Università degli Studi di Padova

Dipartimento di Diritto Pubblico, Internazionale e Comunitario (DiPIC)
Dipartimento di Matematica

Corso di Laurea di Diritto e Tecnologia
a.a 2022/2023

Recruitment systems nowadays: how XAI can improve trust

Tesi di laurea triennale

Relatore:

Roberto Confalonieri

Laureando:

Cesare Gortana

Index

CHAPTER 1: INTRODUCTION

SECTION 1.1 MOTIVATION

In recent years the use of Artificial Intelligence (AI) has exploded. While among the new generations AI tools such as Chat-GPT, Bard... are widely accepted, scholars wonder about the guidelines that these tools should respect and which systems they should adopt for full respect of competitive practices and privacy. More and more companies implement these systems "At 99% of Fortune 500 Companies, job applications are first evaluated by an applicant tracking system instead of a human being[1]". The IBM report "Global AI Adoption Index 2022" reports that 34% of companies say they use Artificial Intelligence in their business. According to respondents, 42% of them say they want to explore the world of AI systems. Moreover, Deloitte report shows 33% employees desire that their workplace will become integrated by AI in the coming future." From the first statistics it is clear, therefore, that these technologies are permeating the social-working fabric of companies. However, caution is needed. While it is undeniable that these tools bring countless advantages, (e.g., they add great value to data-intensive and time-consuming processes), on the other hand they expose users to a technology whose decision-making process is not always transparent and clear. This is why they are often associated with the term "black box" systems to indicate the inscrutable character of its content. In a fast-paced and ever-changing world where information that is new today will be old tomorrow, it is necessary to provide a set-up of suitable rules and legal provisions to ensure respect for the universal rights of the individual, including the principles of privacy and fair treatment. The massive use, available without limitations of time or space, very often free, of this kind of technology requires serious and immediate reflection on the nature, risks, impacts that these technologies have on society, in order to establish an appropriate regulatory framework.

SECTION 1.2 OBJECTIVE

We are beginning to see a large number of academic references, scientific studies and authoritative opinions on this subject, but we are far from guaranteeing precise and timely documentation in this regard. In this thesis we analyze the impact of Artificial Intelligence systems in the context of corporate recruitment. The hope of this thesis is that academic and scientific research on the proposed topic can progress at the same rhythm and prevent the creation of critical situations. Moreover, we wish that the proposed reflections can be of great help to companies that want to use Artificial Intelligence systems in the context of corporate recruitment.

[1] James Hu, ‘Report: 98% of Fortune 500 Companies Use ATS’, *Jobscan*, 2018 <https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/>.

Section 1.3 Approach followed

The study of Artificial Intelligence systems should be carried out on the basis of two analyses. The first, which is more technical, serves to understand in detail the origins and composition of the system. In order to improve it, it is first necessary to know closely what elements characterize it and what computer implications derive from the use of these. To accomplish this, it is necessary to construct a correct basic taxonomy to which the reader can refer. A correct investigation of an Artificial Intelligence system then provides a second analysis of a more legal nature. The use of these technologies has a strong impact on everyday society. Reflecting what legal requirements these systems must meet and, where possible, identifying minimum guidelines to be followed are minimum targets.

Section 1.4 Structure of the thesis

The thesis is divided into five chapters. In chapter II we present the necessary technical background and notions that the reader needs to master in order to fully understand the technical analysis of the models. Chapter III proposes an overview on the legal context that concerns the AI systems. Chapter IV analyzes different AI recommendation systems that are used in recruitment. Finally, Chapter V presents the remarks and conclusion of the thesis.

CHAPTER 2: BACKGROUND

SECTION 2.1 AI AND BLACK BOX MODELS

“Some expect efficient automation via AI applications to increase overall productivity, while others fear that AI systems will completely replace human roles due to their high efficiency"[2]. Before we ask ourselves what the impact and consequences of using an Artificial Intelligence system in the recruitment phase are, it is necessary to provide a common interpretation of Artificial Intelligence. Unfortunately, the literature is still debating the definition to be adopted. Among the various hypotheses presented, this thesis will refer to Artificial Intelligence such as: "The frontier of computational advancements that references human Intelligence in addressing ever more complex decision-making problems[3]. Thus, AI refers to machines performing a spectrum of cognitive tasks and intelligent behavior patterns commonly associated with human intelligence[4]. AI comprises a variety of methods, such as machine learning (ML) and rule-based symbolic logic, which differ in their complexity and suitability for different tasks[5]".

This technology first appeared in 1956 during a summer seminar at Dartmouth College in Hanover, New Hampshire. Following an initial acceleration in the studying this subject, a sharp braking followed. Some of the reasons for this slowdown are to be found in the computational limit of the machines used to implement these Artificial Intelligence systems. Around the 2000s, the world witnessed a rapid recovery in the implementation of these systems due to technical improvements that could increase the performance of computer processors. However, only in recent years has the use of this technology pervaded the daily lives of web users, who are today able to interface directly with it. In Italy alone, the Artificial Intelligence market grew by 22 % in 2022 and analysts believe that the outlook is 20% year on year for the next five years. It is clear that this type of technology will become increasingly present in the lives of technicians and ordinary people and will be used in new areas[6].

From the most recent studies we note a growing use of Artificial Intelligence systems in the working environment, in particular to automate some analysis and evaluation processes that until now were a

[2] Michael Chui, James Manyika, and Mehdi Miremadi, ‘Where Machines Could Replace Humans—and Where They Can’t (Yet)’, 2016.

[3] Nicholas Berente and others, ‘Managing Artificial Intelligence’, *MIS Quarterly*, 45 (2021), 1433–50 <https://doi.org/10.25300/MISQ/2021/16274>.

[4] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education Limited. (Pearson, 2016).

[5] William B. Rouse, ‘AI as Systems Engineering Augmented Intelligence for Systems Engineers’, *INSIGHT*, 23.1 (2020), 52–54 <https://doi.org/10.1002/inst.12286>.

[6] ‘Intelligenza Artificiale, Nel 2022 Crescita Del 22%: 700 Milioni Nel 2025 - Il Sole 24 ORE’ <https://www.ilsole24ore.com/art/intelligenza-artificiale-2022-crescita-22percento-700-milioni-2025-AEzvvYkC?refresh_ce=1> .

human prerogative. This step follows a trend already known in recent years: the amount of data that is created every day, every second is now impressive. Human capabilities are no longer able to handle similar amounts of information and the use of a data-driven system that processes and returns to the human user a value from which to start significantly reduces the time wasted in data analysis. Moreover, the implementation of Artificial Intelligence systems increasingly structured allows to manage non-linear data.

