with another raises pertinent questions about the balance between model complexity, interpretability, and performance. It calls into question the unnecessary development of opaque approaches when transparent methods suffice, reinforcing the need for clarity and user-friendliness in AI systems.

In conclusion, our study contributes to the ongoing discourse on explainable AI by highlighting that effective counterfactual explanations can be achieved through simpler methods, challenging the prevailing assumption that complexity is always synonymous of better performance. We expect these findings will encourage the development of more transparent and interpretable AI systems, fostering trust and accountability in AI-driven decision-making processes.

In the second part of this thesis, our focus shifts to the user perspective. In this new part, we will delve into the impact of the three explanation methods used in each of the previous

chapters. Before doing that, we reflect on the insights we collected throughout these chapters, primarily focusing on the data perspective.

First, our research journey through these chapters has revealed that there is ample room for improving existing explanation techniques. A noteworthy illustration of this is found in Chapter 2, where we demonstrate how pertinent and meticulous adjustments to the essential components of an explanation method can significantly improve the effectiveness of the resulting explanations. This suggests that further refinements and innovations in these methods hold the potential to enhance the quality and utility of explanations.

Secondly, it is imperative to assess judiciously when a particular explanation technique should be applied. Chapter 3 provides an example in this regard by demonstrating that linear explanations are not universally applicable and that there is a need for context-specific considerations. This insight is expected to extend to other explanation paradigms, highlighting the importance of choosing the right method for the dedicated task. This theme will be explored in the following part of this thesis.

Last but not least, the exploration of counterfactual explanation techniques conducted in this chapter offers an important lesson: the introduction of complexity does not inherently translate into improved performance. This lesson has brought implications not only for counterfactual explanations but extends to other explanation paradigms.

As we look ahead to the next part of this thesis, these insights serve as valuable guidance for our exploration of how to provide the best explanations from a user's perspective. We anticipate that these findings will continue to inform our research and contribute to the ongoing discourse on explainable AI and machine learning.

PART II

WHAT MAKES A GOOD EXPLANATION FROM THE PERSPECTIVE OF THE USER?

CHAPTER 5

USER-CENTERED EVALUATION OF EXPLAINABILITY METHODS

Contents

5.1 Evaluating Explanations with Users 113
5.1.1 Evaluating Explainable AI Systems with Users 113
5.1.2 Guidelines and Metrics to Conduct User Studies 115
5.2 Method 116
5.2.1 Methodological Framework 116
5.2.2 Scales & Metrics 118
5.3 Conclusion 120

In the first part of this thesis, we have explored post-hoc explanation techniques, a widely popular family of methods that have gained significant attention within the XAI community. Despite the proliferation of post-hoc explainability methods, researchers have pertinently highlighted that XAI methods are not focusing sufficiently on the target users [38, 125]. In response to this observation, the second part of this thesis is dedicated to the evaluation of explanation techniques from a user-centric perspective. As highlighted by Doshi-Velez and Kim along with two other XAI surveys [1, 4], a low number of XAI papers justify their novel methods through application-grounded evaluations or human-grounded metrics. Adadi et al. [1] found that in a sample of 381 XAI papers, only 5% had an explicit focus on the evaluation of the proposed

methods through human subjects. In other words, research in XAI is busy producing tons of local feature-attribution, rule-based, and example-based explanation methods for users without evaluating the resulting explanations with real users. This gap in research implies that we know little about the extent to which users understand explanations for AI systems. Likewise, it remains unclear whether the presence of explanations when using AI systems increases or not users' trust. In fact, some research findings have shown that explainable AI may offer limited benefits to users of AI recommendation systems [20, 48, 53, 54, 110]. Similar results have been found for cognitive engagement [29, 40].

To address these pressing issues, this chapter introduces a methodological framework for conducting user studies. The primary objective of this framework is to establish a robust methodology to investigate the impact of explanation techniques on users. Furthermore, we propose a comprehensive set of scales and metrics to gauge users' perceived and behavioral trust, understanding, and satisfaction. This methodological framework, along with the scales and metrics, will be subsequently applied in Chapter 6 to conduct a user study. This study aims to measure the benefits of different explanation techniques and their representations (whether graphical or text-based) on users' comprehension and trust.

This chapter begins by providing in Section 5.1 an overview of existing user studies that have been conducted to evaluate explanation techniques. Then, we introduce the methodological framework, along with the proposed scale and metrics, in Section 5.2. In Section 5.3, we provide a discussion of how this framework may be applied.

5.1 Evaluating Explanations with Users

This contribution lies at the intersection of eXplainable AI and Human-Computer Interaction (HCI) research. Therefore, we discuss the evaluation of XAI systems from a user's perspective. Then, we present existing guidelines and metrics to conduct user studies on XAI tools.

5.1.1 Evaluating Explainable AI Systems with Users

Miller argues that the development of effective explanation modules requires the joint effort of the XAI and HCI research communities [102]. The HCI community has previously focused on the evaluation of XAI models with the end-users [10, 14, 36, 46, 150]. We can group these evaluations into two categories: (a) assessment of novel explanation modules and representations, and (b) impact of the explanation's type on users' perception (e.g., trust,

understanding). In contrast to these works, several surveys have highlighted the scarcity of XAI papers that evaluate novel explanation methods through user studies [1, 4, 38]. Consequently, most of the methods aimed at generating explanations for humans are evaluated without considering a human perspective. Among the studies that did consider the human perspective, most studies either assessed only the validity of their novel explanations method [80, 95, 123, 124, 126, 159] or the impact of the explanation's visual representation [28, 112, 119, 156]. A limitation of these works is that they are typically limited to the evaluation of one kind of explanation technique [80, 112, 126] and one application domain [119, 159]. For instance, Lakkaraju et al. proposed a rule-based explanation method and measured the users' understanding in the healthcare domain [80]. While their results proved that their method increases the users' understanding, it is only valid for the healthcare domain and should not be extended to other domains. Moreover, some prominent explanation methods such as LIME and Anchors, evaluate the quality of the explanations with a small number of computer science students, who are already familiar with machine learning concepts [123, 124]. In our work, we set out to compare three different explanation techniques on two distinct datasets.

To study the impact of explanations more broadly, prior works have evaluated users' trust and understanding in specific explanation conditions [5, 28, 79, 81, 150]. For instance, Arora et al. studied the impact of interactive explanations on users' understanding [5]. In their settings, an interactive explanation means that the user can edit the input text and directly visualize the effect on explanations. Their results confirmed that explanations might help users understand how the model works since they better identify key elements for the prediction. Cheng et al. compared the effect of interactive versus static explanations, as well as black box versus white box, on users' trust and understanding [28]. They observe that having access to the internal mechanism of the algorithm and interactive explanations both help to improve the users' comprehension. Other researchers have studied the influence of explanation's representation on users' perceptions [10, 28, 81]. As an example, Larasati et al. [81] presented four explanation styles: contrastive, general, truthful, and thorough, and measured the impact of each style on human-AI trust. Interestingly, their results show that contrastive explanation styles were highly rated in terms of trust and personal attachment by users compared to general explanations, which offer simpler and broader insights. Van Berkel et al. compared textual and scatterplot representations and showed that the usage of a scatterplot visualization led to lower perceived fairness [10]. Other work has compared the effects of different explanation methods [5, 79, 150], as for instance Van der Waa et al. who compared rule-based and example-based explanations in the domain of diabetes [150]. Explanations were generated artificially by domain experts

and the authors measured the impact of each explanation on users' trust and understanding. However, they were unable to identify significant factors in the explanation impacting the users' perception. Kulesza et al. proposed four different explanations based on various levels of soundness and completeness for a music recommender system [79]. The authors expected that too much completeness and soundness would confuse the users while their results showed that explanations with higher soundness and completeness increase the users' understanding of the recommender system.

To summarise, a majority of research studies suggest that explanations influence users' perceived trust and understanding in human-AI interactions. However, certain studies have failed to reveal the impact of explanations on specific aspects of user perception, such as confidence or comprehension, as shown in previous works [5, 109, 150]. Therefore, in line with recent guidelines for evaluating XAI applications [150], our proposed methodological framework focuses on various dimensions of user interaction with AI systems. Specifically, it investigates the effects of explanations on user trust, understanding, and satisfaction. Moreover, we combine metrics on both users' perceptual and behavioral to measure their trust and understanding when interacting with AI systems.

5.1.2 Guidelines and Metrics to Conduct User Studies

The evaluation of trust and understanding has been a prevalent topic in Psychology and Cognitive Sciences, which has led to numerous guidelines for conducting experiments. Hoffman et al. [66] surveyed several questionnaires from Social Sciences [73] and HCI [146] and combined them into a new satisfaction scale adapted for XAI. Cahour and Forzy [22] developed a trust scale based on three factors: reliability, predictability, and efficiency. This scale made of four questions directly asks users whether they are confident in the XAI system. Finally, Madsen and Gregor [97] proposed an eight-question scale measuring perceived technical competence and understandability.

Ribeira and Lapedriza [125], as well as Doshi-Velez and Kim [38], proposed to group users into three categories: (a) machine learning practitioners, (b) domain experts, and (c) laypeople. Based on these three categories, Doshi-Velez and Kim propose to distinguish between application-grounded and human-grounded evaluations. While the former represents real tasks with computer scientists or domain experts, the latter are simplified tasks such as giving humans access to an input and an explanation and asking them to simulate the model's prediction. Doshi-Velez and Kim also indicated that running evaluations with laypeople offers the advantage of (a) evaluating the impact of the explanations more broadly, and (b) facilitating the conduct

of the experiments since it is easier to control the factors of explanations in simplified tasks. In this part of the manuscript, we evaluate the impact of explanation techniques and their visualizations with laypersons. We also closely followed the advice from Van der Waa et al. by proposing an evaluation framework with several measurements, including trust, understanding, satisfaction, and completion time [150].

5.2 Method

In response to the recognized need for user studies and guidelines to evaluate explanation techniques in XAI, this section presents a methodological framework. This framework is designed to comprehensively assess user interactions with XAI systems through surveys, wherein participants engage in specific tasks. This provides valuable insights into the participants' perceptions and behaviors when interacting with AI models. To ensure a structured approach and reproducibility, we organized the surveys into three distinct phases: the introduction, the task rounds, and the post-questionnaires. Moreover, we introduce a range of scales and metrics, to quantify the impact of different variables. These measurements provide a clear picture of how various factors influence user behavior, trust, and understanding when interacting with AI systems. In the next chapter, we apply this methodological framework, alongside the defined scales and metrics to conduct a user study.

5.2.1 Methodological Framework

The proposed methodological framework consists of online surveys where participants are confronted with tasks. These tasks encompass a range of activities, including making predictions based on provided information (for example, predicting the risk of obesity based on lifestyle habits or determining the topic of a newspaper article based on its title). Participants may also be tasked with replicating the prediction of an AI model based on an explanation for a similar instance. Figure 5.1 outlines the process followed by such a survey. Given a task, the only difference among the surveys is the "Read Explanation" phase. Each survey is composed of three phases: (i) introduction, (ii) task rounds, and (iii) post-questionnaires.

Introduction. The survey begins with an introductory explanation of the tasks assigned to the user and the information used by the AI model to make recommendations. Then the user is asked questions to verify whether they understood the task.

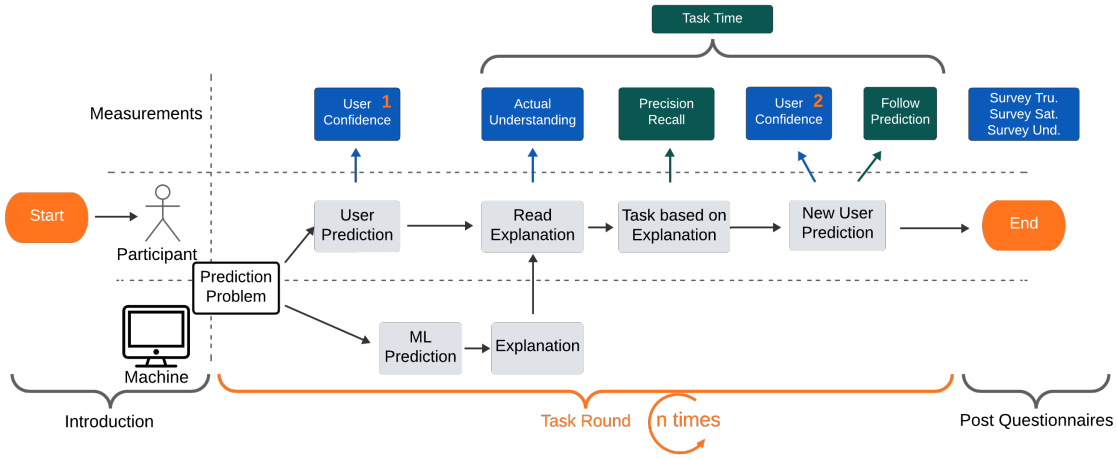


Figure 5.1 – Diagram depicting the experimental workflow proposed in this thesis and used to assess user perception and behavior when interacting with a given explanation technique. Behavioral measurement steps are indicated in green, while self-reported measurement steps are in blue. The task rounds are repeated n times.

Task Round. After the introductory explanations, users are presented with the prediction tasks. Each task is divided into two steps. First, the users are asked to make an assessment based on the given information. After stating their assessment, the participants have access to the AI model's prediction along with its associated explanation. Based on this explanation, we then ask the users to complete a task. This task may be to select the features, among the list of all possible features, that were used by the AI model to make its recommendation. Another task could be to ask the users to provide counterarguments or point out potential drawbacks related to the AI model's recommendation. In the second stage, users can reconsider their prediction and answer two questions on a 5 Likert scale. A Likert scale offers a range of responses, from "strongly disagree" to "strongly agree". Those questions gauge the extent to which users believe they understand the provided explanation and their level of confidence in their prediction.

Post-Questionnaires. After the prediction tasks, our methodology involves the utilization of three standardized questionnaires, each comprising 20 questions. Through these questionnaires, users are encouraged to report their trust, understanding, and satisfaction concerning the AI model. The proposed framework ends with two open questions, designed to get a deeper understanding of the users' perception: (1) "According to the scenarios you have seen and the corresponding explanation, how does the artificial intelligence tool predict?", and (2) "What was good in the explanation? What was bad in the explanation?". Participants should not be restricted by word limits and are encouraged to provide as many points as they feel necessary.

5.2.2 Scales & Metrics

To assess the impact of the independent variables, which are the factors or conditions we control in our study, we propose to employ a range of scales and metrics. In the context of this thesis, the independent variables may include different explanation techniques, their representations, or a combination of paradigms. We propose to evaluate users' behavior and perception regarding trust and understanding of the AI system through various metrics. The proposed methodology seeks to distinguish between the subjective experiences of users and the objective outcomes. Additionally, we measured self-reported satisfaction with the system, which serves as a dependent variable, that reflects how users' experiences are influenced by the independent variables. Furthermore, we recorded the time taken to complete the task round, another dependent variable, which can provide insights into user efficiency and task engagement. Figure 5.1 illustrates the points in the study where these parameters, both independent and dependent, are measured.

This comprehensive approach enables the analysis of the relationships between the independent variables and the observed dependent variables, providing a deeper understanding of how various factors impact users' interactions with the AI system.

Trust. Jarvenpaa et al. [71] have shown that trust influences several factors and should be considered as an essential factor for humans to successfully collaborate with computers. Therefore, we build upon the methodology of Broon and Holmes [16] to measure users' behavioral trust and adapt scales for self-reported trust to align with the XAI domain. We distinguish two categories of trust metrics:

— **Behavioral Trust (Follow Pred.)** This is the proportion of times users modify their initial prediction in favor of the AI's prediction. In this context, we consider only scenarios where the user's initial prediction diverged from the AI's prediction. This metric is particularly relevant when users are prompted to reevaluate their own predictions due to contrast with the AI's prediction, as it provides insights into the level of trust users place in the AI's recommendations.

— **Self-Reported Trust ($\triangle$ Confidence and Survey Tru.)** The assessment of perceived trust is composed of two distinct metrics. Firstly, it includes users' self-reported confidence in the AI model. Secondly, it resorts to a post-survey evaluation based on a questionnaire developed by Cahour and Forzy [22], as exemplified in Section 6.2.2.

- Δ **Confidence.** These are changes in self-reported trust before and after accessing AI predictions and explanations ('User Confidence 2' - 'User Confidence 1' in Figure 5.1).
- **Survey Tru.** This is based on the answers from a four-question questionnaire, originally developed by Cahour and Forzy [22], to assess users' perceptions of the AI system's reliability, predictability, and efficiency. The questionnaire has been adapted for application in XAI by Hoffman et al. [66], and participants rate their responses on a seven-point Likert scale.

Understanding. A widely accepted definition of a good explanation is its capacity to be understood by a human within a reasonable time frame, as defined by Lipton [91]. We thus gauge the users' comprehension of the model through various aspects divided into behavioral and self-reported metrics.

- **Behavioral Understanding (Prec. and Rec.)** Building upon the methodology proposed by Weld and Bansal [154], we assess users' behavioral understanding through a simple task. Therefore, it is important to note that the metrics may be adapted to the nature of the task. In the following chapter, we ask users to identify the features that have the highest impact on the classifier's prediction after viewing an explanation. Subsequently, we measure the behavioral understanding through two metrics:
  - **Precision (Prec.)** Measure of alignment between features identified by users and top features reported in explanations.
  - **Recall (Rec.)** Measure of users' ability to identify all the influential features indicated in explanations.
- **Self-Reported Understanding (Immediate Und. and Survey Und.)** The measurement of perceived understanding combines self-reported comprehension during explanation review and post-survey assessment using Madsen and Gregor's questionnaire [97].
  - **Immediate Und.** Self-reported comprehension of the system's prediction on a five-point Likert scale while looking at the explanation.
  - **Survey Und.** Adapted questionnaire by Madsen and Gregor [97] on perceived technical competence and understandability, using a five-point Likert scale for consistency.

Additional measures. According to Lipton's definition of a good explanation [91], the time taken to understand the explanation is a key element for measuring its quality. Therefore,

we consider the time taken to understand explanations and user self-reported satisfaction as additional measurements to provide a comprehensive evaluation.

— **Task Time.** The time required to complete the requested task based on the explanation.

— **Satisfaction (Survey Sat.)** Users' satisfaction with the AI model is assessed using the Explanation Satisfaction (ES) questionnaire by Hoffman et al. [66], comprising eight items rated on a seven-point Likert scale.

5.3 Conclusion

In this chapter, we have delved into an important yet underexplored area within the field of explainable AI (XAI): user-centered evaluation. We have highlighted a pressing need to measure the real impact of explanation methods on users. To address this gap, we have introduced a set of metrics and a robust methodological framework that can be widely applied across various contexts. This user-centric methodology, tailored for conducting user studies, relies on specific indicators designed to assess understanding, trust, and other aspects related to the explainability of AI systems. It aims to provide a standardized framework for gauging the effects of explanation methods on users, enabling objective and reproducible result comparisons.

Our approach aims to contribute to the advancement of XAI by placing the user at the center of evaluation. With this powerful tool, we are well-prepared to measure the actual influence of explanation methods in AI on users. By understanding how users interact with explanations and how this interaction affects their perception of AI systems, we as a community are more prepared. To shape the future of these technologies we will therefore be able to make them more comprehensible and trustworthy to a wider audience. In the upcoming chapter, we apply this methodology through a user study. This study is designed to measure the impact of three different explanation methods and their representations on users. It aims to provide valuable insights for comparative analysis, enhancing the transparency and acceptance of AI systems.

CHAPTER 6

IMPACT OF EXPLANATION TECHNIQUES AND REPRESENTATIONS ON USERS

Contents

6.1 Explanation Techniques and Representations 123
6.1.1 Datasets & AI models 123
6.1.2 A Common Representation for Explanations 124
6.2 Method 127
6.2.1 Task 128
6.2.2 Scales & Metrics (Illustration for One User) 129
6.2.3 Participants 132
6.3 Results 134
6.3.1 Understanding 134
6.3.2 Trust 137
6.3.3 Additional Measurements 139
6.3.4 Perception vs. Behavior 140
6.3.5 Open Questions 141
6.4 Discussion 142
6.4.1 Impact of Explanation Technique 143

6.4.2 Impact of Representation 144
6.4.3 Limitations & Future Work 145
6.5 Conclusion 146

As we highlighted in the previous chapter, there is a clear lack of comprehensive user studies that investigate how users perceive and interact with these explanations. As a result, the focus of the XAI community is starting to shift toward comprehensive user studies. In the first part of this thesis, our focus was primarily on optimizing explanations from a data perspective. In each chapter of the first part, we focused on a specific explanation paradigm, developing techniques to enhance the quality of explanations. In this chapter, we explore a distinct aspect, namely the impact of the explanation paradigm on users' trust and understanding. Our objective is to gauge how these factors shape users' perceptions and behavior. Consequently, we aim to provide valuable insights that can inform the design of AI interfaces and user-centered XAI systems. We shed light on these issues by addressing the following research question:

RQ1: Which local explanation technique, *i.e.,* feature-attribution, rule-based, or counterfactuals, provides the best explanations in terms of users' trust and comprehension of the AI model?

Moreover, we examined eight different XAI toolkits[1] which employ at least one of the three aforementioned explanation types, and note that each explanation method is commonly represented under a certain form (graphical for feature-attribution explanations and textual for counterfactual and rule-based). This leads us to this research question:

RQ2: Does the explanation's visual representation impact the users' trust and understanding?

To answer these research questions, we employed the methodological framework proposed in the previous chapter to conduct two user studies. These studies evaluate the impact of the three aforementioned explanation techniques and two visual representations (graphical vs. text) on users' trust and understanding. By applying the set of scales and metrics introduced earlier, we obtain multiple findings. Notably, the choice of explanation technique significantly impacts the users' comprehension. Moreover, the selection between graphical and textual representation has a greater impact on users' trust. In summary, we discover that a graphical representation for explanations is perceived as more trustworthy, whereas rule-based explanations are the most effective at conveying the most important features of the AI decision process. Conversely, we find that users confronted with counterfactual explanations exhibit an understanding of the AI

1. AI360, Dalex, H2O, eli5, InterpretML, What-if-Tool, Alibi, Captum.

model comparable to our control group, that is, users who receive an AI-assisted prediction but with no explanation.

Most of the research presented in this chapter was the subject of the paper: Impact of Explanation Techniques and Representations on Users Comprehension and Trust in Explainable AI, submitted at the CHI 2024 conference.

6.1 Explanation Techniques and Representations

We next elaborate on the two datasets used for the studies and the implementation details of the explanation methods. Further, we discuss and present the chosen representation approaches.

6.1.1 Datasets & AI models

Dataset	Features		Instances
	Numerical	Categorical	
Compas	1	7	5364
Obesity	2	13	2111

Table 6.1 – Datasets composition.

Our evaluation is conducted on two widely used datasets among the XAI community, namely the COMPAS [18] and Obesity datasets [114]. These datasets represent two domains where explainability is often deemed crucial: law and healthcare. COMPAS is a tabular dataset used to train a model that predicts a criminal defendant's likelihood of re-offending. The data comprises American defendants.

The Obesity dataset [114] is used to predict the risk of developing obesity based on an individual's Body Mass Index (BMI) and answers to various questions. The data originates from Colombia, Peru, and Mexico. We chose those datasets in line with the recommendations from Van Berkel et al. [10] and Van der Waa et al [150], who suggest that a meaningful application-agnostic XAI evaluation should preferably include more than one domain, and strike a balance between simplicity – users should understand the domain of the AI –, and plausibility – the task should be difficult enough to justify the need for AI assistance. In the spirit of these principles, we conducted some basic feature selection on both datasets to reduce the number of features. We, for instance, removed the BMI index from the obesity dataset, which

otherwise would have oversimplified the prediction task. Table 6.1 contains the final number of features and instances for both datasets as used in our experiments. As both these datasets are public, our results could be used as a baseline for future researchers aiming to evaluate the effectiveness of their explanation method.

AI Model and Explanations. We used our experimental datasets to train an AI model based on a Multi-Layer Perceptron (MLP) classifier [2]. On Compas, the AI model was trained to predict the risk of recidivism among four classes: 'very low risk', 'low risk', 'high risk', and 'very high risk'. The original Obesity dataset considers seven weight categories which we simplified into four ordinal classes (to stay consistent with Compas): 'underweight', 'healthy', 'overweight', and 'obese'. In both cases, we trained the MLP on 70% of the instances and evaluated it on the remaining 30%. We obtained an accuracy of 67% and 78% on Compas and Obesity, respectively. Then, for each instance in the test set, we generated three different types of explanations: a feature-attribution explanation based on LIME [123], a rule-based explanation based on Anchors [124], and a counter-factual explanation using the Growing Fields algorithm [32]. We set the default parameters for each of these methods except that (a) we changed the discretization routine in Anchors to take into account the improvements proposed in Chapter 2, and (b) we reported the importance of all features in the LIME explanation – contrary to the default configuration that only picks the top 6. Moreover, while nothing guarantees that the sum of the coefficients of the linear model trained by LIME is consistent with the final probability prediction of the MLP classifier, we normalized the coefficients with the probability prediction to be more user-friendly.

For each dataset, we selected five target individuals in the test set to be presented to the user — one for each of the four predicted classes plus an additional individual used as an example. Figure 6.1 depicts how information about an individual is shown to the user for both datasets. The grey column represents the various features while the corresponding prisoner or patient data are in the second column. Code, datasets, and results are available on GitHub [3].