While the use of these systems can improve the performance of human being, helping them to carry out a series of tasks independently, on the other hand it raises technical-ethical issues. First, we observe that every Artificial Intelligence systems' decision is intrinsically opaque. The level of opacity depends on many factors, for example the techniques used to process the data and the type of model. The simplest models, such as linear regression, logistic regression and decision tree perform simpler tasks and the choices offered are more understandable for the human subject with which they interface. For this reason, we refer to these models as white-box models, towards which "a person may comprehend simpler machine learning models by glancing at the summary of parameters of the model without the need for an external model to provide an explanation". Other models produce more complex decisions without revealing any information about its decision-making processes. In these cases maintaining an acceptable level of trust with the model is more complicated because of the inscrutable composition of black-box models it is very difficult for them to provide this kind of information. "Given an input, a black-box returns the result of a decision task (classification, prediction, recommendation, etc.), but it does not reveal sufficient details about its internal behavior, resulting in an opaque decision model. For this reason, explainability in machine learning is formulated as the problem of finding an interpretable model that approximates the black-box model as much as possible, typically seeking high fidelity." Like most of these models, they are developed with a dual objective: to improve performance in personnel selection by reducing discriminatory biases and the opacity of decision-making; and to accelerate recruitment processes, automating the phases that may not depend entirely on the human being. The interpretations given by a person are characterized by a, even low degree of uncertainty, equivocality and complexity "based on the amount of information[7]." Since the right amount of information for decision making is an optimal condition, uncertainty refers to a condition with too little information to logically decide, while complexity refers to a condition with too much information to identify the necessary information". In this case, the incidence of equivocality and uncertainty were observed through two motivational indicators: anxiety and trust respectively. It is possible to define anxiety such as "the fear, apprehension and hope that people experience when considering using or using a new

[7] Herbert A. Simon, *Models of Man; Social and Rational*, Models of Man; Social and Rational (Oxford, England: Wiley, 1957), pp. xiv, 287.

technology[8]. Trust instead, is "the attitude that an agent will help to achieve an individual's goal in a situation of uncertainty and vulnerability"[9]. Increasing the trust decision-makers have with an AI system is a crucial point. An AI system is able to provide significant help when it can operate independently of the human subject, sharing the final objectives and being able to operate so that the decision-makers can maximize their performance. An acceptable level of this indicator and a correct implementation of the attributes of fairness and transparency should allow decision-makers to follow the decisions chosen by the algorithm even when the performance of the system is not perfectly reliable, since the human subject could verify the presence of discriminatory biases and correct them[10]. Using an AI system that is also explainable allows a more effective understanding of the decisions taken and the reasons behind them, increasing trustworthiness. Not only that, the explainability factor allows a greater responsibility of the decision-maker, simplifying the process of confirming whether the model functions fairly or ethically by visualizing the feature relations affecting a given result[11] and reducing the technical skills required to understand the effect of an AI recruitment system.

In order to counteract the level of opacity of these systems, they must be designed taking into account three fundamental concepts, underlying the reliability of the models: fairness, accountability, and transparency. The first refers to procedural fairness, which consists of different social statuses and faiths. It is worth pointing out that an algorithm, although it can avoid proposing discriminatory models directly, can be strongly influenced by the discriminatory biases of the raw information it processes. Unfortunately, this possibility frequently occurs in the use of AI because intrinsically the data inherent in human beings carry with them a minimum level of prejudice. An interesting example of this critical situation is the COMPAS case. It is an algorithm used by a federal court in the United States to grant a reduction and/or release into freedom of some detainees present in federal state prisons. This algorithm was established by examining numerous cases of national judgments and on the basis of non-trivial information extracted from them, had to provide an assessment of the prisoner about his possibility of repeating the crime as soon as he was released. To do this, the algorithm took into account a number of key parameters such as: gender, age, ethnicity, place of residence, salary. Analyzing the final assessments expressed by the algorithm, we

[8] Matthew L Meuter and others, 'The Influence of Technology Anxiety on Consumer Use and Experiences with Self-Service Technologies', Journal of Business Research, Strategy in e-marketing, 56.11 (2003), 899–906 <https://doi.org/10.1016/S0148-2963(01)00276-4>.

[9] Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

[10] Andreas Holzinger and others, 'Information Fusion as an Integrative Cross-Cutting Enabler to Achieve Robust, Explainable, and Trustworthy Medical Artificial Intelligence', Information Fusion, 79 (2022), 263–78 <https://doi.org/10.1016/j.inffus.2021.10.007>.

[11] Dena F. Mujtaba and Nihar R. Mahapatra, 'Ethical Considerations in AI-Based Recruitment', in *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 2019, pp. 1–7 <https://doi.org/10.1109/ISTAS48451.2019.8937920>.

note a clear deviation in the judgment between the dark-skinned detainees and the white detainees. The latter were treated favourably because of their ethnicity or a stable financial situation, obtaining a very low assessment of the future recurrence of the crime. A black person, on the contrary, received a very high assessment when compared to the seriousness of the crime committed or the possibility of committing it again due to factors sometimes unrelated to his real will to commit the crime again. In relation to the COMPAS case, therefore, the lack of fairness is evident because the decision-making approach that was implemented was strongly influenced by different discriminatory factors, such as race, gender and religion. A second key aspect is transparency. This attribute helps human decision-makers better understand the use and impact of the algorithm in model presentation. Moreover, it is clear that improved transparency also has a positive impact on the data controller who has suffered, as he can better understand the decision suggested by the AI system, what characteristics were taken into account for the decision and what impact they had. Finally, transparency helps to build fairness and accountability[12], highlighting the presence of unfair decisions and allowing them to be corrected[13]. The third fundamental attribute that can be distinguished is accountability, which is defined like "the responsibility of humans to ensure that their work upholds the common good, such as safety and privacy concerns[14]." You can distinguish two main features that make up accountability: the controllability and the openness. "Controllability refers to the responsibility of the human decision-maker to audit and modify the irrational configuration of the AI decision-maker (Shin, 2021). On the other hand, openness allows non-experts to easily access the decision basis and understand how and why certain decisions were made[15] ".

SECTION 2.2: MACHINE LEARNING

Artificial Intelligence systems that we are going to analyze are machine learning models. In order to understand them, we have to understand the nature of machine learning (ML). We will refer to this subjects like an application of AI that enables systems to learn and improve from experience without being explicitly programmed . It deals with the development of algorithms and techniques aimed at machine learning through computational statistics and mathematical optimization.

[12] Ashraf Abdul and others, ‘Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda’, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18 (New York, NY, USA: Association for Computing Machinery, 2018), pp. 1–18 <https://doi.org/10.1145/3173574.3174156>.

[13] Alejandro Barredo Arrieta and others, ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’, *Information Fusion*, 58 (2020), 82–115 <https://doi.org/10.1016/j.inffus.2019.12.012>.

[14] Nicholas Diakopoulos, ‘Accountability in Algorithmic Decision Making’, *Communications of the ACM*, 59.2 (2016), 56–62 <https://doi.org/10.1145/2844110>.

[15] Alex John London, ‘Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability’, *Hastings Center Report*, 49.1 (2019), 15–21 <https://doi.org/10.1002/hast.973>.

A common definition of a ML mode is the following one: “A ML model is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by p; improves with experience E”[16]. In order to do that, the algorithm is trained over a large amount of examples and assign a predictive output to each one. The objective of a ML model is to maximize or minimize a goal that could represents a classification task, regression task or many others. Models, like linear regression, logistic regression and decision tree can be very accurate in case they process linear data. In cases of non-linear data, they are less accurate and lose effectiveness and more complex models need to be used. In this thesis we will analyze the use of Artificial Intelligence systems in recruitment.