6.1.2 A Common Representation for Explanations

As highlighted in Section 5.1.1, feature-attribution, rule-based, and counterfactual explanations reveal different aspects and insights about an AI's prediction process. Therefore, these explanation types are conveyed using different representations, which also depend on the AI

2. `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html`

3. `https://anonymous.4open.science/r/user_eval-1776`

Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

Gender	Male
Age	26
Race	Other
Number of juvenile major offences	0
Number of juvenile minor offences	4
Number of previous arrest	3 or more
The degree of the charge	major offences
Description of the charge	Aggravated assault with a deadly weapon

Figure 6.1 – Example of two individuals presented to the users for the Obesity (left) and COMPAS (right) datasets. The first column (in grey) represents the features description and the second column (in white) is the patient's or defendant's information.

agent's input data (e.g., image, text, etc.). When it comes to tabular data, existing XAI toolkits [4] opt for a graphical representation based on bars for feature-attribution explanations – as illustrated in Figure 6.2. Conversely, for rule-based and counterfactual explanations, the most common representation is natural language (see Figure 6.2). In order to control for this visual representation in our experiments, users are confronted with common graphical and textual representations for all the explanation types.

Graphical Representation. For each explanation method, we depict the graphical representation through diagrams. The explanation came along within a detailed paragraph reviewed by 20 different people—including 9 computer scientists and 11 laypeople—to verify its comprehensiveness and usefulness. As our AI model predicts four ordinal target outcomes, which range from underweight to obesity and no risk to high risk, we choose a common graphical representation that depicts the spectrum of classes on the x-axis and adds a different background color to the region covered by each of the classes.

— As proposed by SHAP [94] for feature-attribution explanations, the x-axis depicts the contribution of each feature to the predicted class in the form of a directed bar. The length of the bar depends on the magnitude of the attribution, whereas its direction tells towards which side of the spectrum the feature shifts the AI model's prediction (underweight vs. obese, low risk vs. high risk). Furthermore, SHAP computes the attribution score of a limited number of features to reduce the complexity of the explanation. Unlike SHAP,

4. AI360, Dalex, H2O, eli5, InterpretML, What-if-Tool, Alibi, Captum.

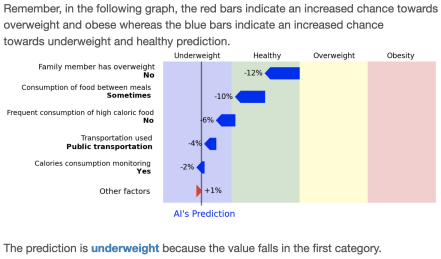
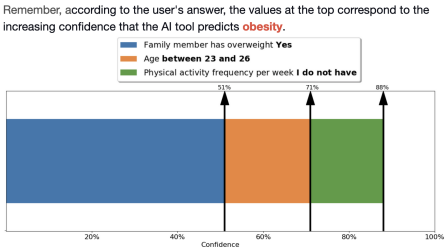
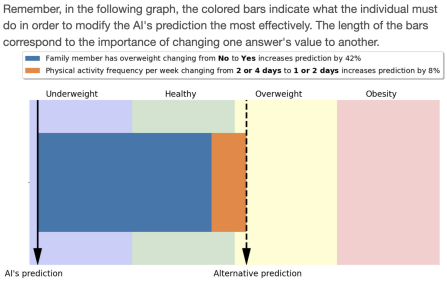
Explanation technique	Graphical representation	Textal representation
Feature-attribution		Based only on the above information, the AI tool has predicted **underweight**. Remember, the AI associates a score to each response. We obtain a value between 0% and 100% by summing these scores. This value falls into one of four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), and **obesity** (above 75%). • First, since **no** family member of this individual **suffers** from overweight, the score decreases by 12%. • Second, since the individual **sometimes** consumes food between meals, the score decreases by 10%. • Third, **no consuming frequently** high caloric food decreases score by 6%. • Fourth, using **public transport** decreases the score by 4%. • Fifth, **monitoring** her calories consumption decreases the score by 2%. Combining all the **other answers** increases the score by 1% and the final value is 17% implying an **underweight** prediction.
Rule-based		Based on the above data, the artificial intelligence (AI) tool has predicted **obesity**. • First, because a family member **suffers** from overweight. • Second, she is **aged between 23 and 26 years old**. • Third, she **doesn't practice** physical activity weekly. All together, it brings an AI's confidence of 95% for this **obesity** prediction
Counterfactual		Based on the above data, the AI tool has predicted **underweight**. To turn the AI prediction into an **overweight** prediction, the individual should **have** (at least) a family member **suffering** from overweight and practice physical activity **1 or 2 days** instead of **2 or 4 days** per week.

Figure 6.2 – Different explanations for a given individual on the Obesity dataset, as split by explanation technique and representation.

our representation groups features with a marginal attribution score under an artificial feature labeled 'Other features'. The aggregated attribution of this label is the sum of the attribution scores of these features. Assuming a ranking of features based on the absolute value of the attribution scores, a feature is considered marginal as soon as the absolute value of its attribution score is less than half the absolute value of the score of the previous feature – all subsequent features in the ranking are deemed marginal from that point. For instance, in Figure 6.2, the features that impact less than 2% are grouped into the last bar and their cumulative attribution score equals 1% toward the obesity class.

— For rule-based explanations, we took inspiration from the representation proposed by Molnar [104]. This representation uses stacked bars as well, where each condition of

the rule is assigned to a bar having a length proportional to the increase in confidence provided by the condition. As an example, imagine the explanation rule in Figure 6.2, stating that "(a) having family antecedents of obesity, (b) an age between 23 and 26, (c) and practicing no physical activity" incur an "obese" prediction with 90% confidence. The blue bar tells that condition (a) on its own predicts obesity with 50% confidence; conditions (a) and (b) increase the confidence to 71%, and all three conditions increase the confidence to 90%.

— For counterfactual explanations we use stacked bars. Each feature in the explanation is associated with a bar. For each feature, the explanation incurs a change in the feature value, hence the length of each bar is proportional to the change incurred by the feature in the model's prediction. In other words, the length of the bar tells us to what extent changing the feature's value shifts the black box answer from one predicted class to another – the counterfactual class. For instance, the counterfactual explanation from Figure 6.2 depicts that if the patient: "(a) had family antecedents of obesity, and (b) practiced less often physical activity" then the AI model would have predicted the patient as being "overweight".

Text Representation. For all explanation types we present the explanation using a bulleted list, where each item describes the effect of each feature on the model answer. This effect can be an increase in the confidence of the prediction (for rule-based explanations), how much the feature contributes to the AI model prediction (feature-attribution explanation), or how sensitive is the AI model prediction in regards to the changes in the input features (counterfactual explanation). For feature-attribution explanations, we used colors to highlight the direction of the impact of each feature. Finally, we highlighted the instance's responses (obesity state, charge) by showing the text in bold.

6.2 Method

While the XAI community has proposed multiple post-hoc explanation methods that fall within the categories of feature attribution, rules, and example-based instances, no user studies have compared the effectiveness of these explanation styles. Hence, our first research question is **RQ1:** "Which local explanation technique, *i.e.,* feature-attribution, rule-based, or counterfactuals, provides the best explanations in terms of users' trust and comprehension of the AI model?". Existing works have shown that explanations improve the users' ability to comprehend a model [123, 5]. Hence, this question underlies our first hypothesis; (H1) explanations improve

the users' trust and understanding of a model. In addition, it has been suggested that explanations based on feature attribution may struggle to consistently assist users in understanding a model [119]. Conversely, decision rules have demonstrated high efficiency in helping users understand the inner mechanism of a model [124, 5]. This leads to our second hypothesis; (H2) rule-based explanations improve users' comprehension of a model the most. Finally, concerning trust, existing works have failed to show significant trust improvement when using feature-attribution [119], as well as exemplars and rule-based explanation [150]. We therefore follow a more explorative approach to studying the impact of the explanation technique on trust and do not propose a hypothesis on this aspect.

As suggested in [10] and [28], the visual representation of an explanation can also impact the users' perception (e.g., trust, understanding, fairness). This leads to our second research question **RQ2:** "Does the explanation's visual representation impact the users' trust and understanding?". As there exists a general tendency to represent feature-attribution explanations graphically and rule-based as well as counterfactual explanations textually, our hypotheses are as follows: for feature-attribution techniques, a graphical representation is preferred (H3), whereas the most preferred explanation for rule-based and counterfactual is text-based (H4).

Therefore, our two independent variables are (i) the explanation style—feature-attribution, rule, and counterfactual, and (ii) the representation—graphical and textual. Conversely, the dependent variables are the users' understanding, satisfaction, and trust in the model.

6.2.1 Task

Our user studies consist of 14 online surveys in which participants are confronted with four prediction tasks. These tasks aim to predict either the risk of recidivism of a defendant given their profile or the risk of obesity of a person given some information about their habits (refer to Figure 6.1). To perform those predictions, users count on the recommendations of the AI models described in Section 6.1.1.

These surveys were developed through the Qualtrics platform [5], with variations in the dataset (Compas and Obesity), explanation techniques (feature-attribution, rule-based, counterfactual) and representation methods (graphical vs. textual). For each dataset, we also created a control group in which no explanations were provided following the AI model's prediction. Furthermore, within the context of the three explanation techniques, two different visualizations were given.

5. https://www.qualtrics.com/

Figure 5.1 outlines the process followed by each of these surveys. With a specific dataset as the common ground, the only difference across the seven surveys is the explanation segment. In other words, except for the explanation, every step of the survey is exactly the same. Each survey is composed of three phases: (i) introduction, (ii) task round, and (iii) post-questionnaires as defined in Section 5.2.1. The task asked to the users is to select based on the explanation, the features among the list of all possible features, that were used by the AI model to make its recommendation.

6.2.2 Scales & Metrics (Illustration for One User)

In this section, we provide a detailed example of how we employed the scales and metrics introduced in Section 5.2.2 for one user from the rule-based explanation group. This example is designed to provide the reader with a detailed explanation of how we assessed various facets of user behavior and perception. We recall that Figure 5.1 shows the times at which these parameters are measured. For this illustration, let us refer to this user as "User J." User J participated in predicting the risk of obesity in response to four distinct scenarios, and their responses are reported in Figure 6.3.

	1st User's **Prediction**	1st User's **Confidence**		AI's Prediction	Top Features According to the Rule-based Explanation	Top Features According to the User	2nd User's **Prediction**	2nd User's **Confidence**	Perceived **Understanding**
Q1: What is the risk of obesity? (Scénario 1)	No Risk	2/5		Low Risk	• Monitoring Calory • Consumption of High-Caloric Food	• Monitoring Calory • Age • Gender	Low Risk	3/5	3/5
Q2: What is the risk of obesity? (Scénario 2)	Low Risk	3/5		Medium Risk	• Family Member has Overweight • Physical Activity Frequency	• Family Member has Overweight • Physical Activity Frequency	Medium Risk	4/5	4/5
Q3: What is the risk of obesity? (Scénario 3)	Medium Risk	1/5		No Risk	• Monitoring Calory • Physical Activity Frequency • Age	• Monitoring Calory • Age	Low Risk	3/5	5/5
Q4: What is the risk of obesity? (Scénario 4)	High Risk	4/5		High Risk	• Consumption of High-Caloric Food • Family Member has Overweight • Transportation Used	• Physical Activity Frequency • Consumption of High-Caloric Food • Smoke	High Risk	3/5	1/5

Figure 6.3 – Example of answers from participant "User J" from the rule-based explanation group. The values within the columns "1st User's Confidence", "2nd User's Confidence", and "Perceived Understanding" are on a 5-Likert scale.

User's Initial Prediction and Confidence. In Figure 6.3, User J's initial predictions, scaled from 1 (no risk) to 4 (high risk), are accompanied by their initial confidence levels, measured on a 5-point Likert scale. The Likert scale spans from "strongly disagree" to "strongly agree."

User J's initial predictions are shown in the "1st User's Prediction" column, and their initial confidence is recorded in the "1st User's Confidence" column.

AI Model Predictions and Explanations. User J's predictions are followed by the AI model's predictions and associated explanations, presented as depicted in Figure 6.2. These explanations comprise lists of the most influential features considered by the AI model for each prediction scenario. For example, in Figure 6.2, the most important features for the feature attribution are *Family member has overweight*, *Consumption of food between meals*, *Consumption of high caloric food*, *Transportation used*, and *Calories consumption monitoring*. In contrast, for counterfactual, this is only the *Family member has overweight* and *Physical activity frequency* while rule-based also includes the *Age* feature.

User's Final Prediction and Confidence. During the task round, User J was asked to select, from the list of features, which features they considered most important for the AI model's prediction. Subsequently, User J was given the opportunity to reevaluate their prediction in the "2nd User's Prediction" column and provide their final confidence in their prediction in the "2nd User's Confidence" column.

User's Perceived Understanding. User J was also asked to rate their "Perceived Understanding" on a 5-point Likert scale to indicate their understanding of how the model made the prediction.

Metrics Calculation. The metrics for User J's responses were calculated as follows:

— **Δ-Confidence:** The Δ-Confidence was computed by subtracting the initial confidence from the final confidence for each scenario. User J's Δ-Confidence values are 1, 1, 2, and -1 for the four scenarios. The average Δ-Confidence for User J is thus $3/4$.

— **Behavioral Trust (Follow Pred.):** We assessed behavioral trust by tracking instances where the user modified their initial prediction to match the AI model's prediction. It is important to note that we only considered scenarios where the user's initial prediction differed from the AI model's prediction. Thus, User J modified their initial prediction to align with the AI model's prediction in 2 out of 3 such scenarios, resulting in a behavioral trust score of $2/3$.

— **Immediate Understanding:** User J's immediate understanding is the average value of their Likert-scale ratings for understanding across all four scenarios. In this case, it is (3 + 4 + 5 + 1) / 4, which equals 13/4.

— **Behavioral Understanding (Precision and Recall.):** To measure User J's precision and recall, we compared the list of features they identified as important to those highlighted in the explanation for each scenario. The precision and recall values for each scenario were calculated as follows:

Scenario Q1:
— Precision = 1/3 (User identified three features, one matched AI explanation),
— Recall = 1/2.

Scenario Q2:
— Precision = 1 (User and AI explanation lists are identical),
— Recall = 1.

Scenario Q3:
— Precision = 1 (User identified 2 features, both matched AI explanation),
— Recall = 2/3.

Scenario Q4:
— Precision = 1/3 (User identified 1 feature, which matched AI explanation),
— Recall = 1/3.

Please note that these are simplified examples, and in practice, the lists of important features in explanations are typically longer.

Trust	What is your confidence in the tool? Do you have a feeling of trust in it?	Are the actions of the tool predictable?	Is the tool reliable?	Is the tool efficient at what it does?		Average
User J's Answers	5/7	6/7	3/7	4/7		4.5/7

Figure 6.4 – Example of answers from one participant to the Trust survey adapted from Cahour and Forzy [22]. We measure the users' perceived trust in the AI system on a scale from 1 to 7.

Post-Questionnaires. In Figure 6.4, we present an example of a survey measuring User J's perceived trust in the AI system. This survey was adapted from Cahour and Forzy [22] and employed a Likert scale ranging from 1 to 7. The average of User J's responses to the four

survey questions provides a representation of their perceived trust, which, in this case, is 4.5 out of 7. Similar procedures were followed for the understanding and satisfaction questionnaires, each consisting of eight questions rated on a five-point Likert scale.

6.2.3 Participants

We recruited participants through the Prolific Academic platform. First, we restricted participation to crowd-workers with at least a high school degree to guarantee a reasonable response quality. Second, we decided not to limit ourselves to a particular geographical location to promote diversity in our sample. Finally, we ensured that participants could participate only once in our study to avoid situations where a participant is in two groups (e.g., control and feature-attribution). After accepting the task, participants were redirected to the corresponding Qualtrics survey, starting with a short introduction to the algorithm and the dataset. Based on a pilot evaluation with 20 people, we estimated a completion time of 20 minutes for non-control groups and 15 minutes for the control group. We remind the reader that the control group has no explanation to extend the AI prediction, therefore, these participants should be faster at filling out the survey. Participants were paid £9.30 per hour, which translated into a payment of £2.25 for the control-group participants, and £3.10 for the non-control group.

To limit Type II errors and not reject the null hypothesis when it is false, we determined the number of respondents on the basis of a power calculation using G*Power [130]. Given the exploratory nature of our investigation, we used medium-to-large effect sizes ($f^2 = 0.2$), an alpha level of 0.05, and a power of 0.8, in line with established methodological recommendations [62]. Based on our a priori multiple linear regression model with two predictors, the required minimum group size is 107 participants. For the sake of precaution and to maintain consistency, we finally recruited 280 participants – 140 participants per dataset, or 20 participants per combination of explanation technique and visual representation.

Table 6.2 presents the demographic information about our participants. Our study follows a between-subject design, meaning that each participant interacts with one representation and one surrogate model.

Following the introduction of the task, we assessed whether the participants had actually read and comprehended the basic information presented through two questions: 'How is Body Mass Index calculated?' for the obesity domain and 'Why is recidivism risk calculated?' for the recidivism dataset. We found 10 incorrect answers for the first question and 30 participants who responded erroneously for the second question. This question had the form 'The algorithm calculates the risk of obesity (resp. recidivism) for an individual by;'. We asked additional users

Domain	**Healthcare**		**Law**	
Factor	N	% sample	N	% sample
Gender				
Female	66	47.14	66	47.14
Male	62	44.29	74	52.86
Prefer not to say	1	0.71	0	0.0
Consent revoked	11	7.86	0	0.0
Age				
< 20	10	7.14	11	7.86
$20 < 30$	81	57.86	88	62.86
$30 < 40$	24	17.14	27	19.29
$40 >$	14	10.0	14	10.0
Nationality				
Africa	45	32.14	37	26.43
Asia	2	1.43	2	1.43
Australia	0	0.0	1	0.71
Europe	77	55.0	82	58.57
North America	5	3.57	15	10.71
South America	0	0.0	3	2.14
Ethnicity (simplified)				
Asian	2	1.43	2	1.43
Black	37	26.43	30	21.43
Mixed	10	7.14	9	6.43
Other	3	2.14	8	5.71
White	77	55.0	91	65.0
Highest education				
Doctorate degree	3	2.14	1	0.71
Graduate degree	27	19.29	24	17.14
High school diploma	47	33.57	37	26.43
Technical college	3	2.14	14	10.0
Undergraduate degree	49	35.0	64	45.71

Table 6.2 – Overview of participants' demographic factors.

to participate in our study until we had 20 responses for each group that validated our two understanding questions. This resulted in a final set of 280 participants.

6.3 Results

We present our findings in four sections. We begin by examining the impact of the application domain (*i.e.*, dataset), explanation technique, and representation on users' understanding in Section 6.3.1. Then, we assess the influence of these factors on users' trust in the AI agent in Section 6.3.2. Additional results concerning satisfaction and completion time are discussed in Section 6.3.3. Subsequently, in Section 6.3.4, we explore the correlation between behavioral and perceived measurements. We conclude with Section 6.3.5, where we present an in-depth qualitative study of the users' perception based on the answers to the open questions. For a more comprehensive understanding of our study, including code, participants' comments, survey details, and analysis notebook, these resources are available on GitHub [6].

	Understanding							
	Recidivism				**Obesity**			
	Self Report		Behavioral		Self Report		Behavioral	
	Immediate	Survey	Prec.	Rec.	Immediate	Survey	Prec.	Rec.
Explanation Technique	0.87	1.20	16.24^{***}	1.58	3.75^{*}	1.35	31.42^{***}	6.37^{***}
Representation	0.96	0.36	0.13	3.00^{-}	0.14	0.55	0.05	2.85^{-}
Age	1.07	0.01	1.88	0.10	0.16	0.06	6.41^{*}	0.02
Education	1.63	0.93	0.94	0.43	0.50	0.34	0.25	1.31
Gender	0.54	1.07	0.35	0.30	0.14	0.03	0.18	0.36
Technique:Representation	0.28	0.87	1.12	0.74	0.48	0.16	0.35	4.99^{**}

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Table 6.3 – F value of the ANOVA Table with understanding measurements grouped by domain and self-reported and behavioral metrics. 'Immediate' corresponds to the perceived comprehension when facing the explanation while 'Survey' represents the perceived understanding measured through the post questionnaire. 'Prec.' and 'Rec.' are respectively the Precision and Recall between the features indicated by the participant and the explanation technique. 'Technique:Representation' denotes the interaction between the explanation technique and the visual representation.

6.3.1 Understanding

To discern the factors that impact users' understanding of the AI agents, we fit a linear model and conduct an ANOVA analysis for each application domain (Recidivism and Obesity). The linear model incorporated six predictors to estimate four metrics. Those predictors included

6. `https://anonymous.4open.science/r/user_eval-1776`

demographic data (age, gender, education level), along with explanation technique and visual representation. These predictors were categorized to construct the model. The ANOVA f-scores of each predictor and target metric can be found in Table 6.3. We show plots only for results that are considered statistically significant according to the ANOVA analysis. The target metrics are elaborated upon in Section 5.2.2.

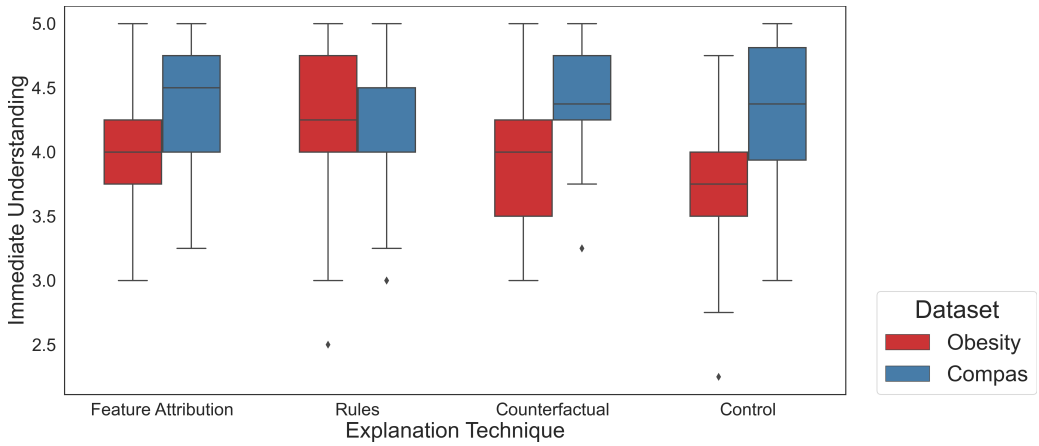


Figure 6.5 – Perceived understanding of the users **(Immediate Und.)** for both the Obesity and Recidivism datasets based on the explanation technique.

We first observe that the users' self-reported understanding of the AI system—based on a post questionnaire **(Survey)**—does not vary across the different explanation techniques, visual representations, and demographic categories. These observations hold both for the recidivism and obesity datasets.

Conversely, we note that the impact of the chosen explanation technique on the users' perceived understanding when exposed to explanations **(Immediate Und.)** is statistically significant *($p<0.05$)* for the Obesity dataset. Figure 6.5 depicts the users' perceived understanding of the AI system across the explanation methods for both domains. While statistically significant differences were not found for the recidivism domain, users confronted with rule-based explanations in the obesity domain report a better understanding of the AI model. Notably, on the obesity domain, the third quartile of the Immediate understanding for the rule-based group was around 4.7 (on a scale of 1 to 5), compared to approximately 4.2 for the counterfactual and feature-attribution groups. In contrast, the control group (without explanation) exhibited a lower self-reported understanding of the AI model (around 4).

We then gauge the behavioral understanding through the precision **(Prec.)** and recall **(Rec.)**. These scores measure the alignment between the features identified as important by the explanation and features marked as important by the user for prediction. Notably, it might

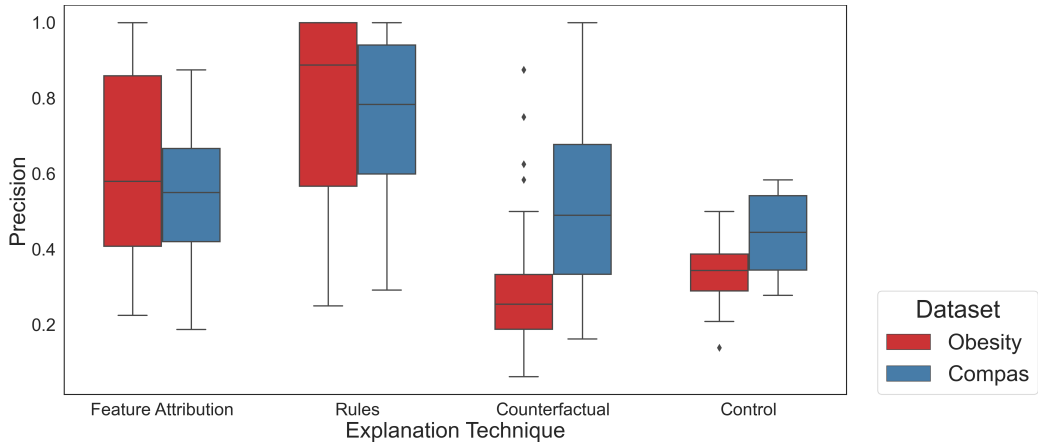


Figure 6.6 – Behavioral precision between the features indicated as important by the users for the AI's prediction and the important features indicated in the explanation. Results are shown for the two domains, the tree explanation methods, and the control group.