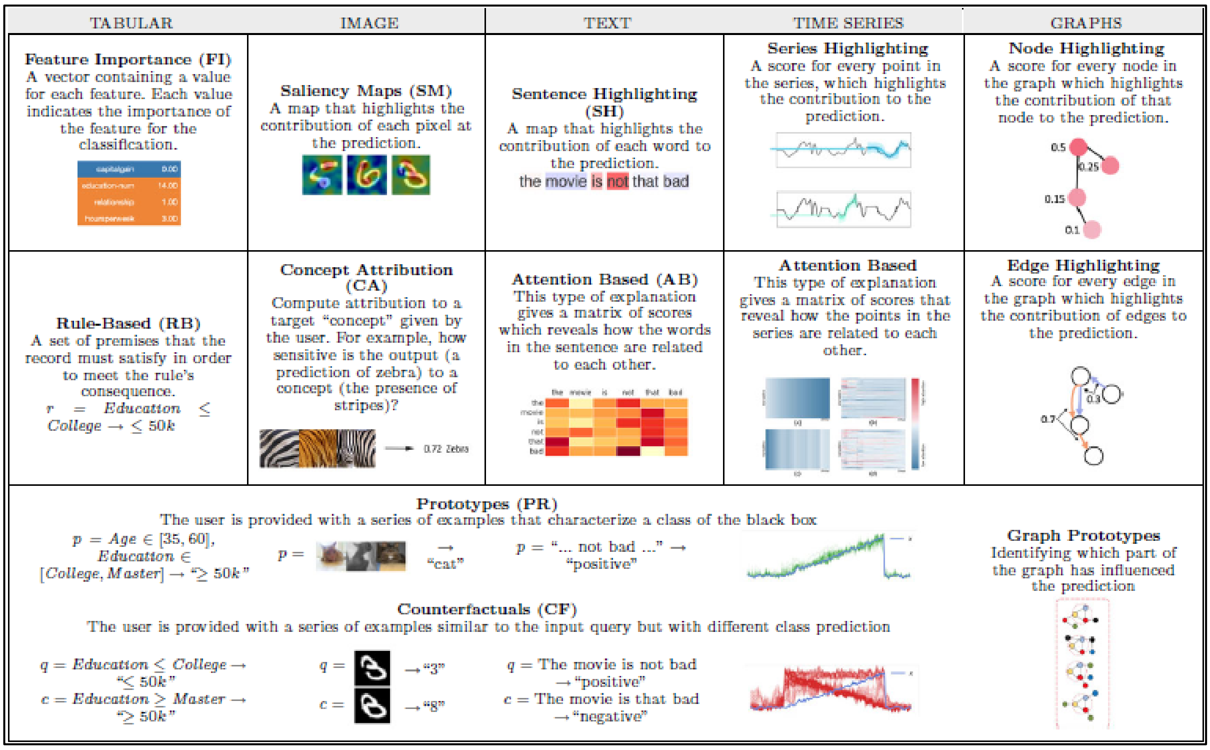
We see that an increasing number of companies are adopting these technologies as they can automate a number of processes, for instance, in recruitment ML can be used for screening resumes and shortlists the best candidates. This will help accelerate the recruitment process and ensure that the best candidates are selected for the job; improve employee retention: ML can be used to identify employees at risk of leaving the company. This information can then be used to implement strategies to improve employee retention; improve performance management: ML can be used to collect employee performance data. This data can then be used to identify areas where employees need improvement; identify training and development needs: ML can be used to identify employees who would benefit from training and development. This information can then be used to design training and development programs tailored to the needs of employees; predict future trends: ML can be used to analyze data and identify trends. This information can then be used to predict future trends in the workforce.

Section 2.3 XAI Approaches

In this section, we propose a survey on most common approaches to explain a black-box model. In particular, we focus on explanations and explanation methods acting on the main used data types: tabular data, images, text, time series and graphs[17].To achieve this, we refer to the analysis proposed in the book "Data Mining and Knowledge Discovery[18]. In particular, we observe that each data type is associated with different types of explanation, as shown in the table below. Columns header identify different data types and rows header distinguish different types of explanations. First, the reader should identify through the column header the data type of her problem setting. After that, the

[16] Tom M. Mitchell, *Machine Learning*, McGraw-Hill Series in Computer Science (New York: McGraw-Hill, 1997).
[17] Riccardo Guidotti and others, ‘A Survey of Methods for Explaining Black Box Models’, *ACM Computing Surveys*, 51.5 (2018), 93:1-93:42 <https://doi.org/10.1145/3236009>; Amina Adadi and Mohammed Berrada, ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, PP (2018), 1–1 <https://doi.org/10.1109/ACCESS.2018.2870052>; Tim Miller, ‘Explanation in Artificial Intelligence: Insights from the Social Sciences’, *Artificial Intelligence*, 267 (2017) <https://doi.org/10.1016/j.artint.2018.07.007>.
[18] Francesco Bodria and others, ‘Benchmarking and Survey of Explanation Methods for Black Box Models’, *Data Mining and Knowledge Discovery*, 37.5 (2023), 1719–78 <https://doi.org/10.1007/s10618-023-00933-9>.

reader should choose one of the proposed type of explanation that are described in the rows. A brief description is associated to each type. For instance, if we are interested in images, we should look to the second column. Here we can find saliency maps and concept attribution as image-specific explanation types.



TABULAR	IMAGE	TEXT	TIME SERIES	GRAPHS
Feature Importance (FI) A vector containing a value for each feature. Each value indicates the importance of the feature for the classification. capitalgain 0.00 education-num 14.00 relationship 1.00 hoursperweek 3.00	**Saliency Maps (SM)** A map that highlights the contribution of each pixel at the prediction.	**Sentence Highlighting (SH)** A map that highlights the contribution of each word to the prediction. the movie is not that bad	**Series Highlighting** A score for every point in the series, which highlights the contribution to the prediction.	**Node Highlighting** A score for every node in the graph which highlights the contribution of that node to the prediction. 0.5 0.25 0.15 0.1
Rule-Based (RB) A set of premises that the record must satisfy in order to meet the rule's consequence. $r = Education \leq College \rightarrow \leq 50k$	**Concept Attribution (CA)** Compute attribution to a target "concept" given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)? 0.72 Zebra	**Attention Based (AB)** This type of explanation gives a matrix of scores which reveals how the words in the sentence are related to each other. the movie is not that bad	**Attention Based** This type of explanation gives a matrix of scores that reveal how the points in the series are related to each other.	**Edge Highlighting** A score for every edge in the graph which highlights the contribution of edges to the prediction. 0.3 0.7
Prototypes (PR) The user is provided with a series of examples that characterize a class of the black box $p = Age \in [35, 60], Education \in [College, Master] \rightarrow$ "$\geq 50k$" $p =$ → "cat" $p =$ "... not bad ..." → "positive" **Counterfactuals (CF)** The user is provided with a series of examples similar to the input query but with different class prediction $q = Education \leq College \rightarrow$ "$\leq 50k$" $c = Education \geq Master \rightarrow$ "$\geq 50k$" $q =$ → "3" $c =$ → "8" $q =$ The movie is not bad → "positive" $c =$ The movie is that bad → "negative"				**Graph Prototypes** Identifying which part of the graph has influenced the prediction

Fig. 1 Explanation-based taxonomy divided for different data types
Source: Data Mining and Knowledge Discovery (2023)

Secondly, we observe the classification of explanation methods. A first division distinguishes: explainable by design methods and black-box explanation methods. The first category refers to methods that return a decision and the reasons for the decision are directly accessible because the model is transparent. Within this category, we can distinguish: global methods, where explanation methods aim at explaining the overall logic of a black-box model and the explanation returned is a global, complete explanation valid for any instance; and local methods, which aim at explaining the reasons for the decision of a black-box model for a specific instance. Similarly, in black-box explanation methods category, that provides explanations for a black-box model, there are two sub-categories: model-agnostic methods which can be used to interpret any type of black-box model and model-specific methods that can be used to interpret only a specific type of black-box model.

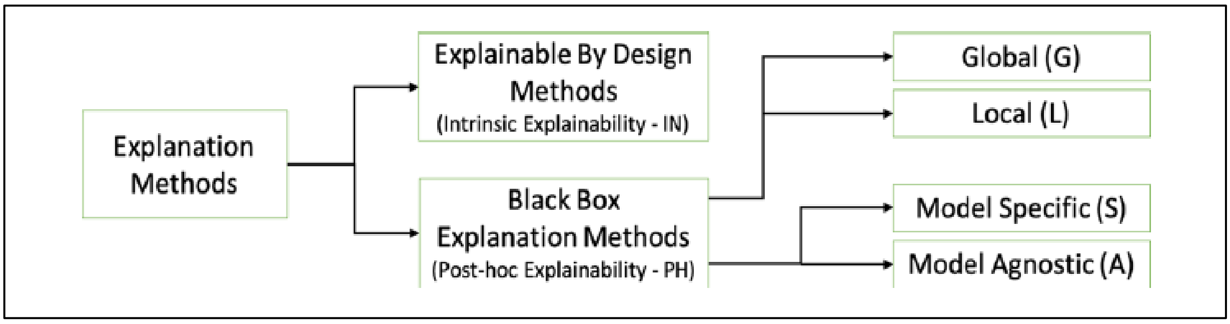

Explanation Methods
Explainable By Design Methods
(Intrinsic Explainability - IN)
Black Box Explanation Methods
(Post-hoc Explainability - PH)
Global (G)
Local (L)
Model Specific (S)
Model Agnostic (A)

Fig. 2 Classification of explanation methods
Source: Data Mining and Knowledge Discovery (2023)

Below, we briefly present some of these methods, in particular the decision trees, SHAP-method and Anchors method. All three methods can be used to explain a black-box model. The first two methods represent the methods implemented in the AI systems for recruitment that we will analyze in the chapter IV, while we selected the third one because it is one post-hoc method able to provide exploit rules explaining cases of a ML model.

Decision Trees are white box-symbolic models and they could be used as post-hoc explanation methods of black box models. The first prototypes were theorized in 1957 by two well-known researchers in the field of Artificial Intelligence, Allen Newell and Herbert Simon through their theory of physical symbol system hypothesis. They suggested that in principle, processing structures of symbols is sufficient to produce Artificial Intelligence in a digital computer. The same process of symbolic manipulations can shaper human Intelligence. For these reasons, classification decisions of Decision Trees models are easily interpretable and simple to understand by humans. Moreover, Decisions Trees can be visualized directly by the user in order to improve his comprehension of the model. They are composed as a “directed acyclic graph consisting of a set of split nodes, usually depicted as rectangles, and a set of leaves, usually depicted as ovals. Each split node in a decision tree has an associated logical test based on the features in the domain. When classifying an instance or example, the role of a split node is to assign the example to one of the outgoing branches of the node. Split nodes may have several branches depending on whether the logical test is over binary, nominal, or real values attributes. The decision as to which branch is selected for an example is determined by the logical test of the node. The way in which split nodes are selected amounts to optimising a reward function, the definition of which can vary depending on the type of induction algorithm used[19].”

[19] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, Fermín Moscoso del Prado Martín, *Using ontologies to enhance human understandability of global post-hoc explanations of black-box models*, “Artificial Intelligence”, 2021.

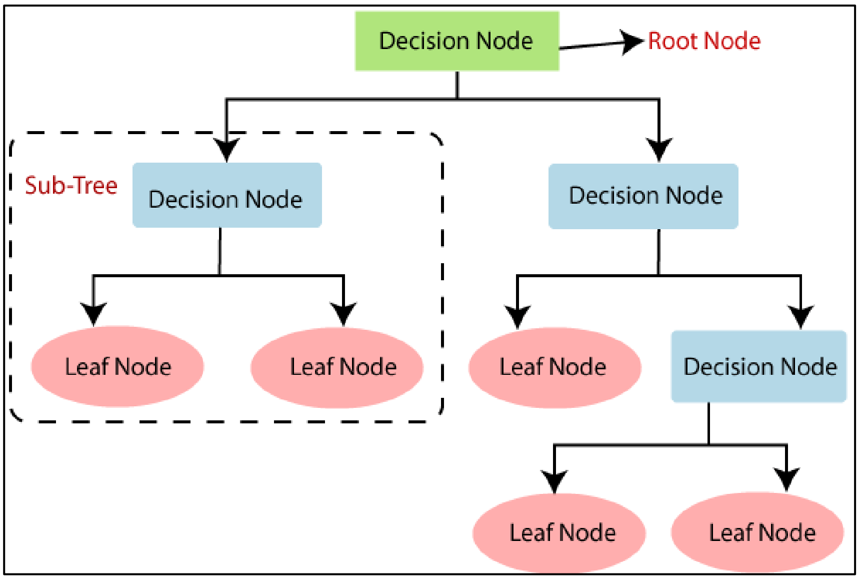


Fig. 3 Structure of Decision Tree models
Source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

SHAP, SHapley Additive exPlanations[20] is a local model-agnostic explanation method. It connects optimal credit allocation with local explanations using the classic Shapley values, a concept from cooperative game theory. This method aims to explain the prediction of an instance/observation by computing the contribution of each feature to the prediction. The advantages offered by this method are: local accuracy, missingness, and stability. Shap can be realized through different explanation models that differ in how they approximate the computation of the shap values.