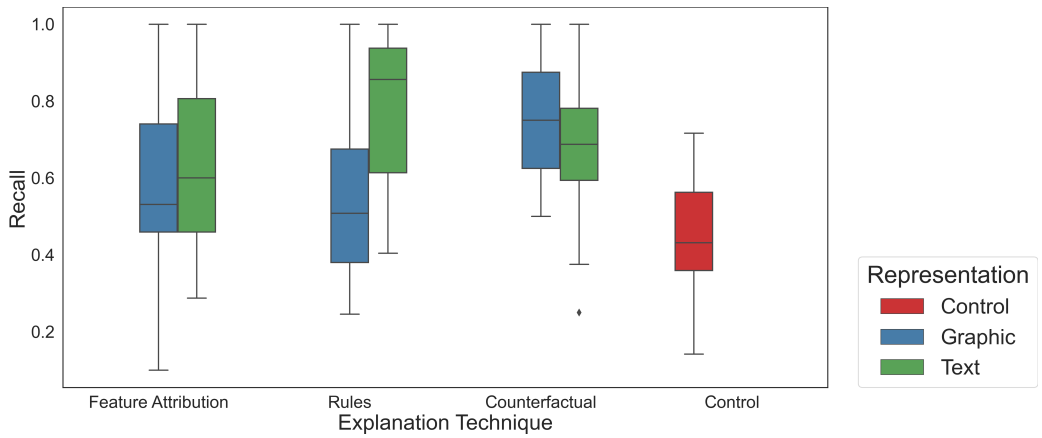


Figure 6.7 – Users' understanding for each explanation technique and representation, computed as the users' recall of the important features according to the explanations in the Obesity dataset.

seem straightforward for a researcher in interpretability to identify the features highlighted as significant by the explanation. However, our findings demonstrate that this is not the case for laypersons. The fact that precision and recall values are far from 1 suggests two aspects. Firstly, participants might exhibit a certain level of distrust in the explanation, leading them to choose features they personally consider impactful. Secondly, they could anticipate a higher degree of complexity, which prompts them to indicate additional features as important. Table 6.3 highlights that precision is significantly affected by the explanation method for both domains *($p<0.01$)*. Figure 6.6 depicts the precision across domains and explanation methods, revealing that rule-based explanations yield the highest precision score in the healthcare domain (median precision of 0.9). On the contrary, counterfactual explanations resulted in poor performances comparable to the control group (precision 0.3). Lastly, feature-attribution explanations fall in between, with median precision scores above 0.5 for both domains. We also note that according to the ANOVA table, the age of the participant significantly impacts the precision in the Obesity dataset.

We also observe in Table 6.3 that the explanation technique and the visual representation exhibit a statistically significant impact on recall for the healthcare task. Figure 6.7 shows the recall of the participants per explanation technique and grouped by representation on the obesity domain. This suggests that for feature-attribution and rule-based explanations, users facing a textual representation are better at selecting important features for the classification than the participants facing a visual presentation. In contrast, users with counterfactual explanations are better at identifying important features when reading a graphical representation than a text representation. We finally note that participants who had an explanation actually achieved a higher recall than those in the control group.

6.3.2 Trust

We then measure the impact of the different factors and domains on the users' trust in the AI system. We proceed similarly as in the evaluation of the user's comprehension. That is, we fitted a linear model on our observations with the same set of predictors and the three metrics defined in Section 5.2.2. We measured the corresponding f-value and reported it in Table 6.4.

We first analyze the users' self-reported trust in the AI model's prediction through the post questionnaire **(Survey)**. Table 6.4 reveals that the perceived trust in the AI system does not show statistically significant differences across the different explanation techniques and visual representations.

	Trust					
	Recidivism			**Obesity**		
	Self Reported		Behavioral	Self Reported		Behavioral
	Δ Confidence	Survey	Follow Pred.	Δ Confidence	Survey	Follow Pred.
Explanation Technique	1.40	0.03	0.78	0.12	0.42	0.38
Representation	0.04	0.32	0.00	8.22^{**}	0.55	0.12
Age	0.46	0.18	2.76^{-}	0.06	0.70	0.00
Education	0.13	1.82	0.34	2.14^{-}	0.69	0.63
Gender	2.16	1.35	0.31	0.12	2.32	1.11
Technique:Representation	0.35	1.23	0.75	0.26	0.23	3.55^{*}

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Table 6.4 – F value of the ANOVA Table with trust measurements grouped by domain and by self-reported and behavioral metrics. The column 'Δ Confidence' is the difference between the self-reported trust before and after facing the AI's prediction. 'Survey' corresponds to the perceived trust in the model assessed through the post-questionnaire. 'Follow Pred.' is the proportion of times the participants changed their prediction to follow the AI's. 'Technique:Representation' refers to the interaction between the explanation technique and the visual representation.

The results in Table 6.4 also suggest that the explanation visual representation has an impact on the difference in self-reported trust before and after seeing the explanation **(Δ Confidence)**. This impact is statistically significant with a p-value inferior to 0.01 for the Obesity dataset. It is noteworthy that, 56% of the users' initial predictions aligned with the AI model's prediction for the Recidivism dataset and 39% for the Obesity dataset. Thus, we limit our evaluation of self-reported trust to scenarios where participants had to adjust their initial predictions. This decision was based on the understanding that assessing behavioral confidence is relevant when participants are prompted to reconsider their own predictions because the AI predicted a different class. Figure 6.8 indicates that for the Obesity dataset, participants exposed to a graphical representation reported an increase in post-explanation trust in their prediction.

Additionally, we conducted a detailed examination of the participants' perceived confidence splitting participants into two groups. These groups are based on whether their initial prediction matched the AI model's prediction. We conducted this in-depth analysis because participants' confidence in the AI system could be influenced by the matching between their predictions and those of the AI system. This analysis revealed that, in the Obesity dataset, participants with higher educational levels and who initially disagree with the AI model's prediction experienced

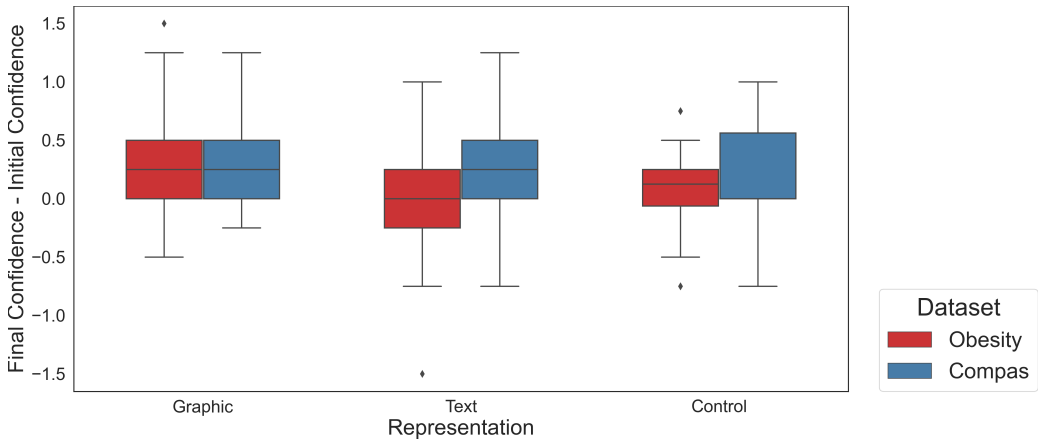


Figure 6.8 – Difference between the self-reported confidence in the users' prediction after and before seeing the AI's prediction and explanation (when provided). Results are shown for the control group and for each dataset and representation.

a decrease in confidence. Conversely, for participants in the Compas dataset, we observed that when the AI model confirmed female participants' prediction, their confidence increased less compared to male participants.

Finally, we find a significant interaction ($p<0.05$) between explanation technique and visual representation on participants' behavioral trust, *i.e.*, users who changed their initial prediction to match the AI agent's **(Follow Pred.)** for the Obesity dataset. The average users' behavioral trust for different explanation methods and representation in the healthcare domain is depicted in Figure 6.9. We observe that users without explanations or with counterfactual explanations are more prone to follow the prediction of the AI system. This suggests that users with feature-attribution and rule-based explanations have lower confidence in the model's prediction.

6.3.3 Additional Measurements

As stated in Section 5.2.2 we also evaluate perceived satisfaction and completion time. First, we measure the participants' self-reported satisfaction through the Explanation Satisfaction questionnaire **(Satis.)**. Second, we measure the completion time required by the user to interpret the explanation, that is, the time to indicate the most important features for the prediction **(Time)**. We fitted a linear model with six factors to predict these two metrics for both domains and report the f-score in Table 6.5.

This table shows that self-reported satisfaction in the AI system is not impacted by the presence of explanations or by the demographic attributes of the participant. Conversely, we

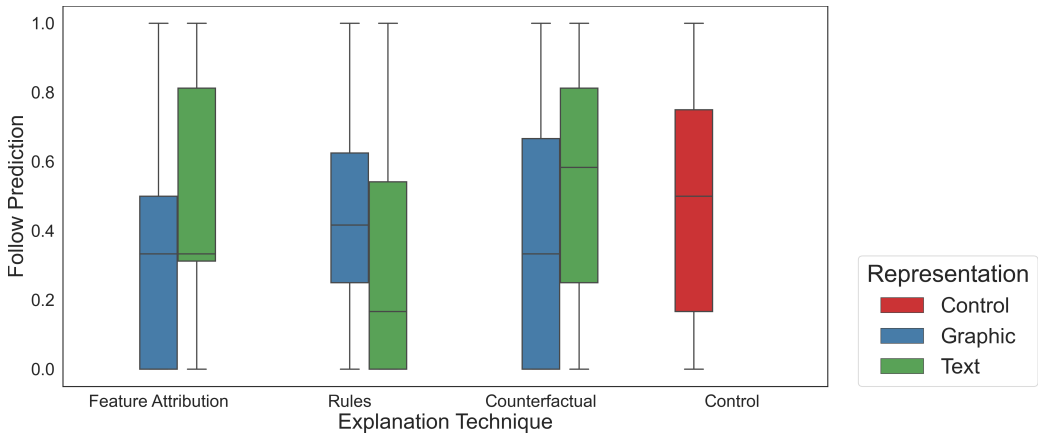


Figure 6.9 – Proportion of cases the participants changed their initial prediction to follow the AI's prediction. Results are shown for the Obesity dataset on the combination of explanation technique and representation.

	Recidivism		**Obesity**	
	Time	Survey Sat.	Time	Survey Sat.
Explanation Technique	2.49^{-}	0.39	0.78	1.60
Representation	1.04	0.00	0.88	0.02
Age	1.76	1.08	8.97^{**}	0.09
Education	1.31	1.75	2.07^{-}	1.73
Gender	2.14	0.05	0.03	1.52
Technique:Representation	0.20	1.60	0.38	1.49

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Table 6.5 – F value of the ANOVA table with additional measurements. The 'Survey Sat.' column is the self-reported satisfaction measured with a post-questionnaire. The column 'Time' represents the time required to complete the task round. 'Technique:Representation' is the interaction between the explanation technique and the visual representation.

observe that the time to indicate the most important features (completion time) for the obesity dataset is influenced by the age of the participant ($p<0.01$).

6.3.4 Perception vs. Behavior

We compare the self-reported and behavioral measurements of our experiment. We thus report the Pearson correlation between perceived trust (resp. understanding) and behavioral trust (understanding). We compare the two perception metrics with the behavioral measures defined in Section 5.2.2. We observed a correlation value of 0.43 and 0.49 between the perceived

trust in AI when facing an explanation **(Δ Confidence)** and the proportion of users following the AI's prediction **(Follow Pred.)** for the recidivism and obesity datasets respectively. This result suggests a moderate positive correlation between these two measurements. Conversely, our results indicate no correlation between the perceived users' understanding **(Immediate Und.)** or **(Survey Und.)** and their actual comprehension of the model as measured by the precision and recall scores.

6.3.5 Open Questions

We performed a qualitative analysis of participants' responses to two open-ended questions after our experiment: (1) "According to the scenarios you have seen and the corresponding explanation, how does the artificial intelligence tool predict?", and (2) "What was good in the explanation? What was bad in the explanation?". Our analysis revealed distinct themes in users' feedback, which were influenced by the explanation style and representation they encountered during the experiment.

For users exposed to counterfactual explanations, their feedback often centered around the scoring or assignment of scores to different categories. They frequently emphasized factors such as physical activity, calorie monitoring, and family history. Additionally, users drew comparisons between individual data and statistical information, as exemplified by comments like, "based on the individual data compared to statistical information of thousands of other individuals" (P68). While counterfactual explanations with graphical representation generally provided clarity, occasional misunderstandings arose. Users commented, "Most of the information was good in the explanation. The bad thing is that previous offenses were not explained in detail." (P147) and "The explanation was good because it listed facts that determine the individual's state of health, but was bad in terms of explaining how something like age can contribute to an individual's state of health."(P31).

In contrast, participants exposed to rule-based explanations tended to identify key risk factors and correlations. They also recognized the influence of lifestyle choices and standard criteria, with comments such as, "It focuses on main risk factors and their correlation between with one another" (P41). These users valued the clarity of rule-based and graphical explanations but expressed a desire for deeper insights. For example, they remarked, "The explanation was all clear, maybe it could use some more information to be more precise." (P06) and "Good at telling us what it was doing but maybe give more information on which specific characteristics had the most weighting" (P222).

Linear explanations often encouraged users to reflect on database reliance and comparisons to previous cases, with family history, food consumption, and physical activity emerging as recurring themes (*e.g.,* "The tool uses the frequency of certain behaviors and circumstances to predict someone's BMI category using percentages" (P101)). In the case of linear explanations with graphical representation, users found them practical but sought clarification on variable selection. They expressed, "The explanation about the AI and how it would review each of the four subject people was good but I didn't understand why the AI kept selecting different criteria to assess each individual." (P81) and "I think the entire explanation is good, but maybe they should give you more information on why some factors have more or less impact on the results" (P199).

Overall, textual representations were well-praised for their clarity, though some users occasionally pointed out biases. For instance, users stated, "The explanation showed me clearly how the AI would work. No doubts about that!" (P22), while also noting, "I think the tool might be a bit biased and too generalized" (P24). Similar trends were observed for the second dataset (COMPAS), with participants appreciating clarity and transparency in explanations. However, some expressed concerns about inherent biases in the AI's decision-making process. One user observed, "The good thing was that the tool scored well in most cases. The bad thing was that it put too much emphasis on previous arrests" (P194).

Participants in the control group found their interaction with the AI system useful but underscored the importance of understanding the prediction criteria. Their remarks included, "The tool predicts randomly" (P126) and "The good is that it was easy to read. The bad was that it did not contain specific information about the tool" (P130). These findings underscore the significance of clear, comprehensive, and unbiased explanations in facilitating effective human-AI interactions across diverse scenarios. We elaborate on these possible biases in the discussion section.

6.4 Discussion

In the subsequent sections, we will draw insights from our results. We studied the effects of three explanation techniques (feature attribution, rule-based, and counterfactual) and two representations (graphical, textual) on users' trust and understanding of AI models trained on data from two domains (recidivism and obesity). Similarly to prior works [7, 28, 156], our results suggest that explanations generally help users to (a) better identify which factors led to an AI's prediction, and (b) increase their perceived comprehension. Our findings also reveal

that the presentation of explanations plays a crucial role in shaping users' trust in the results. Specifically, graphical representations for explanations tend to be more effective in eliciting user acceptance than textual representations. We next discuss these two aspects in more detail.

6.4.1 Impact of Explanation Technique

We assessed the effects of three explanation techniques on participants' trust and understanding of an AI model (RQ1). First, our findings are in line with existing work and support our first hypothesis (H1), namely that explanations increase both (a) the users' comprehension of the AI model and, (b) the trust in the model's predictions. This is supported by the results in Table 6.3 that show a significant impact based on the chosen explanation technique. Furthermore, our study also confirms our second hypothesis that rule-based explanations are the most effective way to explain the inner workings of an AI system. This stands in line with existing results [5, 124]. We suspect that users within the group receiving rule-based explanations achieve better comprehension for two reasons: (a) the rules' alignment with common educational reasoning principles, and (b) the clarity in terms of when these rules are applicable, i.e., their simplicity. This is supported by the results for both self-reported understanding (Fig. 6.5) and precision (Fig. 6.6). Interestingly, we observe that the effects of the presence of explanations in AI-assisted tasks are more pronounced for the obesity dataset than for the recidivism dataset. We hypothesize that this is the result of (a) the number of features in the datasets (8 for the recidivism case and 15 for obesity), and (b) participants' prior knowledge of the field. Although participants are unlikely to have firsthand experience with prisoners, they are more likely to harbor preconceptions about the causes of obesity.

On the other hand, our study revealed a relatively low precision and self-reported understanding of the counterfactual explanations. These results were similar to the precision and self-reported understanding observed for the control group. However, it is important to note that we measure the users' precision and recall based on the ground truth of the explanation. As such, these results suggest that users are able to accurately identify the features mentioned in the explanation. Nevertheless, users do not possess information about whether any additional features play a role in the given classification. Consequently, the low precision results suggest that they tend to indicate additional features based on their preconceptions, which are not marked as important by the counterfactual explanation. This explains the relatively low precision observed with these explanations and underscores that counterfactual explanations may be perceived as less complete than feature attribution or rule-based. This outcome stands in stark contrast to the high scores obtained for both behavioral trust (as shown in Fig. 6.9)

and recall (as illustrated in Fig. 6.7). We attribute these findings to the nature of counterfactual explanations, which have been shown by social sciences to align with humans' ability to comprehend and explain complex events [102, 151]. Our findings align with this concept, as our participants tended to follow the AI model's predictions, as evidenced by the strong behavioral trust. Additionally, they successfully identified the features mentioned in the explanation, as indicated by the recall scores in the case of counterfactual explanations.

Our qualitative analysis in Section 6.3.5 regarding the advantages and drawbacks of explanations, underscores that users found explanations to be practical and were able to discern the influential factors in the prediction task. Nonetheless and similarly to existing results obtained from open discussions [150], some participants expressed concerns about the insufficiency of explanations. Indeed, these explanations do not clarify why specific attributes held importance for the model and their correlation with the assigned score. This qualitative study emphasizes that the current form of explanations is functional but might lack comprehensive depth. Participants also pointed out potential biases within the model that could stem from the model focusing on factors like family history or physical activity. This divergence reveals differences between users' expectations of the model's predictive outcomes and the manner in which these predictions are explained.

Interestingly, we initially anticipated that participants in the control group, who received no explanations, would report a lower perceived understanding of the model. Surprisingly, when we asked them about the benefits and drawbacks of the explanation, we found that some users did not seem to notice the absence of explanations. This observation brings to light two important phenomena: the concept of placebo explanations, which has been previously discussed in the literature [40], and the limited experience of laypersons with AI models. The notion of placebo explanations suggests that users might have attributed their understanding of the model based on the mere presence of explanations, regardless of their actual content. Furthermore, it is possible that these participants did not naturally consider the existence of explanations that have not yet been adapted for explaining ML models.

6.4.2 Impact of Representation

The influence of representation on users' perception has been well-established [10, 28], and our results corroborate this phenomenon. We found that the graphical representation induces a higher perceived trust. Self-reported trust levels were higher for the graphical representation compared to the text representation (Fig. 6.8). We suspect that these results stem from a cognitive bias related to the apparent complexity of a graphical presentation. This complexity

may give the impression of a greater underlying effort, thereby increasing users' trust in the system. It is worth emphasizing that our results do not intend to discourage the use of visual representations. Rather, they underscore the need for improved representation techniques. This is vital to highlight since our experiment studied only one possible visual representation for rule-based and counterfactual explanations.

Nevertheless, we represent the different explanation techniques based on a common representation. In this context, our findings corroborate our hypothesis 4, which states that textual representation appears to facilitate users' understanding of rule-based methods (Fig. 6.7). Similarly, we observe that users' trust in counterfactual explanations rises with textual representation (Fig. 6.9). Surprisingly, we notice an interesting pattern: with rule-based explanations, a textual representation results in higher comprehension but lower trust, while the reverse is observed with a graphical representation. This phenomenon appears to operate inversely with counterfactual explanations. That is, graphical representation leads to lower confidence but higher understanding, as opposed to the textual representation. As such, we find that users tend to trust a model's explanation more when their comprehension of the model is lower. As counterintuitive as this is, this observation stands in line with prior work. For instance, Van der Waa et al. [150] showed that for rule-based explanations, users displayed higher comprehension but lower trust. This contrasts with the results they obtained with example-based explanations, where users showed lower comprehension but higher confidence.

In line with existing work [133], our study evaluated the task completion time. Our results suggest a positive correlation between participants' age and completion time (as shown in Table 6.5). As such, older participants generally take longer to complete the task but achieve a higher recall score. This suggests that younger participants may tend to tackle the task more impulsively, resulting in quicker completion times but lower recall scores.

6.4.3 Limitations & Future Work

We identified several limitations related to the studied application domain and our participant sample. We intentionally presented participants with decision scenarios typically faced by domain experts. This increased reliance of participants on our explanations; this reliance is presumably less pronounced when domain experts use AI recommendation systems. Therefore, these results are not directly transferable to domain experts or computer scientists [33, 103, 125] as they may react differently and prefer different kinds of explanations and visual representations. We focused instead on crowd-sourcing participants to guarantee diversity and reach

a general audience. Further studies could seek to evaluate the effect of different explanation techniques and representations on different user groups.

Prior research has employed questionnaires to assess how explanation techniques impact users' comprehension [150] and how different explanation representations can influence users' trust [153]. However, the results from our three post-questionnaires (trust, understanding, and satisfaction) did not yield any differences across various explanation techniques and representations. This outcome could be due to the fact that users only engaged with the model a limited number of times and encountered instances that were classified differently. It is conceivable that this limited interaction might have contributed to the absence of statistical significance in our findings, as previously suggested by Van der Waa et al. [150]. To gain a more comprehensive perspective on the model's performance, future evaluation could consider either a larger set of instances or a focus on instances with similar classification outcomes.

We evaluated participant understanding through the identification of the most important features in the decision process. Other validation tasks could provide additional insights into participant understanding. For instance, tasks involving using the explanation to reproduce the AI's model behavior on different examples or answering what-if scenarios [13]. Such tasks would have the potential to measure a deeper understanding of the model. However, this would come at the cost of extra participant effort. In our specific experimental context, we observed limited improvements in users' comprehension with counterfactual explanations. We anticipate that what-if scenarios could be particularly beneficial to enhance understanding in the case of counterfactual explanations because these tasks align more closely with the objectives of counterfactual explanations. Finally, our study was conducted on AI models trained on tabular data. While the studied explanation techniques also apply to other data types such as text and images, the explanations on these data types may resort to visual representations not covered in our study.

6.5 Conclusion

The majority of XAI research has focused on the technical aspects of creating accurate explanation methods. In contrast, this study explores the human aspects of presenting explanations for AI agents to users. Specifically, we conducted a user study that aimed to examine the impact of explanation techniques and visual representations on users' trust and comprehension when faced with AI-based recommendations. Our study covered three types of explanations, namely feature-attribution, rule-based, and counterfactual, presented either graphically or as

textual statements. We evaluated these explanation strategies in two domains, namely the prediction of recidivism and the prediction of risk of obesity. Our results indicate that rule-based explanations with textual representation are most effective in terms of precision and self-reported understanding. We also observed a difference in the impact of explanations on users' trust across different domains. Counterfactual explanations presented as text elicited higher levels of trust, while the opposite was observed for graphical feature attribution and textual rule-based explanations. Importantly, our results exhibit some variations across the evaluated domains. This underscores the potential and necessity for future investigations to consider user characteristics, data types, and domains' influence on results.

CONCLUSION AND PERSPECTIVES

This final chapter marks the end of my Ph.D. journey, a moment both long-awaited and filled with apprehension. Embarking on a three-year-long project dedicated to a single research project might initially appear as an eternity for a research novice. However, in retrospect, it becomes evident that the search for more transparency in machine learning is far from reaching its end. In a time when explanations and transparency in machine learning are significant concerns across various fields, numerous techniques have arisen. Yet fundamental questions persist: under which circumstances should each explainability method be employed, and how do they impact the final users?

Therefore, In this concluding chapter, I begin by extracting from my experiences gained during this journey. I synthesize my contributions and glean common insights from the previous chapters of this Ph.D. thesis. Following this, I delve into the valuable lessons learned during this academic experience. From there, I explore some open challenges that require further investigation, while identifying exciting opportunities for future research. Finally, I conclude this thesis.

Thesis Objectives and Research Journey

With the growing demand for transparency and explainability in the field of AI, my primary goal throughout this thesis has been to explore the most effective methods to make AI agents fully understandable. This first objective led me to investigate the inner workings of explanation techniques and, more specifically, how the data used to generate these explanations influences their quality.

The journey began with a focus on enhancing explanation techniques, which resulted in my first publication and Chapter 2. In this research endeavor, I analyzed the impact of discretization techniques on the quality of rule-based explanations. Additionally, I studied the use of pertinent

negatives in rule-based explanations. This study reveals which words, when added, have the most substantial impact on the model's prediction probability.

As my doctoral journey progressed, it became evident that new methods were being introduced without a comprehensive understanding of their optimal application scenarios. Therefore, I switched my focus toward the question of when these explanation approaches should be applied. Choosing appropriate explanations became a pivotal aspect of my research. This inquiry led to the development of APE, a framework designed to discern whether a linear explanation is suitable for approximating a classification decision boundary for a given instance.

As I delved into APE, it guided me toward investigating the nearest decision boundary. This exploration naturally led me into the domain of producing counterfactual explanations. Thus, I developed Growing Fields as well as Growing Net and Growing Language, two novel counterfactual explanation techniques tailored for textual data. This phase of my research was motivated by the increasing quality and complexity of machine learning models in text-related tasks. It led me to reflect on whether adding complexity to methods intended to provide transparency was indeed a worthwhile initiative. This critical question is the foundation of the investigation presented in Chapter 4. In this chapter, I conducted a comparative study between transparent and black-box counterfactual explanation methods and questioned the need for intricate approaches.

During this academic journey, I also considered the end users of explanation techniques. While my initial focus had been on improving the quality of explanations, I recognized that the true impact of any research on eXplainable AI depends, in the end, on whether these improvements are beneficial to the ultimate users. Thus, I collaborated with Dr. Niels van Berkel on the construction of the user study detailed in Chapter 6. This study aimed to explore the impact of three distinct explanation techniques and two different representations on users' trust and understanding. The findings of this study have given us valuable insights into the various effects of different explanation techniques and their representations. This includes a study of the most effective representation for each explanation technique (e.g., textual for example and rule-based, graphical for feature attribution). Also, our study confirms the improved understanding observed when users were presented with rule-based explanations.

Notably, while I was conducting my research at Aalborg University, the lack of user studies measuring the impact of explanation techniques became evident. Conducting such a study presented its own set of challenges. One of those challenges is the irreversible nature of user interactions, which means that mistakes cannot be rectified after deployment. As a result, careful planning and execution were essential.

Key Insights and Lessons

Someone told me before the beginning of my Ph.D. that the true essence of the journey lies not merely in the manuscript itself but in the invaluable lessons acquired by the Ph.D. student. Thus, I conclude this manuscript by sharing some lessons I have learned. Firstly, I will explore the "Explanation El Dorado" and the quest for a one-size-fits-all solution to transparency in machine learning. Second, I will explore the profound impact of explanations on users, particularly the demand for explanations and the evolving trajectory that explanations could take.

The Allure of the "Explanation El Dorado"

The phenomenon I refer to as the "Explanation El Dorado" represents the collective pursuit within the research community to discover an elusive and all-encompassing solution for explainability in machine learning. This quest has become a veritable treasure hunting and focuses on finding a method that can universally render opaque machine learning models transparent to end users. In this era, explanations have emerged as the best way to bridge the gap between humans and machines, offering the promise of understanding the complex inner workings of algorithms. Machine learning models, once opaque and inscrutable, are now expected to elucidate their decision-making processes. Thus, researchers have tirelessly searched for a unique and innovative approach that could uncover the intricate mechanisms of the ML models.

This quest, however, is not without its complexities and challenges. The vast landscape of explanation methods prompts questions about the most suitable approach for specific applications. This issue becomes apparent, as researchers have aimed to develop explanations that are faithful but have yet underexplored the aspect of combining different paradigms for instance. Furthermore, generating explanations necessitates careful consideration of the trade-offs between accuracy and interpretability. In this regard, we exemplified in Chapter 4, that more and more complex methods are developed to explain black box models while simpler methods may be sufficient in certain conditions. As the "Explanation El Dorado" continues to attract, researchers have searched for a one-size-fits-all explanation solution, although such a method may be impossible to catch. It has become increasingly evident that explanations must be tailored to the specifics of the data and the intended users. For instance, in Chapter 3, we demonstrated the necessity of adapting the shape of explanations to the decision boundary for reliable results. In a similar vibe, in Chapter 6, we revealed how the form and representation of explanations can influence the interactions between end users and AI systems.

Upon reflection, it is clear that the quest for explanations holds the potential to revolutionize how we trust and interact with machine learning systems. The allure of the "Explanation El Dorado" continues to guide our endeavors, urging us to push the boundaries of what we can achieve. It underscores the belief that transparency is not a finite destination but an ongoing journey, one that leads us toward a brighter future of collaboration between humans and machines. To attain this, it is imperative to acknowledge the constraints that presently affect explanation techniques, including data types, model architectures, and the diverse requirements of end users.

Users Request for Explanations

My personal experience with explanations as well as the feedback I received through user studies have shown a growing trend: users are no longer satisfied with explanations that merely explain the surface. Indeed, current explanation methods do not provide reasons behind why a particular feature is significant for a model in generating a prediction. Perhaps, it is time to shift the paradigm of explanations towards elucidating the rationale behind the quality of a prediction. For instance, while counterfactual explanations are effective in revealing what needs to be altered to modify a prediction, there is a growing need to understand the underlying reasons behind specific changes. This includes modifying attributes like an individual's gender, and examining the resulting impact on the model's predictions, such as the likelihood of releasing a prisoner. Is this phenomenon a consequence of the training data, the inherent biases within the model, or other intricate factors that drive this outcome?

These questions drive us beyond the domain of traditional explanation methods into a deeper exploration of causality and the mechanisms that influence the decisions made by complex machine learning models [101]. As we delve deeper into the transparency mechanisms, we must not only unveil the "what" of predictions but also the "why", entering an era where interpretability encompasses not just transparency but also a profound understanding of how those AI systems ended up the way they are. By doing so, we equip ourselves with the knowledge required to handle the ethical, societal, and practical implications of machine learning in an increasingly interconnected world. This transformation represents a pivotal shift, where the quest for insight becomes as vital as the quest for prediction accuracy, forging a path towards responsible and trustworthy AI systems.

Open Challenges and Opportunities for Future Research

The contributions of this thesis lead toward several directions for future research. Beyond the perspectives presented in the concluding remarks of each individual chapter, this thesis incorporates upcoming research efforts focused on the interpretability approaches and users' perspectives. While these opportunities may not encompass every worthwhile research avenue, they contain the pivotal paths we regard as essential for advancing the domain of explainability in AI.

Explanations Adapted to Data

Throughout this thesis, I have identified two research directions that hold promise for future exploration. Firstly, it is essential to discern the conditions under which a given explanation technique is adapted to its context. This builds upon the research trajectory initiated by APE (cf Chapter 3). Secondly, the research community must investigate the effectiveness of introducing additional layers of complexity to methods designed to explain the inner workings of already complex models. This involves a more in-depth investigation of the research introduced in Chapter 4.

Traditionally, the pursuit of better explanation techniques has led to the search for a universal solution applicable to all situations. However, the complexity of objectives, user preferences, model architectures, and data types has rendered this approach insufficient. Consider, for example, the diversity of objectives users may have: some may prioritize the maximum fidelity to the underlying model, while others prefer simplicity or are just curious. These objectives are as varied as the individuals themselves. Therefore, a more nuanced approach involves a systematic examination and comparison of these methods, taking into account their impact on different aspects of the explanation process. As demonstrated in Chapter 3, a proper characterization of the decision boundary can guide the selection of adapted explanation methods. This not only provides insights into the suitability of linear surrogates for approximating specific decision boundaries but also serves as a stepping stone for similar investigations with various explanation techniques. Different contexts may require tailored explanations, as exemplified by the contrast between explaining the decisions of a model classifying the toxicity of a text on a social network versus a model predicting the date of a hurricane. While the former may necessitate a simpler, more intuitive explanation, the latter may require a comprehensive and technically precise elucidation of the model's reasoning. By exploring these nuances, we can uncover the true adaptability of explanation methods.

Another promising research direction is the quantification of the benefits derived from employing complex methods to explain the inner workings of black-box models. Although there is a common belief that increasing complexity of the ML models naturally results in better accuracy, Rudin [129] advocates for a rigorous investigation of this hypothesis. Therefore, the extension of this hypothesis to the domain of the trade-off between the complexity of explanations and their accuracy must be thoroughly examined. Consider, for instance, the use of intricate neural network architecture to explain a random forest model with few decision trees. Does the introduction of such sophisticated mechanisms truly enhance the clarity and utility of the explanations, or does it introduce an unnecessary layer of complexity that hinders comprehension?

To address this challenge, future research should aim to develop comprehensive frameworks for evaluating the impact of complexity on the effectiveness of explanation methods. In Chapter 4, our focus was on traditional metrics, however, there arises a necessity to shift towards user-centric assessments, considering factors like cognitive load, user satisfaction, and decision-making accuracy. By rigorously quantifying the advantages and disadvantages of complexity, we can establish a foundation for informed design choices in the development of explanation techniques for black-box models.

Explanations Tailored to Users

It has been largely accepted that explanations should be tailored to factors such as the domain, the AI group [77], and users role [125]. However, I propose considering broader aspects of the users such as their trust in AI, and their purpose when employing the AI systems [26]. While an expert may seek an explanation to understand why a model has failed, a company deploying a system may prioritize providing users with explanations for why the system makes certain predictions. This adaptability can have a profound impact on user interactions.

Throughout this thesis, we have demonstrated that surrogate explanations approximate complex models by training simpler models over interpretable spaces. Among these simpler models, we identified three kinds of surrogate methods: (a) feature-attribution, (b) example-based, and (c) rule-based explanations. Each surrogate approximates the complex model differently, and we have shown in Chapter 6 that the choice of the surrogate impacts how users interpret the explanation. Surprisingly, despite the growing interest in explainability, no prior work has compared the impact of these surrogates on specific user roles (e.g., domain expert, developer). Due to a lack of surrogate explanations comparison, XAI users are presently unable to indicate why they might choose one type of surrogate rather than another. However, the choice of the

surrogate and its representation can significantly affect users (e.g., trust, understanding) [10, 150]. We hence argue that preferring one type of surrogate over another should be driven by criteria and situations rather than for practical reasons.

As a future research direction, I propose to assess the impact of surrogate techniques across different user roles. Building upon our work [33], I outline various user roles to guide researchers and practitioners in investigating the impact of selecting a surrogate and representation depending on user roles.

Most existing research has focused on three types of roles [65, 125]: (a) developers that create or assess AI systems; (b) domain experts, persons with knowledge or authority in a particular area; and (c) lay users, individuals to whom the AI decision is applied (e.g., bank client). Yet, we argue that users and usage scenarios are more complex than those three well-defined categories. Instead, users of AI systems are multi-dimensional (e.g., roles, goals, trust in AI), and various scenarios affect the suitability of different explanation methods (e.g., data types, explanation representation). We thus propose six additional aspects to consider when selecting explanations tailored to users:

— **Motivation for Explanation:** Understanding why a user seeks an explanation, whether out of curiosity, to improve performance or to build trust in the system, is a key criterion for selecting the appropriate explanation model.

— **Trust in AI Systems:** Users' trust in AI systems can vary widely, as some programmers may place excessive trust in the systems they code while others may not have blind faith in it, influencing their reliance on explanations.

— **Use Case:** The context in which users interact with AI systems or explanation methods, whether in a professional setting, educational environment, or everyday life, plays a significant role in choosing the right explanation model.

— **Prior Experience:** Users' prior experience with the domain or explanation techniques can impact their interaction with explanations, as demonstrated in [87], where participants with mistrust in healthcare tended to trust AI systems more compared to doctors.

— **Data Types:** The challenges posed by representing various data types, such as sound or time series, is one of the reasons why few explanation methods exist for these data types [15]. As such, the data type influences the choice of surrogate explanation.

— **Visual Representation:** Selecting one explanation representation over another (e.g., graphical rather than textual) is crucial, as demonstrated in this thesis, as it impacts how users perceive AI systems.

Evaluating how each dimension of the user roles and usage scenarios impacts the effectiveness of surrogate explanations would allow associating surrogate methods tailored to users. Comparing the impact of different surrogate categories over the different aspects of users would benefit the ML sub-community of XAI by allowing them to manage and carefully select the appropriate proxy. As a future work, I envision investigating the impact of the three distinct explanation techniques in collaboration with computer scientists from different research laboratories, specialists either in HCI or ML, and domain experts in relevant domains (e.g., healthcare). These axes can be both continuous (e.g., trust in AI) and categorical (e.g., data types).

Envisioning the Future of Explainable AI

As we progress into the landscape of Explainable AI, it is crucial to pause and envision the path forward. The quest for better explanations should align with the diverse needs and motivations that drive users to seek them. In this higher-level exploration, we contemplate several key aspects that we believe should shape the future of explanations in XAI:

Application-Adapted Explanations

In the pursuit of more effective explanations in XAI, a statement emerges: the notion of a single, one-size-fits-all solution is insufficient. Instead, we must explore the concept of generating application-specific explanations. Consider, for example, that explaining a medical diagnosis should be approached quite differently from clarifying a legal decision. This realization highlights the importance of tailoring explanations to suit the specific context of their application. Likewise, the choice of the explanation architecture should be influenced by factors such as the specific instance being explained, the data types involved, and the characteristics of the black-box architectures. Furthermore, user objectives play a pivotal role in shaping the future of XAI. Explanations can serve a multitude of purposes, ranging from satisfying curiosity to fulfilling legal obligations, each with its own unique requirements (e.g., legal obligations vary [13]). To accomplish this, a thorough analysis of when and for whom a particular explanation technique is adapted is essential. This shift will enable us to transform the elusive dream of a one-size-fits-all solution into a practical reality.

Embracing Causality in Explanations

In the realm of explanations, current XAI techniques excel at identifying influential features in a model's decisions, but we must uncover the "why" behind these decisions. Embracing causality in explanations means going beyond surface-level information and providing insights into the causal relationships between input features and model outcomes. For instance, in the context of a hurricane prediction algorithm, understanding why an increase in temperature or a decrease in rainfall impacts hurricane risk is more crucial than just knowing that it does. This shift from "what" to "why" not only enhances our understanding of AI systems' decision-making processes but also significantly benefits the domain of application. It reduces the impression of being a passive spectator of Artificial Intelligence and paves the way for Augmented Intelligence.

Leveraging Language Models for Explanations

One other aspect that should be integrated into existing approaches for more transparency is the use of large language models. The rise of powerful language models opens up exciting possibilities for generating explanations [141]. However, as we have seen through the second part of this thesis, the explanation can be represented through diverse shapes. All these representations share one important limitation, they are static and non-interactive. Once a graph or a textual explanation is generated to depict the importance of each feature for the final prediction, users cannot seek further clarification or ask follow-up questions. As such, users may perceive themself as spectators not listening and outside of the process. To fill this gap, leveraging models like ChatGPT can lead to the creation of more natural, human-friendly explanations [155]. By integrating language models into XAI, we can bridge the chasm between technical model insights and human comprehension, making explanations more accessible, informative, and adaptable to users' needs.

Navigating the El Dorado of Explanations

The journey for better explanations in XAI can sometimes resemble the quest for the mythical El Dorado, a relentless pursuit that presents both formidable challenges and rich rewards. In this context, we recognize the importance of meaningful collaborations with businesses and organizations that rely on machine learning and XAI within their operations. By extending this invitation, we seek to engage with a diverse array of stakeholders, each with their unique goals and objectives when employing ML and XAI technologies. This collaborative effort is aimed at tailoring explanations to align with the requirements of users and organizations. These objectives

cover a wide range of considerations, from customizing visualization formats to ensuring the consistent provision of explanations. We understand that the needs of various entities can vary significantly, from corporations implementing predictive analytics to government agencies employing AI systems for critical decision-making processes. By actively involving these key players in the process, we can ensure that the journey towards effective explanations is both purposeful and fruitful. This collective effort toward a deeper understanding of the nuanced needs of the community will facilitate the development of more adaptable and user-centered explanation solutions.

Conclusion

This thesis has addressed the fundamental challenge of communication between humans and machine learning models. In an era where machine learning is ingrained into our everyday lives, this communication is not only beneficial in that it results in better models but also becomes a necessity for legal and moral reasons. To this end, we proposed methods to improve the fidelity of explanations for AI models. Then, we observe that there is a tendency to develop increasingly complex explanation methods. However, this complexity does not translate into increasing transparency and may in some conditions even reduce it. Therefore, as I conclude this thesis, it becomes evident that generating the best explanations cannot be divorced from measuring the impact of these explanations on humans. After all, humans are the ultimate users, and their experience should be the central concern of our research.

Conversely, solely assessing how humans perceive explanations is not sufficient. We recognize that current explanation techniques are still far from perfect, and there is much work to be done. Therefore, the combination of improving explanation quality and measuring its effectiveness with final users to guide further improvements represents the pivotal aspect towards responsible and effective machine learning interpretability. Within the complex landscape of human-AI interaction, we are reminded that our goal is not just to build transparent models but also to create trustworthy and impactful AI systems that align with ethical principles. This purpose will continue to guide researchers and practitioners toward a future where AI is not just a tool but a responsible and valuable collaborator in our lives.

Contribution

[23] Antoine Chaffin and Julien Delaunay, « "Honey, Tell Me What's Wrong", Explicabilité Globale des Modèles de TAL par la Génération Coopérative », *in*: *In Proc. le Traitement Automatique des Langues Naturelles*, ed. by Christophe Servan and Anne Vilnat, ATALA, 2023, pp. 105–122, URL: https://hal.science/hal-04130137.

[24] Antoine Chaffin and Julien Delaunay, *"Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Cooperative Generation*, 2023, URL: https://arxiv.org/abs/2310.18063.

[31] Julien Delaunay, Luis Galárraga, and Christine Largouët, « Improving Anchor-based Explanations », *in*: *Proc. CIKM*, ACM, 2020, DOI: https://doi.org/10.1145/3340531.3417461.

[32] Julien Delaunay, Luis Galárraga, and Christine Largouët, « When Should We Use Linear Explanations? », *in*: *Proc. CIKM*, ACM, 2022, DOI: https://doi.org/10.1145/3511808.3557489.

[33] Julien Delaunay et al., « Adaptation of AI Explanations to Users' Roles », *in*: *In Proc. CHI Workshop on Human-Centered Explainable AI*, 2023, pp. 1–7.

[50] Romaric Gaudel et al., « s-LIME: Reconciling Locality and Fidelity in Linear Explanations », *in*: *Proc. IDA*, vol. 13205, Lecture Notes in Computer Science, Springer, 2022, pp. 102–114, URL: https://doi.org/10.1007/978-3-031-01333-1%5C_9.

[155] Joel Wester et al., « On Moral Manifestations in Large Language Models », *in*: *In Proc. CHI Workshop on Moral Agents,* 2023, pp. 1–4.

BIBLIOGRAPHY

[1] Amina Adadi and Mohammed Berrada, « Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) », *in*: *IEEE Access* 6 (2018), pp. 52138–52160, DOI: 10.1109/ACCESS.2018.2870052.

[2] David Alvarez-Melis and Tommi S. Jaakkola, « On the Robustness of Interpretability Methods », *in*: *CoRR* abs/1806.08049 (2018), arXiv: 1806.08049, URL: http://arxiv.org/abs/1806.08049.

[3] Elvio Amparore, Alan Perotti, and Paolo Bajardi, « To Trust or not to Trust an Explanation: Using LEAF to Evaluate Local Linear XAI Methods », *in*: *PeerJ Computer Science* 7 (2021), ISSN: 2376-5992, DOI: 10.7717/peerj-cs.479, URL: http://dx.doi.org/10.7717/peerj-cs.479.

[4] Sule Anjomshoae et al., « Explainable Agents and Robots: Results from a Systematic Literature Review », *in*: *Proc. AAMAS*, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088, DOI: https://dl.acm.org/doi/10.5555/3306127.3331806, URL: http://dl.acm.org/citation.cfm?id=3331806.

[5] Siddhant Arora et al., « Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations », *in*: *Proc. AAAI*, AAAI Press, 2022, DOI: https://ojs.aaai.org/index.php/AAAI/article/view/20464.

[6] Alejandro Barredo Arrieta et al., « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *in*: *Inf. Fusion* 58 (2020), pp. 82–115, DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

[7] Maryam Ashoori and Justin D. Weisz, « In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes », *in*: *CoRR* (2019), DOI: http://arxiv.org/abs/1912.02675, arXiv: 1912.02675.

[8] Gilles Audemard et al., « Trading Complexity for Sparsity in Random Forest Explanations », *in*: *Proc. AAAI*, AAAI Press, 2022, pp. 5461–5469, URL: https://ojs.aaai.org/index.php/AAAI/article/view/20484.

[9] Pierre Baldi, « Autoencoders, Unsupervised Learning, and Deep Architectures », *in*: *Unsupervised and Transfer Learning - Workshop held at ICML*, ed. by Isabelle Guyon et al., vol. 27, JMLR Proceedings, 2012, URL: http://proceedings.mlr.press/v27/baldi12a.html.

[10] Niels van Berkel et al., « Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors », *in*: *Proc. CHI*, ACM, 2021, DOI: 10.1145/3411764.3445365.

[11] Adrien Bibal, Bruno Dumas, and Benoît Frénay, « User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning », *in*: *Workshop on Advances in Interpretable Machine Learning and AI*, 2019.

[12] Adrien Bibal et al., « Explaining t-SNE Embeddings Locally by Adapting LIME », *in*: *Proc. European Symposium on Artificial Neural Networks, ESANN*, 2020, URL: https://www.esann.org/sites/default/files/proceedings/2020/ES2020-105.pdf.

[13] Adrien Bibal et al., « Legal requirements on explainability in machine learning », *in*: *Artificial Intelligence and Law* 29.*2* (2021), pp. 149–169, ISSN: 1572-8382, DOI: https://doi.org/10.1007/s10506-020-09270-4.

[14] Reuben Binns et al., « 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions », *in*: *Proc. CHI*, NY, USA: Association for Computing Machinery, 2018, pp. 1–14, ISBN: 9781450356206, DOI: https://doi.org/10.1145/3173574.3173951.

[15] Francesco Bodria et al., « Benchmarking and Survey of Explanation Methods for Black Box Models », *in*: *CoRR* (2021), DOI: https://arxiv.org/abs/2102.13076.

[16] S Boon and J Holmes, « The Dynamics of Interpersonal Trust: Resolving Uncertainity in the Face of Risk », *in*: *Cooperation and Prosocial Behaviour*, ed. by R Hinde and J Gorebel, Cambridge: Cambridge University Press, 1991, pp. 190–211.

[17] Leo Breiman et al., *Classification and Regression Trees*, Wadsworth, 1984, ISBN: 0-534-98053-8.

[18] Tim Brennan, William Dieterich, and Beate Ehret, « Evaluating the predictive validity of the compas risk and needs assessment system », en, *in*: *Crim. Justice Behav.* 36.*1* (2009), pp. 21–40, DOI: http://dx.doi.org/10.1177/0093854808326545.

[19] Tom B. Brown et al., « Language Models are Few-Shot Learners », *in*: *Proc. NeurIPS*, ed. by Hugo Larochelle et al., 2020, URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[20] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos, « To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making », *in*: *Proc. ACM Hum. Comput. Interact.* 5.*CSCW1* (2021), 188:1–188:21, DOI: 10.1145/3449287, URL: https://doi.org/10.1145/3449287.

[21] Ruth M. J. Byrne, « Counterfactual Thought. », *in*: *Annual review of psychology* 67 (2016), pp. 135–57, URL: https://api.semanticscholar.org/CorpusID:31978325.

[22] Béatrice Cahour and Jean-François Forzy, « Does projection into use improve trust and exploration? An example with a cruise control system », *in*: *Safety Science* 47.*9* (2009), Research in Ergonomic Psychology in the Transportation Field in France, pp. 1260–1270, ISSN: 0925-7535, DOI: https://www.sciencedirect.com/science/article/pii/S0925753509000587.

[23] Antoine Chaffin and Julien Delaunay, « "Honey, Tell Me What's Wrong", Explicabilité Globale des Modèles de TAL par la Génération Coopérative », *in*: *In Proc. le Traitement Automatique des Langues Naturelles*, ed. by Christophe Servan and Anne Vilnat, ATALA, 2023, pp. 105–122, URL: https://hal.science/hal-04130137.

[24] Antoine Chaffin and Julien Delaunay, *"Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Cooperative Generation*, 2023, URL: https://arxiv.org/abs/2310.18063.

[25] Chaofan Chen et al., « This Looks Like That: Deep Learning for Interpretable Image Recognition », *in*: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 8928–8939, URL: https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

[26] Valerie Chen et al., « Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations », *in*: *CoRR* (2023), DOI: 10.48550/arXiv.2301.07255, URL: https://doi.org/10.48550/arXiv.2301.07255.

[27] Ziheng Chen et al., « ReLAX: Reinforcement Learning Agent Explainer for Arbitrary Predictive Models », *in*: *Proc. CIKM*, ed. by Mohammad Al Hasan and Li Xiong, ACM, 2022, DOI: 10.1145/3511808.3557429, URL: https://doi.org/10.1145/3511808.3557429.

[28] Hao Fei Cheng et al., « Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders », *in*: *Proc. CHI, Glasgow, Scotland, UK*, ACM, 2019, DOI: https://doi.org/10.1145/3290605.3300789.

[29] Michael Chromik et al., « I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI », *in*: *Proc. IUI*, ed. by Tracy Hammond et al., ACM, 2021, DOI: 10.1145/3397481.3450644, URL: https://doi.org/10.1145/3397481.3450644.

[30] Belur V. Dasarathy, « Nearest unlike neighbor (NUN): an aid to decision confidence estimation », *in*: *Optical Engineering* 34 (1995), pp. 2785–2792, URL: https://api.semanticscholar.org/CorpusID:122160470.

[31] Julien Delaunay, Luis Galárraga, and Christine Largouët, « Improving Anchor-based Explanations », *in*: *Proc. CIKM*, ACM, 2020, DOI: https://doi.org/10.1145/3340531.3417461.

[32] Julien Delaunay, Luis Galárraga, and Christine Largouët, « When Should We Use Linear Explanations? », *in*: *Proc. CIKM*, ACM, 2022, DOI: https://doi.org/10.1145/3511808.3557489.

[33] Julien Delaunay et al., « Adaptation of AI Explanations to Users' Roles », *in*: *In Proc. CHI Workshop on Human-Centered Explainable AI*, 2023, pp. 1–7.

[34] Jacob Devlin et al., « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding », *in*: *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, ed. by Jill Burstein, Christy Doran, and Thamar Solorio, Association for Computational Linguistics, 2019, DOI: 10.18653/v1/n19-1423, URL: https://doi.org/10.18653/v1/n19-1423.

[35] Amit Dhurandhar et al., « Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives », *in*: *Proc. NIPS*, 2018.

[36] Jonathan Dodge et al., « Explaining models: an empirical study of how explanations impact fairness judgment », *in*: *Proc. IUI*, ACM, 2019, DOI: https://doi.org/10.1145/3301275.3302310.

[37] Finale Doshi-Velez and Been Kim, « Considerations for Evaluation and Generalization in Interpretable Machine Learning », *in*: *Explainable and Interpretable Models in Computer Vision and Machine Learning*, ed. by Hugo Jair Escalante et al., Springer International Publishing, 2018, DOI: 10.1007/978-3-319-98131-4_1, URL: https://doi.org/10.1007/978-3-319-98131-4_1.

[38] Finale Doshi-Velez and Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, 2017, DOI: https://arxiv.org/abs/1702.08608.

[39] Mengnan Du, Ninghao Liu, and Xia Hu, « Techniques for interpretable machine learning », *in*: *Commun. ACM* 63.*1* (2020), pp. 68–77, DOI: https://doi.org/10.1145/3359786.