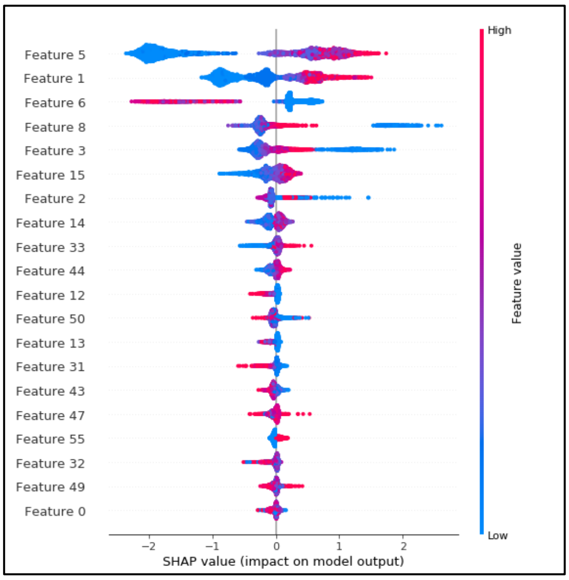


Fig. 4 Shap value analysis plot
Source:https://datascience.stackexchange.com/questions/65307/how-to-interpret-shapley-value-plot-for-a-model

[20] Scott M Lundberg and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), <https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.

Lastly, Anchors is a model-agnostic black-box explanation method. It can be used to interpret any type of black-box model. In this case, the benchmarking domain is recruitment and this system focuses on providing a predictive pipeline for employee attrition in order to enable timely implementation of retention policies. Given a black-box model and an x data point with a particular label, you want to find a "set of conditions on the input characteristics that are necessary and sufficient for the model to predict the label for that data point. These conditions form an Anchor." This model operates in three steps:

“1) Find a positive example: through a binary search, a positive counterfactual example is found x+. This is an example that has a different input but still receives the same prediction as x. This step stops when x and x+ differ by less than a threshold . The goal is to find a counterfactual positive example that is as similar to x as possible;

2) Finding a negative example: again, through a binary search, a negative counterfactual example x- is found. This is an example that has the same input as x but receives a different prediction. This step stops when x and x- differ by less than a threshold .

3) Finding an Anchor: the algorithm finds an Anchor by starting with the input features from the counterfactual positive example x+ and iteratively removing features that are not predictive of the model’s output. This iterative procedure is repeated for all the features in x, until either the Anchor is empty or the Anchor satisfies a minimum-size threshold.”

Using Anchors’ technique provides explanations for any type of machine learning model. In the previous sections we have already stated how important it is for an AI system for recruitment to provide an explanation that is fair, transparent and accountable. The prediction of a system that gives importance to these three factors directly increases the trust that the decision-maker and the potential candidate have towards the system. Secondly, it is clear that the use of simpler and more immediate models allows a greater understanding of the decisions made by the AI system. We believe that the Anchors system reflects all the characteristics specified above as it allows to compose a human-readable explanation of model’s output after having correctly classified and grouped the processed information. For example, one of the possible uses of the Anchors model concerns its application aimed to distinguish which are the leading drivers of attrition among all leaving employees. The advantage offered, in this case, is not only the ability to view data of interest but also to be able to better understand the motivation behind the decision suggested by the system. A quick but complete understanding of this information allows the decision-maker to save time on creating an employee-specific retention policy. Finally, we observe that the more points of control there are in the recruitment process, the less complex is the single steps of the process, as it is possible to divide the phases into sub-phases maintaining the same level of security and trust.

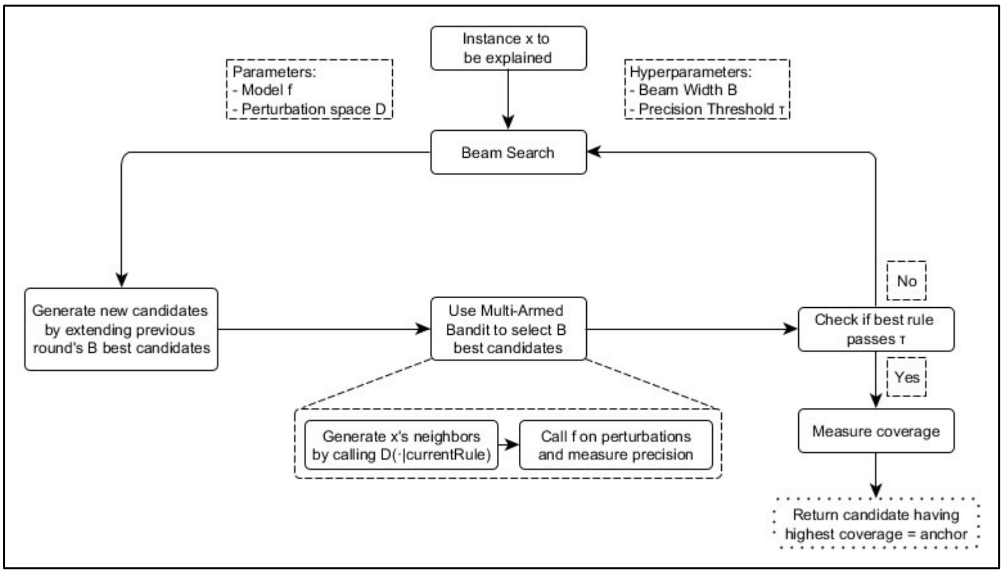

Instance x to be explained
Parameters:
- Model f
- Perturbation space D
Hyperparameters:
- Beam Width B
- Precision Threshold τ
Beam Search
Generate new candidates by extending previous round's B best candidates
Use Multi-Armed Bandit to select B best candidates
Check if best rule passes τ
No
Yes
Generate x's neighbors by calling D(·|currentRule)
Call f on perturbations and measure precision
Measure coverage
Return candidate having highest coverage = anchor

Fig. 5 **The Anchors algorithm's components**
Source: https://christophm.github.io/interpretable-ml-book/anchors.html

A wide range of possible applications however does not eliminate the risk of controversial situations. In the course of the thesis we will bring various examples of uses and implementations that present troubling criticalities. The first that we propose concerns an algorithm used by the Amazon Company to evaluate the job positions of candidates. Amazon's computer models were trained to monitor candidates by looking at templates in resumes submitted to the company over a 10-year period. However, in 2018, Amazon decided to abandon the use of AI for candidate screening. Resumes were primarily sent by men who dominated the technology industry, and AI gave less emphasis to those who included the word "women" and downgraded graduates of two all-women colleges." Numerous studies show how the presence of bias undermines the confidence of less experienced users, slowing the spread and evolution of this kind of technology. In order to reverse this trend, many scholars recommend integrating more fair and transparent recommendation systems to provide an explanation to the user about the processes and decisions recommended by the algorithm. However, The AI community is concerned about the black-box issue and more studies try to propose many ideas to improve trustworthy AIs in order to make them safe to use. One possible solution could be turning Artificial Intelligence (AI) into explainable Artificial Intelligence (XAI) which we define as follows: "given a certain audience, an explainable Artificial Intelligence is one that produces details or reasons to make its operation clear or easy to understand"[21]. XAI refers to a variety of approaches (e.g., reverse engineering) to overcome the opaque nature of certain types of AI-based systems, such

[21] Barredo Arrieta and others.

as deep neural networks[22]". The implementation of XAI improves the indicators of interpretability and explainability. The concept of interpretability "enables developers to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its result" while the second one "provides insight into the DNN's decision to the end-user in order to build trust that the AI is making correct and non-biased decisions based on facts". Key concept of this approach is explainability "which is defined as the ability of an AI to provide information about what its algorithm is doing, has been suggested as an additional role for an AI decision maker to support human decisions"[23]. The introduction of this attribute, we discover that it is crucial for a better cooperation between the human being and AI, establishing a relationship with precise and differentiated objectives and capabilities. Recent studies have differentiated the roles of human decision makers and AI[24]. In particular, human decision-makers have strengths in problem-solving, expansion of ideas and consideration of quality values, while AI decision-makers have strengths in data collection, in the execution of ideas and in considering quantitative values[25]. These two actors are compatible, so that the AI system prepares the basis for the decision and man makes the final decision[26].