[40] Malin Eiband et al., « The Impact of Placebic Explanations on Trust in Intelligent Systems », *in*: *Extended Abstracts CHI*, ed. by Regan L. Mandryk et al., ACM, 2019, DOI: 10.1145/3290607.3312787, URL: https://doi.org/10.1145/3290607.3312787.

[41] David A. Elizondo, « The linear separability problem: some testing methods », *in*: *IEEE Trans. Neural Networks* 17.*2* (2006), pp. 330–344, URL: https://doi.org/10.1109/TNN.2005.860871.

[42] *Estimation of obesity levels based on eating habits and physical condition*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5H31Z, 2019.

[43] European Commission, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016, URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[44] Usama M. Fayyad and Keki B. Irani, « Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning », *in*: *Proc. IJCAI*, 1993, pp. 1022–1029.

[45] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, Bradford Books, 1998.

[46] Carlos Fernandez, Foster J. Provost, and Xintian Han, « Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach », *in*: *CoRR* (2020), DOI: https://arxiv.org/abs/2001.07417.

[47] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim, « On cognitive preferences and the plausibility of rule-based models », *in*: *Machine Learning* 109.*4* (2020), pp. 853–898, ISSN: 1573-0565, URL: https://doi.org/10.1007/s10994-019-05856-5.

[48] Krzysztof Z. Gajos and Lena Mamykina, « Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning », *in*: *Proc. IUI*, ed. by Giulio Jacucci et al., ACM, 2022, pp. 794–806, DOI: 10.1145/3490099.3511138, URL: https://doi.org/10.1145/3490099.3511138.

[49] Damien Garreau and Ulrike von Luxburg, « Explaining the Explainer: A First Theoretical Analysis of LIME », *in*: *AISTATS*, 2020.

[50] Romaric Gaudel et al., « s-LIME: Reconciling Locality and Fidelity in Linear Explanations », *in*: *Proc. IDA*, vol. 13205, Lecture Notes in Computer Science, Springer, 2022, pp. 102–114, URL: https://doi.org/10.1007/978-3-031-01333-1%5C_9.

[51] Sourojit Ghosh and Aylin Caliskan, « ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages », *in*: *CoRR* abs/2305.10510 (2023), DOI: 10.48550/arXiv.2305.10510, URL: https://doi.org/10.48550/arXiv.2305.10510.

[52] Alicja Gosiewska and Przemyslaw Biecek, « iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models », *in*: *CoRR* (2019), URL: http://arxiv.org/abs/1903.11420.

[53] Ben Green and Yiling Chen, « Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments », *in*: *Proc. FAT**, ed. by danah boyd and Jamie H. Morgenstern, ACM, 2019, DOI: 10.1145/3287560.3287563, URL: https://doi.org/10.1145/3287560.3287563.

[54] Ben Green and Yiling Chen, « The Principles and Limits of Algorithm-in-the-Loop Decision Making », *in*: *Proc. ACM Hum. Comput. Interact.* 3.*CSCW* (2019), 50:1–50:24, DOI: 10.1145/3359152, URL: https://doi.org/10.1145/3359152.

[55] Riccardo Guidotti, « Counterfactual explanations and how to find them: literature review and benchmarking », *in*: *Data Mining and Knowledge Discovery* (2022), ISSN: 1573-756X, DOI: https://doi.org/10.1007/s10618-022-00831-6.

[56] Riccardo Guidotti, « Evaluating local explanation methods on ground truth », *in*: *Artif. Intell.* 291 (2021), p. 103428, DOI: 10.1016/j.artint.2020.103428, URL: https://doi.org/10.1016/j.artint.2020.103428.

[57] Riccardo Guidotti et al., « A Survey of Methods for Explaining Black Box Models », *in*: *ACM Comput. Surv.* 51.*5* (2019), 93:1–93:42, DOI: https://doi.org/10.1145/3236009.

[58] Riccardo Guidotti et al., « Black Box Explanation by Learning Image Exemplars in the Latent Feature Space », *in*: *Proc. ECML PKDD*, ed. by Ulf Brefeld et al., Lecture Notes in Computer Science, Springer, 2019, DOI: 10.1007/978-3-030-46150-8_12, URL: https://doi.org/10.1007/978-3-030-46150-8%5C_12.

[59] Riccardo Guidotti et al., « Local Rule-Based Explanations of Black Box Decision Systems », *in*: *CoRR* (2018), DOI: http://arxiv.org/abs/1805.10820.

[60] Suchin Gururangan et al., « Annotation Artifacts in Natural Language Inference Data », *in*: *Proc. NAACL-HLT*, ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent, Association for Computational Linguistics, 2018, DOI: 10.18653/v1/n18-2017, URL: https://doi.org/10.18653/v1/n18-2017.

[61] Victor Guyomard et al., « VCNet: A Self-explaining Model for Realistic Counterfactual Generation », *in*: *Proc. ECML PKDD*, ed. by Massih-Reza Amini et al., Lecture Notes in Computer Science, Springer, 2022, DOI: 10.1007/978-3-031-26387-3_27, URL: https://doi.org/10.1007/978-3-031-26387-3%5C_27.

[62] J F Hair et al., *Multivariate data analysis (Seventh edition Pearson new international)*, Pearson Education Limited, 2014.

[63] Zellig S. Harris, « Distributional Structure », *in*: *WORD* 10.*2-3* (1954), pp. 146–162, DOI: 10.1080/00437956.1954.11659520.

[64] Peter Hase and Mohit Bansal, « Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? », *in*: *Proc. ACL*, ed. by Dan Jurafsky et al., Association for Computational Linguistics, 2020, DOI: 10.18653/v1/2020.acl-main.491, URL: https://doi.org/10.18653/v1/2020.acl-main.491.

[65] Maryam Hashemi, « Who wants what and how: a Mapping Function for Explainable Artificial Intelligence », *in*: *CoRR* abs/2302.03180 (2023), DOI: 10.48550/arXiv.2302.03180.

[66] Robert R. Hoffman et al., « Metrics for Explainable AI: Challenges and Prospects », *in*: *CoRR* (2018), DOI: http://arxiv.org/abs/1812.04608.

[67] Matthew Honnibal and Ines Montani, « spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing », To appear, 2017.

[68] Alon Jacovi, « Trends in Explainable AI (XAI) Literature », *in*: *CoRR* (2023), DOI: 10.48550/arXiv.2301.05433, arXiv: 2301.05433, URL: https://doi.org/10.48550/arXiv.2301.05433.

[69] Alon Jacovi and Yoav Goldberg, « Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? », *in*: *Proc. ACL*, ed. by Dan Jurafsky et al., Association for Computational Linguistics, 2020, DOI: 10.18653/v1/2020.acl-main.386, URL: https://doi.org/10.18653/v1/2020.acl-main.386.

[70] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988.

[71] Sirkka L. Jarvenpaa, Kathleen Knoll, and Dorothy E. Leidner, « Is Anybody Out There? Antecedents of Trust in Global Virtual Teams », *in*: *J. Manag. Inf. Syst.* 14.*4* (1998), pp. 29–64, DOI: https://doi.org/10.1080/07421222.1998.11518185.

[72] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie, « Adversarial Counterfactual Visual Explanations », *in*: *CoRR* abs/2303.09962 (2023), DOI: 10.48550/arXiv.2303.09962, arXiv: 2303.09962, URL: https://doi.org/10.48550/arXiv.2303.09962.

[73] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury, « Foundations for an Empirically Determined Scale of Trust in Automated Systems », *in*: *International Journal of Cognitive Ergonomics* 4.*1* (2000), pp. 53–71, DOI: https://doi.org/10.1207/S15327566IJCE0401_04.

[74] Pride Kavumba et al., « Prompting for explanations improves Adversarial NLI. Is this true? Yes it is true because it weakens superficial cues », *in*: *Findings EACL*, ed. by Andreas Vlachos and Isabelle Augenstein, Association for Computational Linguistics, 2023, URL: https://aclanthology.org/2023.findings-eacl.162.

[75] Been Kim, « Interactive and interpretable machine learning models for human machine collaboration », MA thesis, Massachussetts Institute of Technology, 2015, URL: http://hdl.handle.net/1721.1/98680.

[76] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna, « Examples are not enough, learn to criticize! Criticism for Interpretability », *in*: *Proc. NeurIPS*, ed. by Daniel D. Lee et al., 2016, URL: https://proceedings.neurips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html.

[77] Taenyun Kim et al., « One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles », *in*: *Proc. CHI*, ed. by Albrecht Schmidt et al., ACM, 2023, DOI: 10.1145/3544548.3581340, URL: https://doi.org/10.1145/3544548.3581340.

[78] Mauritz Kop, « EU Artificial Intelligence Act: The European Approach to AI », *in*: (2021), URL: https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/.

[79] Todd Kulesza et al., « Too much, too little, or just right? Ways explanations impact end users' mental models », *in*: *IEEE Symposium on Visual Languages and Human Centric Computing*, IEEE Computer Society, 2013, DOI: https://doi.org/10.1109/VLHCC.2013.6645235.

[80] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec, « Interpretable Decision Sets: A Joint Framework for Description and Prediction », *in*: *Proc. SIGKDD*, ACM, 2016, DOI: https://doi.org/10.1145/2939672.2939874.

[81] Retno Larasati, Anna De Liddo, and Enrico Motta, « The Effect of Explanation Styles on User's Trust », *in*: *Proc. (IUI*, CEUR Workshop Proceedings, CEUR-WS.org, 2020, DOI: http://ceur-ws.org/Vol-2582/paper6.pdf.

[82] Thibault Laugel et al., « Comparison-Based Inverse Classification for Interpretability in Machine Learning », *in*: *Proc. IPMU*, Springer, 2018, DOI: https://doi.org/10.1007/978-3-319-91473-2_9.

[83] Thibault Laugel et al., « Defining Locality for Surrogates in Post-hoc Interpretablity », *in*: *CoRR* abs/1806.07498 (2018), arXiv: 1806.07498, URL: http://arxiv.org/abs/1806.07498.

[84] Thibault Laugel et al., « The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations », *in*: *Proc. IJCAI*, ed. by Sarit Kraus, ijcai, 2019, DOI: 10.24963/ijcai.2019/388, URL: https://doi.org/10.24963/ijcai.2019/388.

[85] Thibault Laugel et al., « Unjustified Classification Regions and Counterfactual Explanations in Machine Learning », *in*: *Proc. ECML PKDD*, ed. by Ulf Brefeld et al., Lecture Notes in Computer Science, Springer, 2019, DOI: 10.1007/978-3-030-46147-8_3, URL: https://doi.org/10.1007/978-3-030-46147-8%5C_3.

[86] Thai Le, Suhang Wang, and Dongwon Lee, « GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction », *in*: *Proc. KDD*, ed. by Rajesh Gupta et al., ACM, 2020, DOI: 10.1145/3394486.3403066, URL: https://doi.org/10.1145/3394486.3403066.

[87] Min Kyung Lee and Katherine Rich, « Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust », *in*: *Proc. CHI*, ACM, 2021, DOI: https://doi.org/10.1145/3411764.3445570.

[88] Chuanrong Li et al., « Linguistically-Informed Transformations (LIT): A Method for Automatically Generating Contrast Sets », *in*: *Proc. of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*, ed. by Afra Alishahi et al., Association for Computational Linguistics, 2020, DOI: 10.18653/v1/2020.blackboxnlp-1.12, URL: https://doi.org/10.18653/v1/2020.blackboxnlp-1.12.

[89] Dongfang Li et al., « Unifying Model Explainability and Robustness for Joint Text Classification and Rationale Extraction », *in*: *Proc. AAAI*, AAAI Press, 2022, DOI: 10.1609/aaai.v36i10.21342, URL: https://doi.org/10.1609/aaai.v36i10.21342.

[90] Ziqiang Li et al., « Interpreting the Latent Space of GANs via Measuring Decoupling », *in*: *IEEE Trans. Artif. Intell.* 2.*1* (2021), pp. 58–70, URL: https://doi.org/10.1109/TAI.2021.3071642.

[91] Zachary C. Lipton, « The mythos of model interpretability », *in*: *Commun. ACM* 61.*10* (2018), pp. 36–43, DOI: https://doi.org/10.1145/3233231.

[92] Yinhan Liu et al., « RoBERTa: A Robustly Optimized BERT Pretraining Approach », *in*: *CoRR* abs/1907.11692 (2019), arXiv: 1907.11692, URL: http://arxiv.org/abs/1907.11692.

[93] Gianluigi Lopardo, Frédéric Precioso, and Damien Garreau, « Understanding Post-hoc Explainers: The Case of Anchors », *in*: *CoRR* abs/2303.08806 (2023), DOI: 10.4855

0/arXiv.2303.08806, arXiv: 2303.08806, URL: https://doi.org/10.48550/arXiv.2303.08806.

[94] Scott M. Lundberg and Su-In Lee, « A Unified Approach to Interpreting Model Predictions », *in*: *Proc. NIPS*, 2017, DOI: https://dl.acm.org/doi/10.5555/3295222.3295230.

[95] Ronny Luss et al., « Leveraging Latent Features for Local Explanations », *in*: *Proc. KDD*, ACM, 2021, DOI: https://doi.org/10.1145/3447548.3467265.

[96] Nishtha Madaan et al., « Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text », *in*: *Proc. IAAI, The Symposium on Educational Advances in Artificial Intelligence, EAAI*, AAAI Press, 2021, URL: https://ojs.aaai.org/index.php/AAAI/article/view/17594.

[97] Maria Madsen and Shirley D Gregor, « Measuring Human-Computer Trust », *in*: *Proc. Computer Science, Psychology*, 2000.

[98] David Martens and Foster J. Provost, « Explaining Data-Driven Document Classifications », *in*: *MIS Q.* 38.*1* (2014), pp. 73–99, URL: http://misq.org/explaining-data-driven-document-classifications.html.

[99] Tom McCoy, Ellie Pavlick, and Tal Linzen, « Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference », *in*: *Proc. ACL*, ed. by Anna Korhonen, David R. Traum, and Lluis Marquez, Association for Computational Linguistics, 2019, DOI: 10.18653/v1/p19-1334, URL: https://doi.org/10.18653/v1/p19-1334.

[100] Tomás Mikolov et al., « Efficient Estimation of Word Representations in Vector Space », *in*: *Workshop Track Proceedings ICLR*, ed. by Yoshua Bengio and Yann LeCun, 2013, URL: http://arxiv.org/abs/1301.3781.

[101] Tim Miller, « Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI », *in*: *Proc. of the Conference on Fairness, Accountability, and Transparency, FAccT*, ACM, 2023, DOI: 10.1145/3593013.3594001, URL: https://doi.org/10.1145/3593013.3594001.

[102] Tim Miller, « Explanation in Artificial Intelligence: Insights from the Social Sciences », *in*: *Artif. Intell.* 267 (2019), pp. 1–38, DOI: https://doi.org/10.1016/j.artint.2018.07.007.

[103] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan, « A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems », *in*: *ACM Trans. Interact. Intell. Syst.* 11 (2021), 24:1–24:45, DOI: 10.1145/3387166, URL: https://doi.org/10.1145/3387166.

[104] Christoph Molnar, *Interprtable machine learning: A guide for making black box models explainable*, 2018.

[105] John X. Morris et al., « TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP », *in*: *Proc. EMNLP*, ed. by Qun Liu and David Schlangen, Association for Computational Linguistics, 2020, DOI: 10.18653/v1/2020.emnlp-demos.16, URL: https://doi.org/10.18653/v1/2020.emnlp-demos.16.

[106] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan, « Explaining machine learning classifiers through diverse counterfactual explanations », *in*: *Proc. FAT*, ed. by Mireille Hildebrandt et al., ACM, 2020, DOI: 10.1145/3351095.3372850, URL: https://doi.org/10.1145/3351095.3372850.

[107] Meike Nauta et al., « From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI », *in*: *CoRR* (2022), DOI: https://arxiv.org/abs/2201.08164.

[108] Arvind Neelakantan et al., « Text and Code Embeddings by Contrastive Pre-Training », *in*: *CoRR* abs/2201.10005 (2022), URL: https://arxiv.org/abs/2201.10005.

[109] Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker, « The Impact of Explanation Dialogues on Human-Computer Trust », *in*: *Proc. Human-Computer Interaction. Users and Contexts of Use*, ed. by Masaaki Kurosu, Springer, 2013, DOI: https://doi.org/10.1007/978-3-642-39265-8_7.

[110] Mahsan Nourani et al., « Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems », *in*: *Proc. IUI*, ed. by Tracy Hammond et al., ACM, 2021, DOI: 10.1145/3397481.3450639, URL: https://doi.org/10.1145/3397481.3450639.

[111] Lampridis O. et al., « Explaining short text classification with diverse synthetic exemplars and counter-exemplars », *in*: *Machine learning* (2022), DOI: 10.1007/s10994-022-06150-7.

[112] Jeroen Ooge and Katrien Verbert, « Explaining Artificial Intelligence with Tailored Interactive Visualisations », *in*: *Proc. IUI*, ACM, 2022, ISBN: 9781450391450, DOI: https://doi.org/10.1145/3490100.3516481.

[113] OpenAI, « GPT-4 Technical Report », *in*: *CoRR* abs/2303.08774 (2023), DOI: 10.48550/arXiv.2303.08774, URL: https://doi.org/10.48550/arXiv.2303.08774.

[114] Fabio Mendoza Palechor and Alexis de la Hoz Manotas, « Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico », *in*: *Data in Brief* 25 (2019), p. 104344, ISSN: 2352-3409, DOI: https://www.sciencedirect.com/science/article/pii/S2352340919306985.

[115] Judea Pearl, « Causal inference in statistics: An overview », *in*: *Statistics Surveys* 3.*none* (2009), pp. 96–146, URL: https://doi.org/10.1214/09-SS057.

[116] F. Pedregosa et al., « Scikit-learn: Machine Learning in Python », *in*: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[117] Vitali Petsiuk, Abir Das, and Kate Saenko, « RISE: Randomized Input Sampling for Explanation of Black-box Models », *in*: *Proc. British Machine Vision Conference BMVC*, BMVA Press, 2018, URL: http://bmvc2018.org/contents/papers/1064.pdf.

[118] Antonin Poche, Lucas Hervier, and Mohamed Chafik Bakkay, « Natural Example-Based Explainability: a Survey », *in*: 2023, URL: https://api.semanticscholar.org/CorpusID:261582519.

[119] Forough Poursabzi-Sangdeh et al., « Manipulating and Measuring Model Interpretability », *in*: *Proc. CHI*, ACM, 2021, DOI: https://doi.org/10.1145/3411764.3445315.

[120] Rafael Poyiadzi et al., « FACE: Feasible and Actionable Counterfactual Explanations », *in*: *Proc. AIES*, ACM, 2020, DOI: https://doi.org/10.1145/3375627.3375850.

[121] Nedeljko Radulovic, Albert Bifet, and Fabian M. Suchanek, « BELLA: Black box model Explanations by Local Linear Approximations », *in*: *CoRR* abs/2305.11311 (2023), DOI: 10.48550/arXiv.2305.11311, URL: https://doi.org/10.48550/arXiv.2305.11311.

[122] Marco Tulio Ribeiro, « Model-Agnostic Explanations and Evaluation of Machine Learning », MA thesis, University of Washington, 2018.

[123] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, « "Why Should I Trust You?": Explaining the Predictions of Any Classifier », *in*: *Proc. SIGKDD*, ACM, 2016, DOI: https://doi.org/10.1145/2939672.2939778.

[124] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, « Anchors: High-Precision Model-Agnostic Explanations », *in*: *Proc. AAAI*, AAAI Press, 2018, DOI: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

[125] Mireia Ribera and Àgata Lapedriza, « Can we do better explanations? A proposal of user-centered explainable AI », *in*: *Proc. IUI*, CEUR Workshop Proceedings, CEUR-WS.org, 2019, DOI: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

[126] Marcel Robeer, Floris Bex, and Ad Feelders, « Generating Realistic Natural Language Counterfactuals », *in*: *Findings EMNLP*, Association for Computational Linguistics, 2021, DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.306.

[127] Alexis Ross, Ana Marasovic, and Matthew E. Peters, « Explaining NLP Models via Minimal Contrastive Editing (MiCE) », *in*: *Findings ACL/IJCNLP*, ed. by Chengqing Zong et al., Association for Computational Linguistics, 2021, DOI: 10.18653/v1/2021.findings-acl.336, URL: https://doi.org/10.18653/v1/2021.findings-acl.336.

[128] Alexis Ross et al., « Tailor: Generating and Perturbing Text with Semantic Controls », *in*: *Proc. ACL*, ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, Association for Computational Linguistics, 2022, DOI: 10.18653/v1/2022.acl-long.228, URL: https://doi.org/10.18653/v1/2022.acl-long.228.

[129] Cynthia Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *in*: *Nature Machine Intelligence* 1.*5* (2019), pp. 206–215, ISSN: 2522-5839, URL: https://doi.org/10.1038/s42256-019-0048-x.

[130] Rosemarie S. Punla Candida C. Farro, « Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC », *in*: *International Multidisciplinary Research Journal* 4.*3* (2022), pp. 50–59.

[131] Gerard Salton and Christopher Buckley, « Term-weighting approaches in automatic text retrieval », *in*: *Information Processing & Management* 24.*5* (1988), pp. 513–523, ISSN: 0306-4573, DOI: https://doi.org/10.1016/0306-4573(88)90021-0,

URL: https://www.sciencedirect.com/science/article/pii/0306457388900210.

[132] Victor Sanh et al., « DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter », *in*: *ArXiv* abs/1910.01108 (2019).

[133] Philipp Schmidt and Felix Bießmann, « Quantifying Interpretability and Trust in Machine Learning Systems », *in*: *CoRR* (2019), DOI: http://arxiv.org/abs/1901.08558.

[134] Mattia Setzu et al., « Global Explanations with Local Scoring », *in*: *Proc. ECML PKDD*, Communications in Computer and Information Science, Springer, 2019, DOI: https://doi.org/10.1007/978-3-030-43823-4_14.

[135] Sharath M. Shankaranarayana and Davor Runje, « ALIME: Autoencoder Based Approach for Local Interpretability », *in*: *CoRR* abs/1909.02437 (2019), URL: http://arxiv.org/abs/1909.02437.

[136] Lloyd S Shapley, « A Value for n-Person Games », *in*: *Contributions to the Theory of Games II*, ed. by Harold W. Kuhn and Albert W. Tucker, Princeton: Princeton University Press, 1953, pp. 307–317.

[137] Yujun Shen et al., « Interpreting the Latent Space of GANs for Semantic Face Editing », *in*: *Proc. CVPR*, Computer Vision Foundation / IEEE, 2020, DOI: 10.1109/CVPR42600.2020.00926, URL: https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Shen%5C_Interpreting%5C_the%5C_Latent%5C_Space%5C_of%5C_GANs%5C_for%5C_Semantic%5C_Face%5C_Editing%5C_CVPR%5C_2020%5C_paper.html.

[138] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, « Learning Important Features Through Propagating Activation Differences », *in*: *Proc. ICML*, PMLR, 2017, DOI: https://proceedings.mlr.press/v70/shrikumar17a.html.

[139] Alban Siffer et al., « Are your data gathered? », *in*: *Proc. SIGKDD*, ACM, 2018, URL: https://doi.org/10.1145/3219819.3219994.

[140] Sameer Singh, Marco Túlio Ribeiro, and Carlos Guestrin, « Programs as Black-Box Explanations », *in*: *CoRR* (2016), DOI: http://arxiv.org/abs/1611.07579.

[141] Dylan Slack et al., « Explaining machine learning models with interactive natural language conversations using TalkToModel », *in*: *Nature Machine Intelligence* (2023), DOI: 10.1038/s42256-023-00692-8, URL: https://doi.org/10.1038/s42256-023-00692-8.

[142] Ilia Stepin et al., « Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers », *in*: *Proc. International Conference on Fuzzy Systems FUZZ-IEEE*, 2020, DOI: 10.1109/FUZZ48607.2020.9177629.

[143] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, « Axiomatic Attribution for Deep Networks », *in*: *CoRR* abs/1703.01365 (2017).

[144] Robert L. Thorndike, « Who belongs in the family? », *in*: *Psychometrika* 18.*4* (1953), pp. 267–276, ISSN: 1860-0980.

[145] Chris Thornton, *Truth from trash: How learning makes sense*, Mit Press, 2002.

[146] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux, « How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies », *in*: *Proc. ACM Hum. Comput. Interact.* 5.*CSCW2* (2021), pp. 1–39, DOI: https://doi.org/10.1145/3476068.

[147] Sahil Verma, John P. Dickerson, and Keegan Hines, « Counterfactual Explanations for Machine Learning: A Review », *in*: *CoRR* abs/2010.10596 (2020), DOI: https://arxiv.org/abs/2010.10596.

[148] Giorgio Visani, Enrico Bagli, and Federico Chesani, « OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms », *in*: *Proc. CIKM*, vol. 2699, CEUR Workshop Proceedings, CEUR-WS.org, 2020, URL: http://ceur-ws.org/Vol-2699/paper03.pdf.

[149] Giorgio Visani et al., « Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models », *in*: *J. Oper. Res. Soc.* 73.*1* (2022), pp. 91–101, DOI: 10.1080/01605682.2020.1865846, URL: https://doi.org/10.1080/01605682.2020.1865846.

[150] Jasper van der Waa et al., « Evaluating XAI: A comparison of rule-based and example-based explanations », *in*: *Artif. Intell.* 291 (2021), p. 103404, DOI: https://doi.org/10.1016/j.artint.2020.103404.

[151] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell, « Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR », *in*: *CoRR* (2017), DOI: http://arxiv.org/abs/1711.00399.

[152] Xiao-Yong Wei and Chong-Wah Ngo, « Ontology-enriched semantic space for video search », *in*: *Proc. International Conference on Multimedia*, ACM, 2007, URL: https://doi.org/10.1145/1291233.1291447.