In the next section we will address the use of AI systems for recruitment from a legislative point of view, observing the guidelines provided for by the AI ACT, a European regulation aimed at regulating the use of AI systems.

[22] Guidotti and others; Christian Meske and others, 'Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities', *Information Systems Management*, 39.1 (2022), 53–63 <https://doi.org/10.1080/10580530.2020.1849465>.

[23] London.

[24] Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi, 'Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges and Research Agenda', *International Journal of Information Management*, 48 (2019), 63–71 <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>.

[25] Mohammad Hossein Jarrahi, 'Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making', *Business Horizons*, 61.4 (2018), 577–86 <https://doi.org/10.1016/j.bushor.2018.03.007>.

[26] Mark Sendak and others, '"The Human Body Is a Black Box": Supporting Clinical Decision-Making with Deep Learning', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (New York, NY, USA: Association for Computing Machinery, 2020), pp. 99–109 <https://doi.org/10.1145/3351095.3372827>.

CHAPTER 3: LEGAL FRAMEWORK

Section 3.1 History and motivation

As we have already noted in the first chapter, the use of Artificial Intelligence systems is constantly growing. The causes of this increase are many and must be found in the easy accessibility of which these users can enjoy and in the many advantages they offer. In some cases you can access this technology by connecting to a site and with a few clicks here is that this technology becomes viral and usable by anyone who can surf the net. In this chapter we will look at the legal impact of these technologies on the population and the individual.

Let us first analyze the legal framework for the design and use of an Artificial Intelligence system. We observe that on the subject, the world legislature is still struggling to take a clear and clear position. The spread and design of Artificial Intelligence systems that we have seen in the last three years has undergone a sudden and unexpected acceleration. A technology that was usually intended for technicians and scholars in the field, soon became at the mercy of every user able to connect to the network. A phenomenon of this magnitude caught the governments of the world, unable to formulate rapid regulation in this regard. There was no, and there is still no, common legislation on the subject, each Country is moving in the direction that best cares for its interests. The USA, in fact, initially adopted a lenient approach and then followed a different line of thinking because the calls for regulation have recently been mounting. In China, the regulation in this regard has been immediately under the attention of The Cyberspace Administration of China" which is also consulting on a proposal to regulate AI. At international level, the Organisation for Economic Co-operation and Development (OECD) adopted a (non-binding) Recommendation on AI in 2019, UNESCO adopted Recommendations on the Ethics of AI in 2021, and the Council of Europe is currently working on an international convention on AI. Furthermore, in the context of the newly established EU-US tech partnership (the Trade and Technology Council), the EU and USA are seeking to develop a mutual understanding on the principles underlining trustworthy and responsible AI. [27]

The European Union, always very attentive to the guarantee of issues such as privacy and human rights protections, has immediately become a world spokesperson, seeking first to regulate the issue of Artificial Intelligence in a decisive and rapid manner. In 2019 the European Commission published a non-binding act concerning ethical models for the reliability of Artificial Intelligence. The Member States of the Union transposed the act. The approach suggested by the Commission was a "soft law", a type of non-legally binding approach that suggests a series of behaviors that can prevent hypothetical conflicts of interest by relying on the spontaneous adherence of the subjects to whom

[27] European Commission, ‘Artificial Intelligence Act’, 2023.

such rules could benefit in finding a solution appropriate or inappropriate. But it was soon necessary to legislate on this subject in a more decisive and detailed manner. The need to harmonise the legislative approach within and between Member States has always been one of the most important objectives for European bodies. It became necessary to take a step forward in regulating the use of Artificial Intelligence systems. In 2020 the European Commission issued the White Paper on Artificial Intelligence in order "to promote the uptake of AI and address the risks associated with certain uses of this new technology". Following the White Paper, "the Commission launched a broad public consultation in 2020 and published an Impact Assessment of the regulation on Artificial Intelligence, a supporting study and a draft proposal, which received feedback from a variety of stakeholders. In its impact assessment, the Commission identifies several problems raised by the development and use of AI systems, due to their specific characteristics."

SECTION 3.2 AI ACT STRUCTURE

Given these premises, let us look in more detail at the issues dealt with in the AI ACT.

First, it is advisable to understand the goals it aims to achieve. "The general objective of the proposed AI act unveiled in April 2021 is to ensure the proper functioning of the single market by creating the conditions for the development and use of trustworthy AI systems in the Union."[28]. A key aspect of this process is the uniform creation of common legislation for all European States. A harmonised legal framework allows greater fluidity in the application of laws. Since all Member States have a common rule to refer to, the application of the same rule is simpler and more immediate, eliminating doubts about the relevant legislation. This dynamic offers many advantages, but it should be noted that it requires a much more complex and lengthy process of drafting, debating and creating the standard than any European citizen is accustomed to. If not efficiently optimised, this process risks slowing down the regulation of AI, thus creating a non-optimal legislative stalemate for Member States, or worse, creating a state of legislative conflict in which each Member State promotes its own laws contrary to the vision of other States. The harmonisation process has been one of the priorities of the organs of the European Union since the creation of the European Union. The practical objective of the AI ACT is to define a common mandatory requirements applicable to the design and development of AI systems before they are placed on the market and harmonises the way ex-post controls are conducted. The approach suggested by the AI ACT is implemented in two stages: the first, ex-ante to the creation of the Artificial Intelligence system, aims to describe the minimum legal requirements that are required by the system to legitimize its placing on the market; The second one, ex-post to its design, is aimed at standardizing

[28] European Commission, 'Artificial Intelligence Act', 2023.

the control and review tools used to verify the systems. “The Commission proposes to follow the logic of the new legislative framework (NLF), i.e. the EU approach to ensuring a range of products comply with the applicable legislation when they are placed on the EU market through conformity assessments and the use of CE marking.”[29]

Secondly, let us consider the scope of the act. As it itself states: “The new rules would apply primarily to providers of AI systems established within the EU or in a third country placing AI systems on the EU market or putting them into service in the EU, as well as to users of AI systems located in the EU. To prevent circumvention of the regulation, the new rules would also apply to providers and users of AI systems located in a third country where the output produced by those systems is used in the EU.” The applicability of this act therefore extends to all services offered to the population of the EU within the borders of the Member States and to all systems whose output is used within the EU. The purpose of using the output is not specified. The legislator chooses to formulate the norm in this way because it allows him to include implicitly within this expression all the fields worthy of note, avoiding to number them in detail one by one. In doing so, it will not be necessary to amend the act from time to time if a new scope is identified. This prediction is supported by particular cases, precisely identified by the legislator: “the draft regulation does not apply to AI systems developed or used exclusively for military purposes, to public authorities in a third country, nor to international organisations, or authorities using AI systems in the framework of international agreements for law enforcement and judicial cooperation.”[30]

SECTION 3.3 AI ACT APPROACH

The legislator then details the committee’s approach to regulating the use of Artificial Intelligence systems. It is a risk-based approach whereby legal intervention is tailored to concrete level of risk. At the design stage, each AI model should incorporate a certain type of instruments or ethical values that comply with the minimum forecasts suggested by the AI ACT. Depending on the objective of the system, its complexity and its scope, it will be placed in a category to describe the level of inherent risk it entails. There are 4 categories: *unacceptable risk, high risk, limited risk, and low or minimal risk*.