[153] Katharina Weitz et al., « "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design », *in*: *Proc. International Conference on Intelligent Virtual Agents*, IVA '19, Paris, France: Association for Computing Machinery, 2019, ISBN: 9781450366724, URL: https://doi.org/10.1145/3308532.3329441.

[154] Daniel S. Weld and Gagan Bansal, « Intelligible Artificial Intelligence », *in*: *CoRR* (2018), DOI: http://arxiv.org/abs/1803.04263.

[155] Joel Wester et al., « On Moral Manifestations in Large Language Models », *in*: *In Proc. CHI Workshop on Moral Agents,* 2023, pp. 1–4.

[156] James Wexler et al., « The What-If Tool: Interactive Probing of Machine Learning Models », *in*: *IEEE Trans. Vis. Comput. Graph.* 26.*1* (2020), pp. 56–65, DOI: https://doi.org/10.1109/TVCG.2019.2934619.

[157] Mike Wu et al., « Optimizing for Interpretability in Deep Neural Networks with Tree Regularization », *in*: *J. Artif. Intell. Res.* 72 (2021), pp. 1–37, DOI: https://doi.org/10.1613/jair.1.12558.

[158] Tongshuang Wu et al., « Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models », *in*: *Proc. ACL/IJCNLP*, ed. by Chengqing Zong et al., Association for Computational Linguistics, 2021, DOI: 10.18653/v1/2021.acl-long.523, URL: https://doi.org/10.18653/v1/2021.acl-long.523.

[159] Linyi Yang et al., « Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification », *in*: *Proc. COLING*, International Committee on Computational Linguistics, 2020, DOI: https://doi.org/10.18653/v1/2020.coling-main.541.

[160] Muhammad Rehman Zafar and Naimul Mefraz Khan, « DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems », *in*: *CoRR* abs/1906.10263 (2019), URL: http://arxiv.org/abs/1906.10263.

[161] Yu Zhang et al., « A Survey on Neural Network Interpretability », *in*: *IEEE Trans. Emerg. Top. Comput. Intell.* 5.*5* (2021), pp. 726–742, DOI: 10.1109/TETCI.2021.3100641, URL: https://doi.org/10.1109/TETCI.2021.3100641.

APPENDIX A

WHEN SHOULD WE USER LINEAR EXPLANATIONS?

This appendix is divided into four parts. In Appendix A.1 we present more information about the datasets and classifiers used in our experimental evaluation. In Appendix A.2 we compare Local Surrogate (LS) and our extension LS_{APE} used by APE to compute the reported linear explanation. In the following section, we report additional results from the evaluation of APE's building blocks: Appendix A.3 extends the analysis presented in Section 3.5.2 (Table 3.3) by showing the effect of unimodality (as reported by APE) on the performance of Local Surrogate. Appendix A.4 provides additional experiments that compare Growing Fields with Growing Spheres in terms of their impact when used for counterfactual search in Local Surrogate.

A.1 Datasets and Classifiers

This section provides useful information to reproduce the presented experimental results.

— The source code is available in a repository on github [1];

— Each dataset can be downloaded from the link listed in Table A.1;

— The parameters required to replicate the scikit-learn artificial datasets are provided in Table A.2.

1. https://github.com/j2launay/APE-Adapted_Post-Hoc_Explanations

Name	Link
Adult	https://archive.ics.uci.edu/ml/datasets/adult
Titanic	https://www.kaggle.com/c/titanic/overview
Blood	https://www.openml.org/d/1464
Diabete	https://www.openml.org/d/37
Compas	https://github.com/propublica/compas-analysis/
Mortality	https://github.com/suinleelab/treeexplainer-study/tree/master/notebooks/mortality

Table A.1 – Dataset links.

Cat Blobs dataset is generated through the make_blobs function (Blob, Blobs or M Blobs from Table A.2) where 4 features have been randomly discretized into binary variables depending on either the feature value is (1) superior or (0) inferior to the mean feature value of the dataset.

Table 3.1 summarizes for each dataset – i.e.: synthetic and real – the number of instances as well as the number of categorical and numerical attributes.

Regarding the classifiers, 6 black-box models from scikit-learn have been used. Table A.3 shows the hyperparameters set for learning each classifier. We split each dataset into training (70%) and testing (30%) sets.

A.2 LS vs. LS$_{APE}$

This experiment aims to evaluate various local surrogate models on five datasets (Blood, Diabetes, Blobs, Moons, and Circles). Four different variants of Local Surrogate (LS) have been tested and compared on the accuracy.

Local Surrogate (LS): It corresponds to the version in [83] in which a linear regression model is learned around the closest counterfactual over the classifier's binary answers.

Name	Parameters
Blob	$make\_blobs(1000, n\_features = 2, random\_state = 0, centers = 2, cluster\_std = 1)$
Blobs	$make\_blobs(n\_samples = 5000, n\_features = 12, random\_state = 0, centers = 2, cluster\_std = 5)$
Circles	$make\_circles(n\_samples = 1000, noise = 0.05, random\_state = 0)$
M Blobs	$make\_blobs(7500, n\_features = 20, random\_state = 0, centers = 2, cluster\_std = 5)$
Moons	$make\_moons(n\_samples = 2000, noise = 0.2, random\_state = 0)$

Table A.2 – Parameters set to generate the artificial datasets.

Local Surrogate logistic regression ($LS_{log.}$): We introduce a novel version of Local Surrogate based on logistic regression as an explanation model. Contrary to LS, the surrogate is trained on the classifier's output probabilities.

Local Surrogate raw data (LS_{raw}.): In this method, we use LS with a linear regression model learned over the raw dataset, including no discretization of continuous variables.

Extended Local Surrogate (LS_{APE}): This last method employs a Local Surrogate with a logistic regression, over the classifier's output probabilities, where we extend the size of the sphere and learn a new classifier while the accuracy does not drop. This is the explanation method used by APE.

We obtained the average score for each of the 4 explanation methods on 4 of the black-box models used in Section 3.5 and presented in Table A.4. We remove the Ridge Classifier from the list of classifiers since it can not return classification probabilities and is thus unable to be combined with LS and $LS_{raw.}$.

Accuracy is computed over 50 instances for each linear explanation. Table A.4 shows that LS_{APE} obtains the best accuracy results on 12 out of 20 before LS_{raw} with 7 better accuracy results. Logistic Regression embedded in LS ($LS_{log.}$) obtains worse results than the traditional Local Surrogate method.

To conclude, we observed that the proposed method LS_{APE} consists of training the linear model explanation over instances from the hyperfield and extending the radius of the hyperfield

Name	Parameters
GB	GradientBoostingClassifier($n\_estimators$ = 20, $learning\_rate$ = 1.0, $random\_state$ = 1)
MLP	MLPClassifier($random\_state$ = 1)
RF	RandomForestClassifier($n\_estimators$ = 20, $random\_state$ = 1)
VOT	VotingClassifier($estimators$ = [('lr', $LogisticRegression()$), ('gnb', $GaussianNB()$), ('svm', $svm.SVC(probability = True)$)], $voting$ = "$soft$")
RC	RidgeClassifier($random\_state$ = 1)
SVM	SVC($probability = True, random\_state$ = 1)

Table A.3 – Parameters to set the classifiers.

as long as the accuracy of the linear explanation is above the 95% threshold, gives the best results.

A.3 Additional Linear Suitability Evaluation

Similarly to Table 3.3 and as mentioned in Section 3.5.2, we furnish in respectively Table A.5 and Table A.6 the average accuracy score over 100 instances depending on either APE claims that a linear explanation is adapted (Uni) or not (Mul) for (a) Extended Local Surrogate dealing with Uni and Mul situation depending only on the Thornton's *separability index* introduced in Section 3.4.2 and (b) Local Surrogate classic over continuous datasets.

While Extended Local Surrogate (LS_{APE}) selected for unimodal cases with the Thornton acquires the highest accuracy score in 34 cases over 44 times, we highlight that for the categorical datasets, LS_{APE} obtains lower scores 10 out of 19 cases. Hence, we argue that the usage of the unimodality test over the set of friends and enemies introduced in Section 3.4 is necessary to detect the suitability of linear explanations for datasets composed of categorical features.

	GB				MLP				RF				VC			
	LS	$LS_{log.}$	$LS_{raw.}$	LS_{APE}	LS	$LS_{log.}$	$LS_{raw.}$	LS_{APE}	LS	$LS_{log.}$	$LS_{raw.}$	LS_{APE}	LS	$LS_{log.}$	$LS_{raw.}$	LS_{APE}
Blood	0.52	0.59	0.62	**0.92**	0.74	0.74	0.75	**0.92**	0.56	0.53	0.60	**0.90**	0.85	0.74	0.85	**0.91**
Diabetes	0.67	0.40	0.65	**0.86**	0.75	0.72	0.74	**0.91**	**0.65**	0.36	0.64	0.50	**0.86**	0.52	0.84	**0.86**
Blobs	0.91	0.90	**0.97**	0.90	0.96	0.91	**0.98**	0.91	0.91	0.89	**0.95**	0.91	0.95	0.90	**0.96**	0.89
Moons	0.68	0.83	0.73	**0.93**	0.77	0.80	0.85	**0.89**	0.65	0.75	0.70	**0.91**	0.93	**0.98**	0.93	0.91
Circles	0.80	0.52	0.81	**0.92**	0.98	0.49	**0.99**	0.90	0.89	0.50	**0.90**	**0.90**	0.87	0.85	**0.93**	0.81

Table A.4 – Average accuracy score for Local Surrogate, a Local Surrogate employing a logistic regression $LS_{log.}$, a Local Surrogate training over raw data $LS_{raw.}$ and our extended Local Surrogate $LS_{APE.}$ on multiple datasets and classifiers.

	GB		MLP		RF		VC		SVM	
	Uni	Mul	Uni	Mul	Uni	Mul	Uni	Mul	Uni	Mul
Adult	**0.614**	0.293	**0.402**	0.276	0.590	\	0.226	**0.362**	**0.626**	0.250
Blob	**0.896**	0.790	**0.890**	0.759	**0.876**	0.729	**0.899**	0.748	**0.894**	0.746
Blobs	**0.855**	0.636	**0.723**	0.606	**0.783**	0.655	**0.745**	0.610	**0.717**	0.599
Blood	**0.679**	0.438	**0.580**	0.498	**0.602**	0.282	**0.598**	0.224	**0.651**	0.624
Cat Blobs	**0.905**	0.899	0.849	**0.897**	0.892	**0.950**	0.883	**0.944**	**0.883**	0.794
Circles	0.945	\	0.958	\	0.949	\	0.948	\	0.949	\
Compas	0.729	\	0.881	**0.981**	0.761	**0.995**	0.128	\	0.474	**0.955**
Diabetes	**0.625**	0.400	**0.785**	0.585	**0.512**	0.456	0.698	\	0.735	\
M Blobs	\	0.833	\	0.967	**0.863**	0.845	\	0.947	**0.944**	0.942
Moons	**0.925**	0.706	**0.914**	0.802	**0.919**	0.728	**0.916**	0.881	**0.918**	0.750
Mortality	0.819	**1.000**	**1.000**	**1.000**	**0.867**	0.727	0.513	**0.556**	0.414	**0.573**
Titanic	**0.761**	0.667	0.919	\	0.974	\	**0.999**	0.997	0.715	\

Table A.5 – Average accuracy computed on 100 instances per black-box model and dataset of LS_{APE} when the Thornton's *separability index* indicates unimodal (Uni) and multimodal (Mul) decision boundaries. A \ denotes a non-meaningful accuracy score, i.e., there were less than 5 instances falling in that case.

Results in Table A.6 show that the average score of Local Surrogate in situations where APE indicates that a linear explanation fits the given situation obtains the highest accuracy score in 20 cases out of 23. Hence, we can conclude that (a) the fidelity of a linear explanation depends on the complexity of the decision border and (b) APE is able to determine whether a linear will perform well.

	GB		MLP		RF		VC		SVM	
	Uni	Mul	Uni	Mul	Uni	Mul	Uni	Mul	Uni	Mul
Blob	**0.727**	0.600	**0.757**	0.621	**0.729**	0.597	**0.750**	0.608	**0.743**	0.612
Blobs	**0.638**	0.605	**0.608**	0.571	**0.661**	0.617	**0.613**	0.584	**0.607**	0.570
Blood	\	0.435	\	0.481	\	0.279	\	0.222	\	0.616
Circles	**0.712**	0.673	0.702	\	**0.711**	0.654	0.685	\	0.716	\
Diabetes	0.390	**0.399**	**0.607**	0.604	\	0.453	0.249	**0.260**	0.508	**0.517**
M Blobs	\	0.793	\	0.925	**0.850**	0.781	\	0.885	**0.942**	0.883
Moons	**0.714**	0.692	**0.796**	0.772	**0.729**	0.715	**0.890**	0.880	**0.751**	0.719

Table A.6 – Average accuracy computed on 100 instances per black-box model and dataset of Local Surrogate when APE indicates unimodal (Uni) and multimodal (Mul) decision boundaries. A \ denotes a non-meaningful accuracy score, i.e., there were less than 5 instances falling in that case.

A.4 Growing Fields vs. Growing Spheres

Table A.7 presents the accuracy of the Local Surrogate methods centered on the counterfactual generated by Growing Fields (GF) and Growing Spheres (GS). We focus only on linear explanations, on five datasets having no categorical attributes, five black boxes, and 50 instances.

	RF		VC		GB		RC		MLP	
	GS	GF	GS	GF	GS	GF	GS	GF	GS	GF
Blobs	0.55	**0.87**	0.55	**0.87**	0.55	**0.87**	0.58	**0.88**	0.51	**0.89**
Blood	**0.88**	0.80	0.48	**0.72**	0.71	**0.81**	0.44	**0.73**	0.50	**0.77**
Circles	0.52	**0.85**	0.57	**0.78**	0.42	**0.85**	0.57	**0.93**	0.53	**0.93**
Diabetes	**0.62**	0.47	0.65	**0.88**	**0.69**	0.32	**0.59**	0.21	**0.78**	0.26
Moons	0.41	**0.89**	0.62	**0.91**	0.34	**0.87**	0.61	**0.92**	0.68	**0.91**

Table A.7 – Average accuracy score of Local Surrogate with GF vs. Local Surrogate with GS on multiple datasets.

Results show that the Local Surrogate with GF outperforms the Local Surrogate with GS 20 out of 25 times. Thus, we can conclude that GF is better than GS at generating an artificial counterfactual on which we center the GF generation process of artificial instances to train our linear surrogate model compared to the GS sampling.

APPENDIX B

IMPACT OF EXPLANATION TECHNIQUES AND REPRESENTATIONS ON USERS

This appendix is divided into two parts. In Appendix B.1 we present more information about the code, classifier, and datasets used in our experimental evaluation. Thereafter, in Appendix B.2, we present the different questions and surveys used throughout the entire experimental process.

B.1 Code and Data Processing

This section provides useful information to reproduce the presented experimental results. The source code is available in a repository on GitHub [1].

Compas: In order to generate explanations meaningful to the users, we removed some features and kept this subset of features {Gender, Age, Race, Juvenile felony count, Juvenile misdemeanor count, Priors count, Charge degree, Charge description}. We also removed 508 individuals having a charge description that occurred less than 5 times in the whole dataset. The dataset can be downloaded online [2].

1. https://github.com/j2launay/user_eval
2. https://github.com/propublica/compas-analysis/

Obesity: This dataset is originally composed of 16 features and a target obtained from questions detailed in [114]. However, we removed the weight since it would be too easy for the model and the user to predict the BMI with both the height and weight. We binaries five features: Gender, family history with overweight, does the user smokes, calorie consumption monitoring, and does the user frequently consumes high-caloric food. The other features were one hot encoded, the original data can be downloaded on this link [42] [3].

B.2 Questionnaire

In our survey, we ask the online users to complete three different questionnaires, each one evaluating a given criteria. We present in this section the question and where each questionnaire comes from.

B.2.1 Satisfaction Scale

We now present the questions to evaluate the users' perceived satisfaction with the system from Hoffman et al. [66]. This questionnaire is composed of 8 questions:

1. From the explanation, I understand how the tool works.
2. This explanation of how the tool works is satisfying.
3. This explanation of how the tool works has sufficient detail.
4. This explanation of how the tool works seems complete.
5. This explanation of how the tool works tells me how to use it.
6. This explanation of how the tool works is useful to my goals.
7. This explanation of the tool shows me how accurate the tool is.
8. This explanation lets me judge when I should trust and not trust the tool.

For each of these questions, Hoffman et al. recommend these 5 Likert scales:

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

B.2.2 Trust Scale

We first use the Cahour-Forzy scale [22] composed of these 4 items with a 7 Likert scale as follows:

1.

What is your confidence in the tool? Do you have a feeling of trust in it?						
I do not trust it at all.	2	3	4	5	6	**I trust it completely.**

2.

Are the actions of the tool predictable?						
It is not at all predictable.	2	3	4	5	6	**It is completely predictable.**

3.

Is the tool reliable?						
It is not at all reliable.	2	3	4	5	6	**It is completely reliable.**

4.

Is the tool efficient at what it does?						
It is not at all efficient.	2	3	4	5	6	**It is completely efficient.**

B.2.3 Understanding Scale

We now present the questions to evaluate the users' perceived understanding of the system from Madsen and Gregor [97]. This questionnaire is composed of 8 questions:

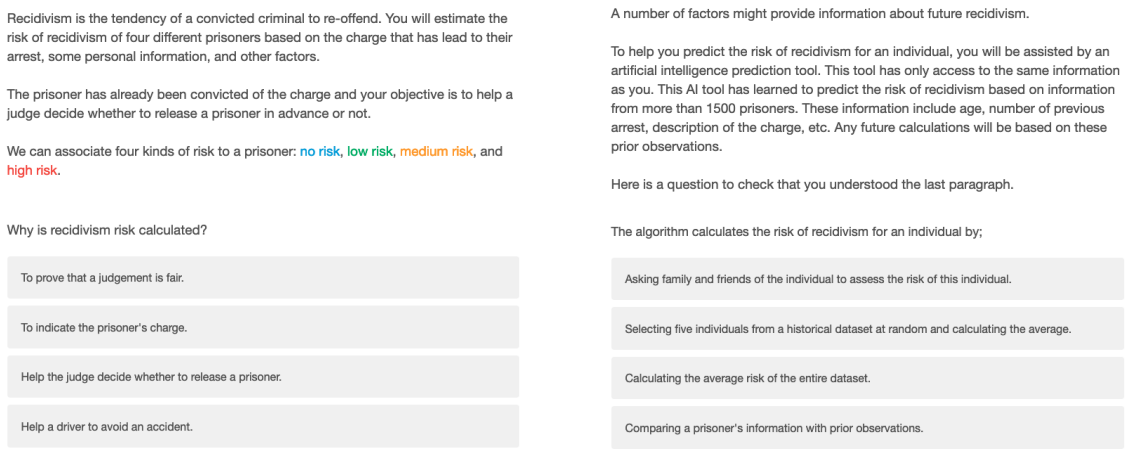
1. The system uses appropriate methods to reach decisions.
2. The system has sound knowledge about this type of problem built into it.
3. The advice the system produces is as good as that which a highly competent person could produce.
4. The system makes use of all the knowledge and information available to it to produce its solution to the problem.
5. I know what will happen the next time I use the system because I understand how it behaves.
6. I understand how the system will assist me with decisions I have to make.

7. Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.

8. It is easy to follow what the system does.

For each of these questions, Madsen and Gregor recommended this 5 Likert scale:

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly



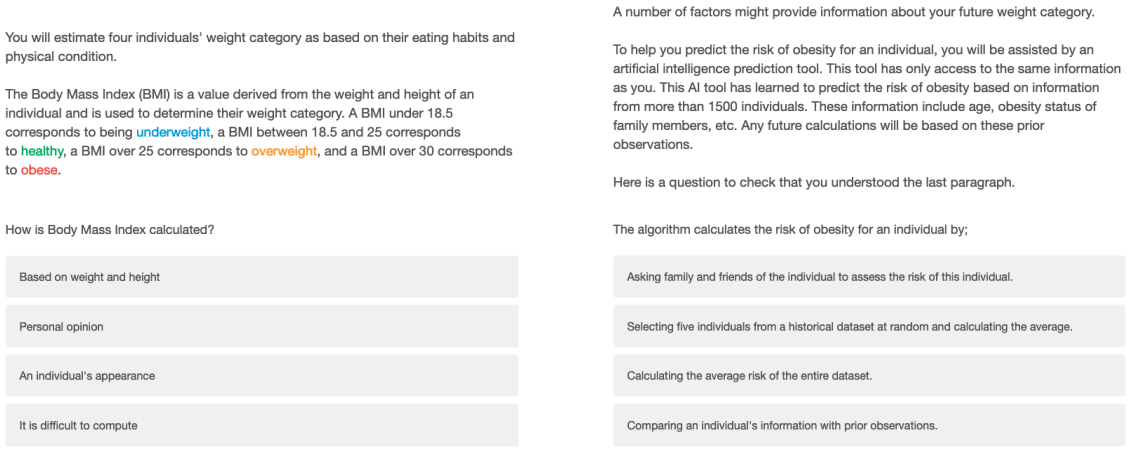
(a) (b)

Figure B.1 – Detailed presentation of the two verifying questions at the end of the Compas dataset survey.

B.2.4 Question to verify user's validity

We ask the user two questions in order to verify that they understand and will try efficiently to complete the questionnaire.

Following the task introduction, we assessed whether the participants had actually read and understood the task through two questions: 'How is Body Mass Index calculated?' for the Obesity dataset and 'Why is recidivism risk calculated?' for COMPAS. We found 10 and 30 incorrect answers for the first and second questions, respectively. This question had the form 'The algorithm calculates the risk of obesity (resp. recidivism) for an individual by;'. We asked additional users to participate in our study until we had 20 responses for each group that validated our two understanding questions resulting in a final set of 280 participants.



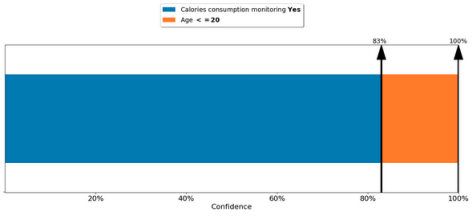
(a) (b)

Figure B.2 – Detailed presentation of the two verifying questions at the end of the obesity dataset survey.

B.2.5 Explanation Paragraph in Example Round

During the introduction and more precisely when we showed for the first time an explanation to the participant, we described more completely how to grasp the different graphics as follows:

Based only on the above information, the artificial intelligence (AI) tool has predicted **healthy**.
The following graph shows the criteria that impacted the AI's prediction. Each of the colored bars represent the importance of one particular user's answer to the final prediction.
The numerical values at the top correspond to the increasing confidence that the AI tool predicts **healthy** for this user.

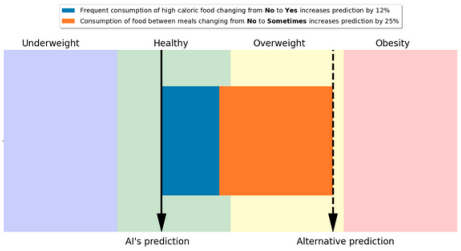


You now know everything required to proceed to the tasks!

(a) Rule-based.

As highlighted in the graph below and based only on the above information, the AI tool has predicted **healthy**.

The following graph shows the criteria that impacted the AI's prediction. The AI computes a value between 0% and 100% to classify the individual. This value corresponds to the "AI's prediction" vertical black bar and falls into one of the four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), and **obesity** (above 75%).



The colored bars indicate what the individual must do in order to modify the AI's prediction the most effectively. The length of the bars correspond to the importance of changing one answer's value to another.

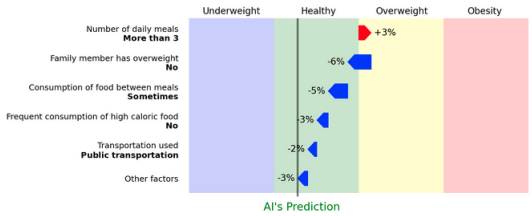
You now know everything required to proceed to the tasks!

(b) Counterfactual.

Based only on the above information, the AI tool has predicted **healthy**.

The following graph shows the criteria that impacted the AI's prediction. The red bars indicate an increased chance of being overweight and obese. The blue bars indicate an increased chance of being underweight or healthy.

The values on the side of the bars correspond to the impact of the specific factor on the prediction. The "Other parameters" bar indicates the impact of all other factors not presented in the graph.



By summing the values associated with each response by the AI, we obtain a value between 0% and 100%. This value corresponds to the vertical black bar and falls in one of the four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), **obesity** (above 75%).

You now know everything required to proceed to the tasks!

(a) Linear.

Figure B.4 – Detailed presentation of the three graphs presentation in the introduction and more precisely the first time the participant had access to an explanation in the survey.

LIST OF FIGURES

1 Illustration of eXplainable AI 12

1.1 Illustration of two models, a linear model and a decision tree, employed to predict loan approval likelihood based on applicants' age and salary. 19
1.2 Taxonomy of the explanation techniques. 22
1.3 Rule-based explanation for a sentiment classification model. 25
1.4 Feature-attribution explanation for a sentiment classification model. 27
1.5 Counterfactual explanations for a sentiment classification model that predicts a text as positive. 29
1.6 Schema of the research plan followed in the thesis. 38

2.1 Two anchors (depicted as green lines) learned with different discretizations of the numerical features. The target instance is marked as a violet star 41
2.2 Trade-off between F1 score and anchor length for three neighborhood generation methods 42

3.1 Two explanation scenarios for a classifier and a target instance (the filled star): (a) a suitable single linear explanation; (b) three contradictory linear explanations. 55
3.2 Illustration of a linear explanation for a classifier and a target instance x. . . . 57
3.3 Illustration of the tree structure of the algorithm used to iteratively perturb the target document. 63
3.4 Representation of the mechanism employed to generate sets of potential word replacements in Growing Language and Growing Net. 64
3.5 Illustration of the Adapted Post-hoc Explanation framework. 66
3.6 Description of APE Oracle assessing the suitability of a linear surrogate for approximating a classifier locally. 68
3.7 An illustrative example of the moon dataset. 72

3.8 Average Kendall's rank correlation coefficient of the LS_{APE}'s explanations computed on 100 instances for 7 tabular datasets and 4 "glass" black-box models across the Oracle's outcomes. 79

3.9 Average accuracy per black-box model on 100 instances of the experimental tabular datasets for APE_a and APE_t. 81

3.10 Average accuracy achieved by APE_a and APE_t on each black-box model based on 100 instances drawn from the textual datasets. 82

3.11 Average accuracy per black box computed on 100 instances of the tabular datasets for APE_a in (a) and for APE_t in (b) when we remove the Libfolding unimodality and linear separability tests from the APE Oracle. 83

3.12 Average accuracy attained by APE_a in (a) and for APE_t in (b) across different black-box models and textual datasets. 84

3.13 Average accuracy comparison between the outcomes of APE_a and APE_t based on the space used for the APE Oracle's test. 85

3.14 Average Mahalanobis distance between the Growing Spheres (GS) and Growing Fields (GF) counterfactual instances and their closest real enemy. 88

4.1 Illustration of an artificial decision boundary around a target text and its counterfactuals. 94

4.2 Illustration of the mechanism employed to perturb the target documents by the transparent and opaque methods. 95

4.3 Spectrum for counterfactual explanation techniques that goes from the most transparent methods on the left to the most opaque methods, passing by our methods in red. 99

4.4 Minimality as the distance between the closest counterfactual and the target document (the lower the better). 103

4.5 Outlierness, quantified as the measure of distance between the counterfactuals and the nearest instance within the test set. 104

4.6 Perplexity as the MSE loss of a GPT model on the generated counterfactuals. 105

4.7 Stability as the average similarity between the generated counterfactuals for five successive runs on a single text document. 106

5.1 Diagram depicting the experimental workflow proposed in this thesis and used to assess user perception and behavior when interacting with a given explanation technique. Behavioral measurement steps are indicated in green, while self-reported measurement steps are in blue. The task rounds are repeated n times. 117

6.1 Example of two individuals presented to the users for the Obesity and COMPAS datasets. 125
6.2 Different explanations for a given individual on the Obesity dataset, as split by explanation technique and representation. 126
6.3 Example of answers from participant "User J" from the rule-based explanation group. 129
6.4 Example of answers from one participant to the Trust survey 131
6.5 Perceived understanding of the users **(Immediate Und.)** for both the Obesity and Recidivism datasets based on the explanation technique. 135
6.6 Behavioral precision between the features indicated as important by the users for the AI's prediction and the important features indicated in the explanation. 136
6.7 Users' understanding for each explanation technique and representation, computed as the users' recall of the important features according to the explanations in the Obesity dataset. 136
6.8 Difference between the self-reported confidence in the users' prediction after and before seeing the AI's prediction and explanation (when provided). 139
6.9 Proportion of cases the participants changed their initial prediction to follow the AI's prediction. 140

B.1 Detailed presentation of the two verifying questions at the end of the Compas dataset survey. 186
B.2 Detailed presentation of the two verifying questions at the end of the obesity dataset survey. 187
B.4 Detailed presentation of the three graphs presentation in the introduction and more precisely the first time the participant had access to an explanation in the survey. 188
B.5 Schema du plan de recherche suivi dans la thèse. 204

LIST OF TABLES

1.1 Notation used in the thesis. . . . 24
1.2 Example of counterfactual, feature attribution, and rule-based explanations on tabular data. . . . 24

2.1 F1 score for Anchors using different discretization methods. MR denotes the mean rank of the method. . . . 46
2.2 Anchor length using different discretization methods. MR denotes the mean rank of the method. . . . 47
2.3 Length of textual Anchors for different neighborhood generation strategies. . . . 49
2.4 F1 score of textual Anchors for different neighborhood generation strategies. . . . 49

3.1 Number of instances, numerical and categorical features for each dataset. . . . 74
3.2 Information about the experimental datasets with textual data. . . . 75
3.3 Average accuracy per black-box model and tabular dataset for LS_{APE} in relation to both Oracle outcomes. . . . 77
3.4 Average accuracy per black-box model and textual dataset for LS_{APE} with respect to both oracle outcomes. . . . 78
3.5 The average difference in the model prediction's probability between the original text and the same text without the words identified as important by LS_{APE}. . . . 80
3.6 Average runtime (in seconds) of Growing Spheres (GS) and Growing Fields (GF) over 100 instances per black box and dataset. . . . 89

4.1 Information about the experimental datasets. . . . 101
4.2 Average label flip per dataset and black box of the six counterfactual methods. 105
4.3 Average runtime in seconds for an instance (and standard deviation) per dataset and black box of the six counterfactual methods. . . . 107

6.1 Datasets composition. . . . 123

6.2 Overview of participants' demographic factors. 133
6.3 F value of the ANOVA Table with understanding measurements grouped by domain and self-reported and behavioral metrics. 134
6.4 F value of the ANOVA Table with trust measurements grouped by domain and by self-reported and behavioral metrics. 138
6.5 F value of the ANOVA table with additional measurements. 140

A.1 Dataset links. 178
A.2 Parameters set to generate the artificial datasets. 179
A.3 Parameters to set the classifiers. 180
A.4 Average accuracy score for various versions of Local Surrogate on multiple datasets and classifiers. 181
A.5 Average accuracy per black-box model and dataset of LS_{APE} when the *separability index* indicates unimodal and multimodal decision boundaries. 181
A.6 Average accuracy per black-box model and dataset of Local Surrogate when APE indicates unimodal and multimodal decision boundaries. 182
A.7 Average accuracy score of Local Surrogate with GF vs. Local Surrogate with GS on multiple datasets. 182

RÉSUMÉ EN FRANÇAIS

Contents

A.1 Datasets and Classifiers 177
A.2 LS vs. LS$_{APE}$ 178
A.3 Additional Linear Suitability Evaluation 180
A.4 Growing Fields vs. Growing Spheres 182

Contexte

La mode de l'intelligence artificielle n'est plus un rêve depuis l'augmentation des puissances de calcul et la disponibilité exponentielle de données permettant l'émergence des modèles d'apprentissage automatique. En effet, leur application est partout, allant de la recommandation de films et de musiques à la découverte de nouveaux médicaments en passant par la prédiction de spam. Ces algorithmes prennent en entrée des informations, ici un profil utilisateurs ou un courriel et, suite à un entraînement sur différents exemples, apprennent à déterminer si cette entrée appartient davantage à une classe qu'à une autre. Tant que ces modèles étaient peu utilisés ou dans des domaines à faible risque tels que la recommandation de films, le principal objectif était d'améliorer ces modèles. Les modèles sont ainsi devenus de plus en plus complexes, pouvant comporter des milliards de paramètres [4]. Cependant, leurs installations dans des domaines à risques tels que la santé ou la loi [78] ont conduit à de mauvaises interprétations. Par exemple, aux Etats-Unis, l'utilisation du logiciel COMPAS pour prédire le risque de récidive des prisonniers américains a alimenté les préjugés raciaux [5]. Aux Pays-Bas, l'État a eu recours

4. Modèle BERT: 110 million de paramètres; GPT4: 1.76 trillion de paramètres.
5. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

à un algorithme permettant de détecter des fraudes [6]. Cet algorithme a injustement accusé des familles, souvent binationales ou immigrées, ce qui a donné lieu à des sanctions abusives et à des troubles financiers. De même, Amazon a développé un projet visant à automatiser le processus d'embauche de son personnel à l'aide d'un algorithme d'apprentissage automatique. Il s'est avéré que cet algorithme présentait des préjugés sexistes, engendrant une discrimination à l'égard des femmes et soulevant des préoccupations en matière de justice et d'équité [7]. Une des raisons pour justifier l'arrivée de ces biais est le manque de transparence des modèles [43]. En effet, il est difficile voire impossible pour un humain de vérifier qu'un modèle qui possède des milliards de paramètres fonctionne vraiment et utilise uniquement les attributs souhaités, au lieu de biais dans le jeu de données d'entraînement. La situation est suffisamment sérieuse pour justifier que des chercheurs travaillent à résoudre cette question. Le domaine de l'explicabilité des algorithmes d'apprentissage automatique est revenu au centre des débats économique, politique, juridique et même philosophique, avec des applications dans tous les domaines, allant des sciences sociales à la médecine et aux mathématiques [68].

Dans cette thèse, nous nous concentrons sur la génération d'explications permettant d'identifier les principaux facteurs influençant les prédictions d'un modèle d'apprentissage automatique. En clarifiant le processus de prise de décision, ces explications visent à renforcer la confiance des utilisateurs et à améliorer la fiabilité et la responsabilité des modèles d'apprentissage automatique.

Besoin de transparence

L'utilisation de méthodes d'apprentissage automatique, et en particulier de modèles d'apprentissage profond, pose de nombreux challenges qui empêchent de pouvoir profiter pleinement de leurs capacités. Un des problèmes majeurs est celui de la transparence qui vient de la nature complexe des modèles les plus performants. En effet, l'amélioration des capacités des modèles d'apprentissage automatique vient conjointement avec l'augmentation de leur complexité, les rendant opaques à leurs utilisateurs. Les modèles les plus performants étant fréquemment des réseaux de neurones avec des millions de paramètres ou des modèles ensemblistes combinant les avantages de différents modèles, il est devenu impossible de déterminer quel est l'élément qui impacte le plus leurs prédictions. Or, comprendre le fonctionnement de ces algorithmes

6. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

7. https://aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against

devrait permettre une meilleure utilisation passant par des possibilités de résolution de biais avant le déploiement et ainsi une plus grande confiance des utilisateurs [91, 129].

L'apparition de ces modèles dans des milieux à risques a obligé les différents États à prendre des décisions et à définir des lois ou des restrictions sur l'usage des algorithmes. En Europe en particulier, nous avons vu l'apparition en 2016 du Règlement Général sur la Protection des Données, ou RGPD, qui permet à n'importe quel individu de l'Union Européenne de demander une explication lorsqu'une décision l'impactant est prise en utilisant un algorithme. De la même manière, l'acte européen sur l'intelligence artificielle a été adopté par le Parlement européen en 2021. Ces différentes exigences ont donc poussé à un renouveau sur l'éthique de l'intelligence artificielle, amenant un nouvel essor des méthodes d'explication.

Expliquer les modèles d'apprentissage automatique

Au cours de cette dernière décennie, les méthodes d'explication sont devenues omniprésentes, à l'image des algorithmes qu'elles cherchent à expliquer [68]. Ces méthodes ont pour objectifs de montrer quels sont les éléments les plus importants ayant amené le modèle à faire telle ou telle prédiction. Pour illustrer, prenons l'exemple simplifié présenté dans la Figure 1, représentant un modèle conçu pour détecter les fausses informations à partir de titres de journaux. Ce modèle est entraîné sur des titres existants, chacun étant étiqueté comme vrai ou faux. Cette étape d'entraînement permet au modèle de prédire l'étiquette d'une nouvelle instance, comme un titre de journal dans notre exemple. La prédiction résultante est ensuite transmise à l'utilisateur final, dans ce cas, un journaliste, qui doit prendre une décision en se basant sur cette prédiction. L'étape d'explicabilité de ce processus permet au journaliste d'explorer quels éléments de l'article d'origine contribuent à la décision du modèle. Par exemple, une méthode d'explicabilité peut indiquer les mots spécifiques dans le titre du journal qui ont conduit le modèle à le classer comme vrai. Alternativement, une autre technique d'explication pourrait identifier des mots sensibles qui, s'ils étaient supprimés ou modifiés de certaines manières, auraient conduit à prédire "faux".

Au moment de l'essor de l'explicabilité, la plupart des recherches sur la génération d'explications pour les modèles d'IA étaient menées par des chercheurs en apprentissage automatique. Ainsi, la première préoccupation de la communauté XAI (Explainable AI) a été de générer des explications précises et fiables pour les prédictions de modèles. Dans ce contexte, cette thèse se concentre sur l'étude des explications locales, qui visent à expliquer la prédiction d'un modèle pour une *instance cible*. Cette instance peut revêtir diverses formes, telles que des informations sur un individu ou des données textuelles, comme illustré dans la Figure 1. La communauté

XAI a développé de nombreuses méthodes diverses pour expliquer la prédiction d'un modèle pour une instance donnée [15, 68]. Ces méthodes vont des règles de décision simples [80, 124], aux modèles linéaires [52, 123], et à la présentation d'entrées similaires générant des sorties différentes du modèle [82, 151]. Une observation notable émerge : la recherche existante s'est principalement concentrée sur la génération des meilleures techniques d'explication censées fonctionner pour chaque instance, modèle et utilisateur. Cependant, comme le démontre cette thèse, la quête de telles solutions universelles peut ressembler à la recherche du mythique El Dorado. Aucune technique unique ne peut s'adapter à tous les contextes de données et d'utilisateurs.

Étant donné l'abondance des techniques d'explication existantes, cette thèse vise à identifier certaines limites des méthodes d'explication actuelles. Puisque la qualité de l'explication peut être impactée par différents aspects tels que le type de modèle à expliquer ou le profil des utilisateurs, cette thèse est divisée en deux parties. Tout d'abord, nous nous concentrons sur la génération d'explications adaptées aux données. Par conséquent, notre première question de recherche est : **Comment générer la meilleure explication d'un point de vue des données ?** De la même manière, une explication ne pourra pas être identique qu'elle soit donnée à un utilisateur lambda, un professionnel du domaine ou un informaticien [125, 38]. Or, bien qu'il soit largement accepté que les explications devraient être adaptées aux utilisateurs qui les reçoivent, l'aspect centré sur l'utilisateur a été sous-représenté dans la littérature [1, 4]. Ainsi, nous étudions dans la deuxième partie de cette thèse, l'impact des techniques d'explication et des représentations choisies sur les utilisateurs en répondant à la question : **Comment générer la meilleure explication du point de vue des utilisateurs ?**

Comprendre les différents types d'explications

Parmi les méthodes d'explication, on peut identifier différentes stratégies que je vais définir ici pour comprendre la suite de ce résumé.

Explications auto-explicables vs. explications a posteriori

On va commencer par différencier les méthodes qui sont transparentes ou génèrent directement des explications, des méthodes d'explications qui seront utilisées a posteriori sur des modèles déjà entraînés. L'avantage des méthodes directement transparentes est qu'elles sont plus fidèles au modèle car elles utilisent des informations telles que l'architecture du modèle ou se basent sur le même jeu de données que le modèle pour générer des explications. On retrouve

parmi ces méthodes, les modèles linéaires simple, qui associent un poids à chaque paramètre de l'entrée. La valeur de ce poids indique si le paramètre a un impact positif ou négatif sur la prédiction du modèle. De la même manière, un chemin dans un arbre de décision peut être utilisé pour construire des règles permettant de déterminer la classe à laquelle appartient un individu. [80, 134] (cf Figure 1.1). Les méthodes d'explication a posteriori, en revanche, entrent en jeu après qu'un modèle complexe, généralement un algorithme de boîte noire, a été entraîné pour générer des prédictions. Elles sont utiles lorsque nous n'avons pas accès aux mécanismes internes du modèle. Cette inaccessibilité peut être due par exemple à des contraintes de confidentialité ou parce que le jeu de données n'est pas disponible. Ainsi, dans cette thèse, nous allons nous concentrer sur ces méthodes qui ont l'avantage de pouvoir être appliquées sur un modèle déjà déployé, permettant d'être utilisées sur un modèle sans avoir besoin de le ré-entraîner, pouvant causer des coûts importants, qu'ils soient économiques ou environnementaux. Ces méthodes a posteriori peuvent s'avérer particulièrement utiles également pour les modèles de production critique qui font partie intégrante d'une entreprise et qui ne peuvent être remplacés d'un jour à l'autre.

Explication globale vs. explication locale

Une autre distinction à faire lorsque l'on parle de méthode d'explication est de savoir sur quelle partie du modèle l'explication se porte. Est-ce que l'on cherche à expliquer un modèle dans son ensemble, et dans ce cas, on parle d'explication globale ? Où cherche-t-on à obtenir une explication locale, se concentrant sur la prédiction d'un individu en particulier ? Les méthodes d'explication locale vont générer une explication pour une instance ou une donnée précise, se concentrant uniquement sur une localité autour de cette instance [123, 124, 94]. Les méthodes d'explications locales sont particulièrement utiles pour les modèles complexes qui ne sont pas directement interprétables telles que les réseaux de neurones ou les modèles ensemblistes. A l'inverse, les explications globales sont précieuses pour obtenir des informations sur le comportement global du modèle et identifier des schémas ou des biais éventuels qui pourraient exister [80, 134, 140]. Parmi les techniques couramment utilisées pour générer des explications globales, on retrouve les graphiques de dépendance partielle, les scores d'importance des caractéristiques et la visualisation des frontières de décision [104].

Un défi majeur dans la conception des techniques d'explication consiste à trouver le bon équilibre entre la fidélité au modèle complexe et la compréhensibilité ou simplicité de l'explication. Ainsi, les explications globales peuvent manquer de fidélité au modèle complexe, car des méthodes simples telles que les arbres de décision ou la régression logistique couram-

ment utilisés pour générer des explications, ne peuvent pas approximer entièrement ou expliquer chaque variation dans la sortie d'un modèle comportant un grand nombre de paramètres. En revanche, l'approximation locale de la frontière de décision permet aux techniques d'explication de conserver une grande fidélité au modèle dans un contexte local tout en restant facilement compréhensible. Pour cette raison, dans cette thèse nous allons nous concentrer uniquement sur les méthodes d'explication locale.

Explications dépendantes aux modèles vs. agnostiques aux modèles

Pour terminer cette taxonomie, je vais différencier les méthodes d'explications qui sont applicables sur n'importe quel modèle d'apprentissage automatique et qui leur sont donc agnostiques de celles qui sont dépendantes aux modèles et utilisent des informations sur le modèle à expliquer pour générer de meilleures explications. Ces dernières profitent par exemple de la structure des arbres pour générer des explications plus fidèles aux modèles de forêts aléatoires [8, 157]. D'autres peuvent bénéficier de l'architecture des réseaux de neurones pour mieux approximer leur fonctionnement [138]. À l'inverse, les explications agnostiques aux modèles font la présomption de n'avoir aucune information sur le type de modèle à expliquer et sont ainsi applicables sur n'importe quelle architecture. Cette flexibilité permet à ces méthodes d'être utilisées sur plus de domaines, tâches et scénarios. Ces techniques donnent la priorité à la transparence et à la généralisation, permettant une approche plus polyvalente et inclusive de l'interprétabilité des modèles. En outre, les explications agnostiques fournissent une approche cohérente et uniformisée pour comprendre les comportements des modèles, ce qui les rend inestimables dans les scénarios où le déploiement de différents modèles est courant. Ces techniques permettent aux chercheurs et aux praticiens de mieux comprendre les processus décisionnels des différents modèles sans avoir besoin d'adaptations spécialisées ou de modifications de la méthodologie d'explication. C'est pour ces raisons, que dans cette thèse nous allons uniquement étudier les explications agnostiques aux modèles.

Au cours de cette thèse, j'ai co-introduit une méthode d'explication globale pour les modèles de langue que nous avons nommé Therapy [23]. Therapy utilise la génération coopérative pour générer des textes qui soient le plus représentatif des différentes catégories que le modèle à expliquer peut prédire (par exemple spam ou fausse information). Therapy mesure ensuite parmi ces textes générés quels sont les mots les plus fréquents et donc importants pour que le modèle les classe comme appartenant à une certaine catégorie. Therapy est ainsi la première méthode d'explication globale à la fois agnostique au modèle et qui ne nécessite pas d'information sur

le jeu de données d'entraînement du modèle à expliquer. Therapy ne sera pas plus développé dans ce manuscrit puisqu'étant une méthode d'explication globale.

Les trois types d'explications courants

Comme indiqué au début de ce chapitre, l'interprétabilité peut être atteinte à des degrés différents, certaines méthodes étant considérées comme plus interprétables en raison de leur simplicité et du nombre limité de paramètres (régressions linéaires ou arbres de décision par exemple). Parmi ces méthodes, trois types d'approches se distinguent par leur utilisation répandue et leur pertinence pour diverses applications [15, 57, 161]. Chacun de ces types d'explications offre un éclairage unique sur le fonctionnement des modèles, permettant aux utilisateurs de mieux comprendre leurs décisions et leurs prédictions.

Cette thèse se concentre sur l'explication des prédictions d'un modèle d'apprentissage automatique, noté f, qui attribue des catégories Y à des instances x du domaine X. Ces instances peuvent être des données tabulaires avec des caractéristiques ou des données textuelles sous forme de séquences de mots. Pour expliquer les prédictions de $f(x) = y$, on utilise des méthodes d'explication par substitution. Celles-ci entraînent un modèle de substitution g qui imite le comportement de f à proximité d'une instance x. Le voisinage de x est déterminé par une fonction ν_x qui indique quelles instances sont considérées comme voisines. La plupart des méthodes de substitution génèrent des explications locales en entraînant g sur des instances artificielles et, si disponibles, sur des instances réelles du voisinage de x. Ainsi, dans cette section, nous explorerons en détail, les explications basées sur des règles, les explications d'attribution de caractéristiques et les explications contrefactuels.

Explication à base de règles

Les règles logiques sont largement reconnues pour leur interprétabilité et ont fait l'objet de nombreuses recherches. Par conséquent, l'extraction de règles constitue une approche attrayante et structurée pour expliquer le fonctionnement des modèles d'apprentissage automatique. Ces règles peuvent être interprétées comme des "si... alors..." qui révèlent les conditions spécifiques sous lesquelles le modèle prend des décisions particulières. Les méthodes fondées sur des règles déterminent les conditions nécessaires sur les caractéristiques de l'instance cible qui permettent à l'IA de prédire un résultat particulier. Ces conditions prennent la forme d'une ou de plusieurs règles de décision appliquées aux caractéristiques d'entrée.

On retrouve par exemple LORE [59] et consor [58, 111], des méthodes qui perturbent aléatoirement l'instance pour laquelle on cherche à expliquer la prédiction d'un modèle et qui vont ensuite entraîner un arbre de décision sur cet ensemble de données artificielles. La structure de l'arbre de décision permet à ces méthodes de générer des règles en suivant le chemin dans l'arbre qui mène à l'instance cible. Similairement, Anchor [124] apprend une règle de décision sur un ensemble de données perturbées. Ces approches présentent l'avantage de produire une explication claire et facilement interprétable.

Explication par attribution de caractéristiques

Dans le domaine de l'explicabilité, l'attribution de caractéristiques met en lumière l'importance relative de chaque attribut, permettant ainsi aux utilisateurs de mieux comprendre comment le modèle prend ses décisions. On retrouve les deux méthodes les plus populaires [107, 68], LIME [123] et SHAP [94] mais aussi de nombreuses variantes [143, 148, 50]. Ces méthodes perturbent l'instance à expliquer et mesurent les changements dans la prédiction du modèle à expliquer en fonction des différentes perturbations. Bien que la représentation de l'explication pour ces deux méthodes est similaire, leur fonctionnement varie légèrement. LIME approxime une boîte noire au moyen d'un modèle linéaire tandis que SHAP mesure la contribution de chacune des caractéristiques dans la prédiction finale à partir de théorie des jeux [136]. L'ampleur de la contribution révèle l'importance de la caractéristique pour un résultat de prédiction particulier, qui peut être corrélé positivement ou négativement avec la réponse fournie par la boîte noire.

Explication par contrefactuels

Les explications contrefactuels sont une approche qui est justifiée à la fois par les sciences sociales [102], mais également par les juristes [151]. Elles explorent les "et si" du modèle d'apprentissage automatique en générant des scénarios alternatifs qui répondent à la question : "Qu'aurait-il fallu changer dans les données d'entrée pour que le modèle prenne une décision différente ?" On retrouve par exemple Growing Spheres [82], une méthode qui perturbe itérativement l'instance cible à l'intérieur d'une hypersphère jusqu'à trouver l'instance la plus proche classifiée différemment. On retrouve d'autres méthodes telles que Polyjuice [158] et Tailor [128] qui utilisent un modèle de langue et des codes de contrôle pour modifier un document textuel. Ces codes de contrôle peuvent par exemple changer la temporalité de la phrase, rajouter de la négation ou modifier le lieu. Ces explications offrent des informations

précieuses sur le comportement du modèle en mettant en évidence les ajustements nécessaires pour influencer ses prédictions.

Comme nous l'avons vu dans ce résumé, de nombreux aspects sont importants pour générer une explication adaptée à la situation. Ceux-ci englobent le type d'explication (règles, attribution de caractéristiques ou contrefactuels), les critères et mesures qu'ils visent à maximiser (décrits dans la version étendue du manuscrit) et le modèle à expliquer (agnostique ou dépendant d'un modèle). Par conséquent, cette thèse aborde deux dimensions essentielles décrites dans la Figure B.5 pour générer des explications adaptées. Tout d'abord, cette thèse aborde la perspective d'améliorer la génération d'explications à centrer sur les données. Chaque chapitre étudie un type d'explication particulier.

Au-delà de la perspective des données, un aspect crucial à prendre en compte lors de la génération d'une explication est la personne qui la reçoit. Ainsi, la deuxième partie de cette thèse étudie comment générer la meilleure explication du point de vue de l'utilisateur. Nous avons donc mené des études sur des utilisateurs lambda avec diverses représentations et techniques d'explication, comme illustré dans la Figure B.5.

Génération d'explications d'un point de vue des données

Cette section résume les travaux que j'ai menés afin de déterminer comment générer de meilleures explications. Cette section est composée de trois sous-sections, chacune synthétisant une partie de mes travaux de thèse. Dans la première, nous avons proposé deux améliorations pour la génération d'explication à partir de la méthode Anchor [124]. Dans la seconde, nous avons introduit un nouvel outil permettant de déterminer si une explication linéaire est adaptée ou non pour approximer un modèle. La dernière contribution de cette partie est une étude comparative entre deux grandes familles d'explication par contrefactuels.