[29] European Commission, ‘Artificial Intelligence Act’, 2023.
[30] European Commission, ‘Artificial Intelligence Act’, 2023.

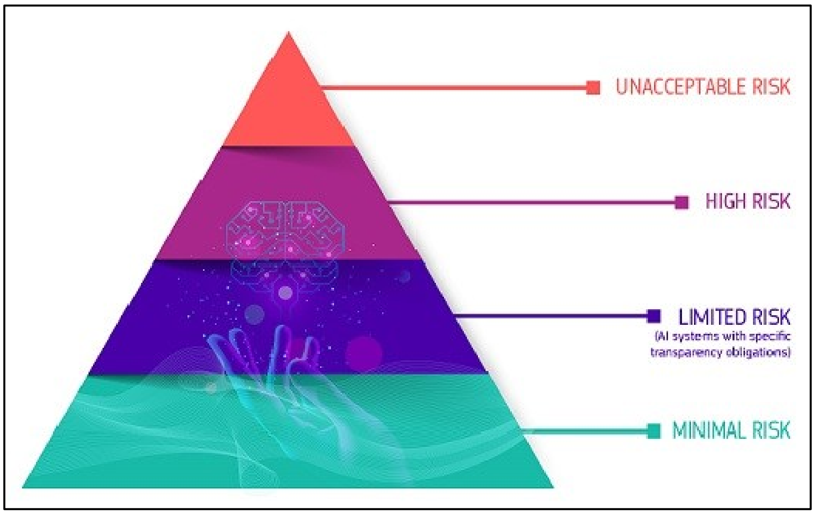


Fig. 5 Source: European Commission

AI applications would be regulated only as strictly necessary to address specific levels of risk. The legislator chooses this type of approach because it is strongly based on the practical impact that an Artificial Intelligence system has on the population. At the lowest level, low or minimal risk, "AI systems could be developed and used in the EU without conforming to any additional legal obligations. However, the proposed AI act envisages the creation of codes of conduct to encourage providers of non-high-risk AI systems to voluntarily apply the mandatory requirements for high-risk AI systems."[31] If AI systems presenting 'limited risk', such as systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorisation systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes), would be subject to a limited set of transparency obligations."[32] We observe that the character of transparency, of which we will talk in more depth in the next chapter, turns out to be a fundamental attribute for the design of a proper Artificial Intelligence system. The penultimate category is reserved for high-risk systems. The use of these systems has a serious impact on the safety of individuals or the protection of their fundamental rights. We distinguish two categories:

- Systems used as a safety component of a product or falling under EU health and safety harmonisation legislation (e.g. toys, aviation, cars, medical devices, lifts).
- Systems deployed in eight specific areas identified in Annex III, which the Commission could update as necessary through delegated acts:
  - Biometric identification and categorisation of natural persons;
  - Management and operation of critical infrastructure;
  - Education and vocational training;

[31] European Commission, 'Artificial Intelligence Act', 2023.
[32] European Commission, 'Artificial Intelligence Act', 2023.

- Employment, worker management and access to self-employment;
- Access to and enjoyment of essential private services and public services and benefits;
- Law enforcement;
- Migration, asylum and border control management;
- Administration of justice and democratic processes.[33]

As mentioned above, the recruitment phases of a possible candidate are activities that make an Artificial Intelligence system "high risk". In this thesis we will refer to this category for the analysis of future models. Unlike the previous categories, in this case the legislator provides for a number of precise requirements. First they “would be required to register their systems in an EU-wide database managed by the Commission before placing them on the market or putting them into service. Any AI products and services governed by existing product safety legislation will fall under the existing third-party conformity frameworks that already apply. Providers of AI systems not currently governed by EU legislation would have to conduct their own conformity assessment (self-assessment) showing that they comply with the new requirements and can use CE marking”[34]. We specify that these are not the only provisions introduced by the AI ACT. “Such high-risk AI systems would have to comply with a range of requirements particularly on risk management, testing, technical robustness, data training and data governance, transparency, human oversight, and cybersecurity.”[35] An additional obligation is provided for providers from outside the EU. In this case is required an authorised representative in order to ensure the conformity assessment, establish a post-market monitoring system and take corrective action as needed. The legislator illustrates a last category, that of the unacceptable risk. In this case the legislator identifies practices about the use of an Artificial Intelligence system not allowed. The use of these means in certain areas would entail a risk to the safety of individuals or the protection of their fundamental rights. The AI ACT prohibits the trade or use of such systems where they: deploy harmful manipulative 'subliminal techniques'; exploit specific vulnerable groups (physical or mental disability); used by public authorities, or on their behalf, for social scoring purposes.

Subsequently the European Commission indicates two figures of reference for the control of the correct application of the AI ACT. They are located on two different levels. At EU level is designated the EAIB, (European Artificial Intelligence Board) which is composed of representatives from the Member States and the Commission. This figure represents a first common barrier for all Member States. The second reference figure is at national level. Each country will have the obligation to

[33] European Commission, ‘Artificial Intelligence Act’, 2023.
[34] European Commission, ‘Artificial Intelligence Act’, 2023.
[35] European Commission, ‘Artificial Intelligence Act’, 2023.

establish an internal body or designate an existing one capable of supervising at national level certain events involving Artificial Intelligence systems. As the legislator suggests “National market surveillance authorities would be responsible for assessing operators' compliance with the obligations and requirements for high-risk AI systems. They would have access to confidential information (including the source code of the AI systems) and subject to binding confidentiality obligations. Furthermore, they would be required to take any corrective measures to prohibit, restrict, withdraw or recall AI systems that do not comply with the AI act, or that, although compliant, present a risk to health or safety of persons or to fundamental rights or other public interest protection.”[36] The identified figure, therefore, will not deal directly with the dynamics of low or low risk systems, operating on higher level systems. The competent body is empowered to act directly on systems that do not comply with the minimum requirements set out in the AI ACT, taking all appropriate restrictive or punitive measures.

After having observed the reference legislative framework, in the next chapter we will examine in detail some recommendation systems used in the field of job recruitment.

[36] European Commission, ‘Artificial Intelligence Act’, 2023..