Amélioration des explications basées sur Anchor

Dans cette partie portant sur la génération d'explications adaptées aux données, nous proposons deux améliorations pour la méthode d'explication Anchor [124]. La première amélioration porte sur la technique utilisée pour discrétiser les données tabulaires. Dans la méthode originale, l'entropie, le décile et le quartile sont utilisés alors que nous proposons une approche plus adaptée nommé MDLP [44]. Cette amélioration permet de générer des explications qui ont à la fois une meilleure fidélité au modèle et une plus grande couverture. La seconde amélio-

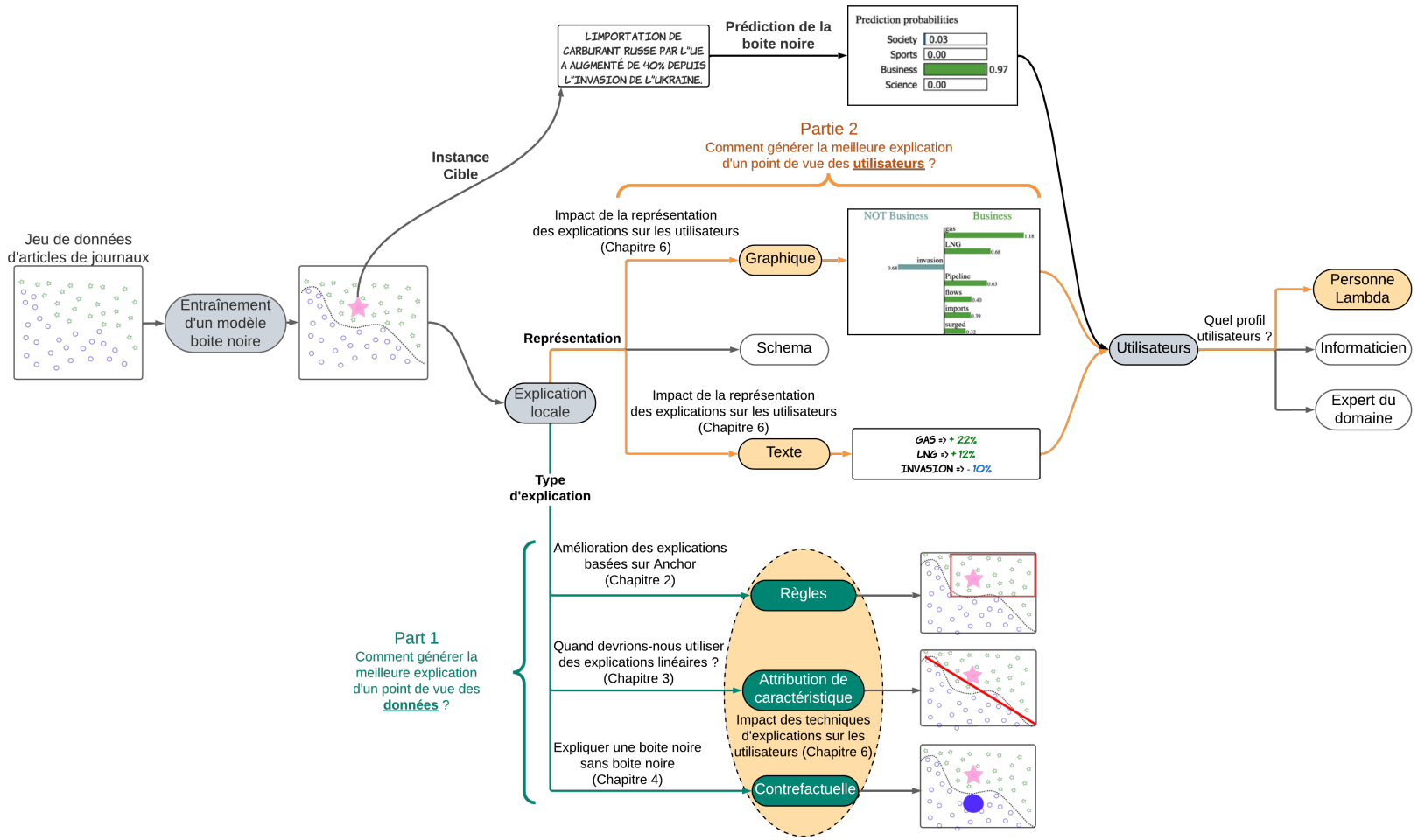


Figure B.5 – Diagramme décrivant les aspects de l'explicabilité étudiés dans cette thèse. La première partie, représentée par les cases vertes, se concentre principalement sur les différentes techniques d'explication. La seconde partie, représentée par les ovales orange, se concentre davantage sur les aspects liés aux utilisateurs.

ration concerne la version d'Anchor sur les données textuelles et permet une recherche plus large d'explication en se basant sur des conditions négatives [35]. Cette approche permet de générer des explications qui montrent que l'absence d'un mot fréquemment associé à un mot du document à expliquer modifierait la prédiction. Ce travail a conduit à la publication d'un article présenté dans le cadre de la conférence internationale sur la gestion de l'information et des connaissances (CIKM) 2022.

Quand devrions-nous utiliser des explications linéaires ?

Dans cette recherche, nous avons exploré la pertinence de l'utilisation d'une explication linéaire pour comprendre le fonctionnement des modèles de prédiction. La question centrale était de savoir si un modèle linéaire est toujours approprié pour approximer la frontière de décision d'un modèle à expliquer. Nous avons eu l'intuition qu'en fonction de la forme de cette frontière, un modèle linéaire pourrait ne pas fournir une approximation précise. Pour répondre à cette question, nous avons développé un outil, que nous appelons un oracle, capable de déterminer si une explication linéaire est adaptée pour une instance donnée.

L'oracle commence par localiser la frontière de décision la plus proche de l'instance à expliquer en utilisant une méthode appelée Growing Spheres [82] que nous avons améliorée. Ensuite, cet oracle analyse la distribution des données autour de cette frontière pour déterminer si une explication linéaire est appropriée. Pour ce faire, l'oracle utilise un test d'unimodalité, qui évalue si les données peuvent être regroupées en un ou plusieurs groupes distincts [139]. Si le test suggère que les données sont regroupables en plusieurs groupes, l'oracle conclut qu'une explication linéaire n'est pas adaptée. À l'inverse, en cas de distribution unimodale, l'oracle procède à un simple test de séparabilité linéaire. Si ce dernier est réussi, cela confirme que l'explication linéaire est appropriée pour l'instance considérée. En résumé, l'oracle offre une approche systématique pour évaluer la pertinence d'une explication linéaire en fonction de la structure de la frontière de décision et de la distribution des données environnantes.

Nous avons ensuite introduit deux nouvelles méthodes, que nous avons appelées APE (Explication Post-hoc Adaptée), tirant parti de cet oracle. Ces méthodes fournissent une explication linéaire lorsque celle-ci est indiquée comme adaptée par l'oracle, sinon elles renvoient une explication basée sur des règles. Les expériences que nous avons menées démontrent que le fait de toujours fournir une explication linéaire, sans évaluer son adaptabilité à la situation et au modèle à expliquer, entraîne une nette diminution de la fidélité de l'explication. Ce travail, dont les résultats ont été publiés lors de la conférence internationale CIKM en 2020, constitue

ainsi la première pierre pour le développement de méthodes d'explications adaptées à chaque situation.

Exploration du spectre des méthodes d'explication contrefactuel

Dans cette troisième étude, nous avons effectué une comparaison entre différentes méthodes d'explications contrefactuels appliquées sur les données textuelles. Nous avons commencé par différencier les méthodes existantes en fonction de l'approche utilisée pour perturber le document cible. Cette distinction nous a conduit à identifier deux grandes familles. D'une part, les approches qui perturbent le document à expliquer au niveau des mots [127, 98], par exemple, en les retirant, en les remplaçant, ou en ajoutant des mots. D'autre part, il existe des méthodes qui convertissent le document cible dans un espace latent. Ces espaces latents résultent de la conversion d'un texte ou d'un mot en un vecteur de haute dimension, généralement à l'aide de modèles de langues complexes. Cela permet d'effectuer des perturbations plus subtiles dans cet espace avant de revenir à l'espace du texte d'origine [111, 128, 96]. Nous avons proposé de qualifier les premières méthodes comme étant transparentes, tandis que les secondes sont considérées comme opaques.

Nous avons mené des expériences sur ces différentes méthodes et nos résultats montrent que les approches transparentes produisent des exemples contrefactuels aussi bons voire meilleurs que les méthodes opaques sur un ensemble de métriques, de tâches et de modèles. Ainsi, la question qui est soulevée au travers de ce travail est de déterminer si le recours à des méthodes complexes pour expliquer le fonctionnement d'un modèle complexe présente un réel intérêt.

Conclusion

Tout au long de cette partie, j'ai identifié deux pistes de recherche prometteuses pour des explorations futures. Premièrement, il est essentiel de discerner les conditions dans lesquelles une technique d'explication donnée s'adapte à son contexte. Cela s'appuie sur la trajectoire de recherche initiée par APE (cf. Chapitre 3). Deuxièmement, la communauté sur l'explicabilité doit étudier l'efficacité de l'introduction de couches de complexité supplémentaires dans les méthodes conçues pour expliquer le fonctionnement interne de modèles déjà complexes. Il est important de ne pas oublier que l'objectif de l'explicabilité est d'améliorer la confiance que l'on peut avoir dans les modèles d'apprentissage automatique, pas de rajouter de la complexité. Cela nécessite une investigation plus approfondie de la recherche présentée dans le Chapitre 4.

Les perspectives utilisateurs dans la génération d'explications

Dans cette première partie, nous avons exploré les techniques d'explication post-hoc, une famille de méthodes très courante qui a bénéficiée d'une attention significative au sein de la communauté XAI. Malgré la prolifération de ces méthodes d'explicabilité, les chercheurs ont pertinemment souligné que les méthodes XAI ne se concentrent pas suffisamment sur les utilisateurs cibles [38, 125]. En réponse à cette observation, la deuxième partie de cette thèse est consacrée à l'évaluation des techniques d'explication dans une perspective centrée sur l'utilisateur. Comme l'ont souligné Doshi-Velez et Kim ainsi que deux autres enquêtes XAI [1, 4], un faible nombre d'articles XAI justifient leurs nouvelles méthodes par des évaluations utilisateurs. Adadi et al. [1] ont constaté que dans un échantillon de 381 articles sur l'explicabilité, seuls 5% mettaient explicitement l'accent sur l'évaluation des méthodes proposées par des sujets humains. En d'autres termes, la recherche en XAI est occupée à produire des méthodes d'explication basées sur l'attribution de caractéristiques, des règles et des exemples pour les utilisateurs, sans évaluer les explications qui en résultent avec les utilisateurs finaux. Cette lacune dans la recherche implique que nous en savons peu sur la mesure dans laquelle les utilisateurs comprennent les explications des systèmes d'IA. De même, on ne sait toujours pas si la présence d'explications lors de l'utilisation de systèmes d'IA augmente ou non la confiance des utilisateurs.

Cette partie introduit un cadre méthodologique pour la réalisation d'études utilisateurs. L'objectif de cette méthodologie est d'établir un cadre solide pour étudier l'impact des techniques d'explication sur les utilisateurs. En outre, nous proposons un ensemble complet de mesures pour évaluer la confiance, la compréhension et la satisfaction perçues et comportementales des utilisateurs. Ce cadre méthodologique, ainsi que les mesures, seront ensuite appliqués dans le Chapitre 6 pour mener une étude utilisateurs. Cette étude vise à mesurer les avantages des différentes techniques d'explication et de leurs représentations (graphiques ou textuelles) sur la compréhension et la confiance des utilisateurs. Ces expériences sont conduites sur des utilisateurs inscrits sur une plateforme de crowdsourcing, permettant d'avoir un échantillon de participants représentatif.

Évaluation des méthodes d'explications centrée sur l'utilisateur

Dans ce chapitre, on donne un aperçu des études utilisateurs existantes qui ont été menées pour évaluer les techniques d'explication. Nous proposons ensuite en réponse au besoin en études utilisateurs et lignes directrices pour évaluer les techniques d'explication dans la XAI,

un cadre méthodologique. Cette méthodologie est conçue pour évaluer de manière exhaustive les interactions des utilisateurs avec les systèmes d'IA par le biais d'enquêtes, dans lesquelles les participants effectuent des tâches spécifiques. Cela permet d'obtenir des informations précieuses sur les perceptions et les comportements des participants lorsqu'ils interagissent avec des modèles d'IA. Pour garantir une approche structurée et la reproductibilité, nous avons organisé les enquêtes en trois phases distinctes : l'introduction, les tâches et les post-questionnaires. En outre, nous avons introduit une série de mesures pour quantifier l'impact des différentes variables. Ces mesures fournissent une image claire de la manière dont divers facteurs influencent le comportement, la confiance, la compréhension et la satisfaction de l'utilisateur lorsqu'il interagit avec des systèmes d'IA.

Impact des techniques et représentations sur les utilisateurs en XAI

Dans cette étude utilisateurs, nous avons comparé les explications à base d'attribution de caractéristiques, celles à bases de règles et les exemples contrefactuels. Pour chacune de ces méthodes, nous avons construit deux types de représentation, soit graphique ou textuel. Nous avons déployé ces questionnaires sur deux domaines, la prédiction d'obésité en rapport à la santé et la prédiction de récidive pour le domaine judiciaire. Chaque participant a eu accès à un mode d'explication et une représentation. Nous avons ensuite mesuré les effets de chaque type d'explication et chaque représentation sur la confiance et la compréhension du participant dans le système d'IA. Nous avons établi différentes métriques en se basant à la fois sur la perception de l'utilisateur et son comportement.

Nos résultats ont montré qu'à la fois le type d'explication et sa représentation ont un impact sur l'utilisateur. Nous avons remarqué par exemple que les participants ont plus tendance à faire confiance à une explication si sa représentation est graphique. De la même manière, nous avons remarqué que les explications à base de règles sont celles qui amènent la meilleure compréhension.

Conclusion

Il est surprenant de constater que, malgré l'intérêt croissant pour l'explicabilité, aucun travail antérieur n'a comparé l'impact des différentes méthodes d'explication sur des rôles d'utilisateurs spécifiques (par exemple, expert du domaine, développeur). En raison du manque de comparaison des techniques d'explication, les utilisateurs de l'explicabilité sont actuellement incapables d'indiquer pourquoi ils choisissent un type d'explication au détriment d'un autre.

Nous soutenons que la préférence pour un type d'explication par rapport à un autre devrait être motivée par des critères et des situations plutôt que par des raisons pratiques. Ainsi, dans cette thèse, nous avons cherché à determiner l'impact des méthodes d'explication les plus courantes et leurs représentations sur les utilisateurs. En particulier sur la confiance et la compréhension lorsqu'un utilisateur non expert interagit avec un système d'IA. Il est important de noter que nos résultats présentent des variations entre les domaines évalués (santé ou judiciaire). Cela souligne le potentiel et la nécessité d'études futures pour prendre en compte les profils des utilisateurs, les types de données et l'influence des domaines sur les résultats.

REMERCIEMENTS

Presque quatre années se sont écoulées depuis mes premières intentions jusqu'à ce text, fruit de multiples collaborations aussi riches qu'inspirantes. Je souhaite exprimer toute ma gratitude envers ceux qui m'ont accompagné et encouragé dans la poursuite de ce travail et qui ont ainsi contribué à enrichir ma recherche. Ce moment marque l'apogée d'une période intense qui va profondément influencé mon devenir, tant sur le plan professionnel que personnel.

A l'instar des modèles étudiés dans cette thèse, je ne suis pas parfait et sollicite donc votre indulgence si par inadvertance j'omettais de vous citer. D'avance, acceptez mes excuses et contactez-moi afin que nous puissions travailler à réparer l'oubli.

Je tiens à exprimer mes premiers remerciements à mes encadrants, Christine Largouët et Luis Galarraga. Nos réunions ont toujours été empreintes d'une atmosphère propicice à une recherche exigeante n'ont jamais empêchées les digressions qui permettent d'attendre nos objectifs avec envie. Tout d'abord, Christine, qui a su me cadrer et m'enseigner les rouages du métier de chercheur. Ta capacité à remettre en question l'intérêt de chacune de mes idées, ainsi que ta franchise et ta rigueur dans les moments cruciaux m'ont propulsé là où j'en suis aujourd'hui. Luis, ensuite, un immense merci pour le soutien sans faille que tu m'as apporté au cours de ces années. Tu m'as accordé ta confiance dès mon premier stage, puis m'as offert la chance de réaliser cette thèse avec vous deux. Merci de m'avoir donné le sens de l'organisation, tant à grande échelle que dans les détails. Tu m'as appris que la communication, à l'écrit comme à l'oral, est un vecteur essentiel de la recherche. Tes retours sans cesse plus complets m'ont fait progresser plus rapidement que je ne l'aurais imaginé. Par ton enthousiasme, tu m'as fais aimé la recherche ! Au-delà d'être un encadrant hors pair, ton amitié n'a pas de prix.

I would like to express my gratitude to the reviewers, Marie-Jeanne Lesot and Andrea Passerini, as well as the members of the jury, Elisa Fromont, Pierre Marquis, Katrien Verbert, and Niels van Berkel, for the honor they have accorded to me by taking a second look at this work.

A special acknowledgment goes to Niels, for his kindness and hospitality when I approached him to collaborate on a subject in which I was a foreigner. My time in Aalborg was truly enjoyable, thanks to him and, of course, all the other members of the Human-Centered Computer group. To begin with, I want to express my gratitude to my co-writer, Joël. I value the late nights spent working together at the office and the enriching conversations we shared as we delved into each other's cultures and scientific backgrounds. Next, my coworkers, Naja and Sander, I look forward to continuing our drawing games for a long time! Naja, thank you for your joy and assistance in helping me integrate into the team; your Christmas party will be a fond memory. Sander, it was a pleasure to watch the World Cup with you and explore Aalborg and its pub together.

De même, ma gratitude s'adresse à l'équipe LACODAM, qui depuis 2018 me transmet sa passion des sciences et m'encourage dans mes travaux. Mes remerciements vont en premier lieu aux doctorants : les nouveaux (Olivier, Pierre, Lucie, Julie, Ambre et Isseïnie) et les anciens (Colin, Yi-Chang, Maël, Johanne, Camille, Gregory et Antonin) ainsi que ceux qui ont partagé mon bureau : Simon, Lenaïg, Heng, Gwladys et Paul. Merci pour ces discussions toutes plus originales les unes que les autres durant les pauses-déjeuner ou café, ainsi que les soirées jeux que nous avons pu partager. Je tiens ensuite à remercier les permanents, en particulier Laurence et Véro, avec qui j'ai partagé un bureau lors de mon premier stage, puis Alexandre et Romaric, avec qui j'ai enseigné dans de bonnes conditions.

Le temps de la thèse, c'est aussi la vie qui se poursuit avec ses moments mémorables. Comme le mariage de Jeanne et Géraud. Merci, mon ami, de m'avoir constamment poussé à me poser les bonnes questions et à m'avoir aidé à prendre les bonnes décisions. Ces qualités, essentielles à la recherche, ne se transposent pas aisément dans la vie privée. Ensuite, un remerciement particulier à mes amis de longue date, Guénolé, Hugues, Nolan et Josselin et leur bonne humeur salvatrice pendant la période de confinement où je démarrais la thèse. Je tiens également à remercier mes amis de la fac : Mani, Romiche, Adrien, Simon, Elise, Thomas, et tous les autres. La MIAGE, c'est le partage, et je suis convaincu que toutes les expériences que nous avons vécues ensemble resteront gravées dans nos mémoires. En particulier, Thomas et Pauline, vous avez été très présents ces dernières années, votre sens de l'humour et votre hospitalité m'ont permis de déconnecter si souvent si rapidement. Merci beaucoup !

Il y a tant de personnes que je souhaite remercier, notamment l'équipe des Champs Libres : Kenza, Juliette et Carolane, (Maureen tu fais partie de l'équipe maintenant). Mais aussi celle de Sherbrooke dont les retrouvailles me semblent toujours improbables. Malgré la distance

de plus de 5000 kilomètres et le temps écoulé depuis notre rencontre, nous avons réussi à maintenir le contact, et cela me semble incroyable ! Merci à Lolo, Jojo, Nono, Vico, Nana, Juju, Lulu et Ali. Évidemment, je ne peux pas oublier mes chers camarades de master, en particulier Chaffin et Taha. Nos repas hebdomadaires ont été de véritables moments d'échanges sur nos expériences respectives et de décompression pendant des périodes parfois difficiles. Chaffin, j'ai adoré collaborer avec toi, et j'espère que nous aurons l'occasion de poursuivre.

Je passe maintenant à ceux qui sont là depuis le tout tout début, mes parents ! Ma mère, naturellement, qui a toujours su voir le positif en tout, y compris les enseignements à tirer des soumissions rejetées. Ton soutien infaillible a été une lumière dans ce marathon qu'est le doctorat. Mon père, qui s'est toujours engagé dans des discussions stimulantes et a toujours voulu challenger mes idées, m'a enseigné à ne pas baisser les bras et à lutter pour obtenir ce que je veux. Je ne peux que vous remercier pour ce que vous avez fait, et je suis fier d'être votre fils.

Je tiens également à exprimer toute ma reconnaissance à mon grand petit frère, Teddy, qui m'a toujours inspiré et a été le premier à me montrer le chemin de la thèse. Je me souviens de ce moment en amphi où j'ai décidé de me lancer dans la recherche. J'ai pensé à toi, ta vie, tes histoires, et je me suis dis que je pouvais suivre la voie de mon super grand frère. Grâce à toi, j'ai eu la chance de rencontrer Amelie tout d'abord. Ton intégration dans la famille, ta franchise et ton intelligence m'ont permis d'en apprendre davantage sur ce que je suis et sur cette famille qui compte beaucoup pour moi. En parlant de famille, elle s'est agrandie dernièrement, je te souhaite la bienvenue, Ismaël. J'ai hâte de te voir grandir et de découvrir ce que tu deviendras. Vu l'amour qui t'entoure, je suis sûr que tu seras un véritable Ismamour.

Je ne serai pas complet si je ne remerciais pas mon grand grand frère JC. Toi qui m'as doté d'un premier ordinateur et donné une première attirance pour l'informatique, et qui t'es si bien occupé de moi toutes ces années. J'ai eu la chance de voir ta famille s'agrandir, d'abord Nanou, dont la patience et la gentillesse sont un exemple pour moi. Ensuite, Arthur, mon filleul, aussi gentil et attentionné qu'actif et intelligent ; sa petite soeur, Alice, pas très lourde, mais avec un grand coeur et un sacré caractère.

Enfin, je tiens à exprimer ma gratitude envers mes grands-parents qui ont su par leur générosité, contribuer à fonder une famille unie et solide.

Pauline, tu t'es dévouée pour me soutenir et créer les conditions optimales pour que je puisse conclure cette thèse, et je suis incapable de trouver les mots justes pour exprimer ma gratitude. Ta joie de vivre quotidienne, ton expressivité et ta motivation sont des exemples pour

moi vers lesquelles j'aspire à tendre. S'engager avec quelqu'un en pleine thèse n'est pas chose facile, et je suis déterminé à tout mettre en oeuvre pour te montrer que ton investissement en valait la peine. L'avenir nous réserve de belles surprises, et j'ai une certitude : partager ces moments avec toi ne pourra qu'apporter du bonheur. J'attends avec impatience de découvrir la direction que prendra notre projet, mais une chose est sûre, nous serons heureux de vivre ces moments ensemble.

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Explicabilité des modèles d'apprentissage automatique : De l'adaptabilité des données à la perception de l'utilisateur

Mot clés : Explicabilité ; Interprétabilité ; Interaction Homme-Machine

Résumé : Cette thèse se concentre sur la génération d'explications locales pour les modèles de machine learning déjà déployés, en recherchant les conditions optimales pour des explications pertinentes. L'objectif principal est de développer des méthodes produisant des explications à la fois fidèles au modèle sous-jacent et compréhensibles par les utilisateurs qui les reçoivent. La thèse est divisée en deux parties. Dans la première, on améliore une méthode d'explication basée sur des règles. On introduit ensuite une approche pour évaluer l'adéquation des explications linéaires pour approximer un modèle à expliquer. Enfin, cette partie présente une expérimentation comparative entre deux familles de méthodes d'explication contrefactuelles, dans le but d'analyser les avantages de l'une par rapport à l'autre. La deuxième partie se concentre sur des expériences utilisateurs évaluant l'impact de trois méthodes d'explication et de deux représentations différentes. Ces expériences mesurent la perception en termes de compréhension et de confiance des utilisateurs en fonction des explications et de leurs représentations. L'ensemble de ces travaux contribue à une meilleure compréhension de la génération d'explications, avec des implications potentielles pour l'amélioration de la transparence, de la confiance et de l'utilisabilité des systèmes d'IA déployés.

Title: Explainability for Machine Learning Models: From Data Adaptability to User Perception

Keywords: Explainability, Interpretability, Human-Computer Interaction

Abstract: This thesis explores the generation of local explanations for already deployed machine learning models, aiming to identify optimal conditions for producing meaningful explanations. The primary goal is to develop methods for generating explanations faithful to the underlying model and comprehensible to the users. The thesis is divided into two parts. The first enhances a widely used rule-based explanation method. It then introduces a novel approach for evaluating the suitability of linear explanations to approximate a model. Additionally, it conducts a comparative experiment between two families of counterfactual explanation methods to analyze the advantages of one over the other. The second part focuses on user experiments to assess the impact of three explanation methods and two distinct representations. These experiments measure how users perceive their interaction with the model in terms of understanding and trust, depending on the explanations and representations. This research contributes to a better explanation generation, with potential implications for enhancing the transparency, trustworthiness, and usability of deployed AI systems.