CHAPTER 4: REVIEW

In this chapter, we will analyze in more detail some recommendation systems used in recruitment, in order to fully understand their advantages and possible vulnerabilities. Each recommendation model aims to provide a quantitative and qualitative value in the matching between the job proposal of the company and the application of a subject. Most of the recruitment process is managed by a ML (or AI) systems, while the human recruiter is entrusted with the final stage of screening and selection of candidates.

4.1 ANALYSIS OF THE FAT-CAT MODEL

In this section we will analyze in detail which attributes regarding the explainability of the output model should be considered in the construction phase of an AI system. To achieve this, we mention a first case study on AI recruitment-system adoption, the FAT-CAT model37. This model "describes the path from explainability to AI system adoption considering augmentation, assuming that the capability of the AI decision maker to explain the basis of its decision and interact with the human decision maker is crucial for AI recruitment system adoption". The model's name refers to the objectives that the model aims to ensure, such as fairness, accountability, and transparency (FAT); and to the motivational indicators used to replace human-AI augmentation attributes, such as complexity, anxiety and trust (CAT). The advantages offered by this model are many: it allows to be applied on any AI adoption case systems and provide a guideline for developing user-friendly ones. The study conducted on the effectiveness of the FAT-CAT model reveals which role and what impact the attributes had: fairness, transparency, accountability on motivational indicators such as: complexity, anxiety and trust.

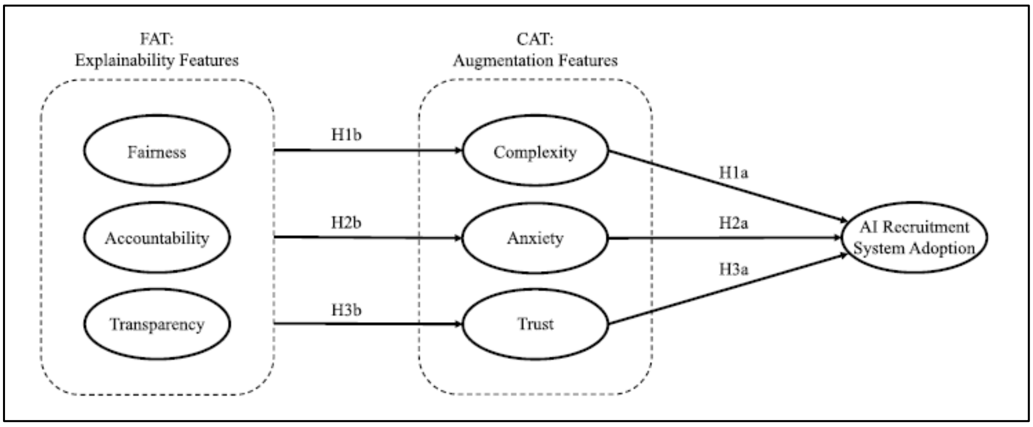


Fig. 6 Proposed FAT-CAT model
Source: FAT-CAT—Explainability and augmentation for an AI system: A case study on AI recruitment-system adoption,2023

[37] ChangHyun Lee and KyungJin Cha, 'FAT-CAT—Explainability and Augmentation for an AI System: A Case Study on AI Recruitment-System Adoption', *International Journal of Human-Computer Studies*, 171 (2023), 102976 <https://doi.org/10.1016/j.ijhcs.2022.102976>.

To verify the model, the developers collected survey data from human-resource managers in the Republic of Korea who have utilized or are considering adopting AI recruitment systems, collecting 224 responses. The questionnaire was composed by 31 items about the role of FAT attributes in modifying CAT indicators. "To evaluate the reliability of the analysis, we tested the convergent and discriminant validities. We first assessed that the convergent validity is reliable with a 0.5 or higher average variance extracted (AVE) value and with a 0.7 or higher composite reliability and Cronbach's alpha values." These values have been chosen because the average variance extracted should be higher than the minimum threshold of 0.5 to be acceptable. However even if AVE is less than 0.5, but composite reliability is higher than 0.6, the convergent validity of the construct is still adequate38. Regarding Cronbach's alpha values, statistical studies show that obtaining a value of 0.70 is acceptable while starting from the value 0.80 and above is very good. We conclude that the level of reliability on convergent tests is acceptable but not optimal. "The discriminant validity is assessed to be reliable when the latent variable correlation of each construct is lower than either 0.85 or the square root of the corresponding AVE value39". On the contrary, we note that the degree of reliability offered by the test on discriminant validity is optimal. Standardized path coefficients were used to confirm the explanatory power of each construct, and bias-corrected confidence intervals were used to confirm their significance. Compared to the complexity of the hiring decision, it is shown that it will negatively affect AI hiring adoption and that an AI hiring system designed to be transparent and responsible will decrease complexity. The implementation of a fair system does not contribute significantly to this process. With regard to anxiety towards the hiring decision, it is shown that an AI recruitment system designed to be fair and transparent will decrease anxiety towards the hiring decision. The implementation of an accountable system does not contribute significantly to this process. However, it should be noted that there is no evidence that anxiety about the recruitment decision will adversely affect the adoption of the AI recruitment system. Finally, it has been shown that the trust will positively influence the adoption of the AI recruitment system. You can also increase this indicator by implementing a fair, transparent and responsible AI recruitment system.

In Chapter II, we have theoretically described the concepts of fairness, accountability and Transparency. Observing what has been demonstrated by the analyses carried out on the FAT-CAT model, these attributes are fundamental also for the recruitment personnel, those who themselves should interface with these systems in order to automate the recruitment process.

[38] 'Evaluating Structural Equation Models with Unobservable Variables and Measurement Error - Claes Fornell, David F. Larcker, 1981' <https://journals.sagepub.com/doi/abs/10.1177/002224378101800104>.

[39] 'Evaluating Structural Equation Models with Unobservable Variables and Measurement Error - Claes Fornell, David F. Larcker, 1981'; Muhammad Shakaib Akram and others, 'Exploring the Interrelationships between Technological Predictors and Behavioral Mediators in Online Tax Filing: The Moderating Role of Perceived Risk', *Government Information Quarterly*, 36.2 (2019), 237–51 <https://doi.org/10.1016/j.giq.2018.12.007>.

4.2 Analysis of AaJeeViKa model

The second recommendation model that we will analyze is called AaJeeVika40. It is a "trusted and decentralized solution [...] which integrates blockchain and explainable AI (XAI) to integrate trust analysis into recruitment and personnel processes"41. The use of distributed registers, the main element behind blockchain technology, can be a potential solution to drive transparent, chronological and unchangeable human resource management (HRM) processes. Following this model, each candidate is evaluated, assigning a value to a series of attributes that refer to certain areas of interest for his job evaluation, such as education profile, unique identification number (social security identifier), criminal history, details from previous employers and social media profiles respectively. This information is only considered in relation to public repositories. A content key is generated for each attribute. The data protected by the content key and the content key undergo a hashing process, generating a reference. References are recorded in a blockchain, so that the integrity of the information they refer to can be guaranteed. The company that intends to post a job will undergo a similar process, with the difference that its attributes refer to the description of the offer must. The algorithm allows you to automatically associate candidate attributes with any company job posts based on job description. The company receives a notification about the number of associations made by the algorithm and sends each candidate a request to share the details of their content key, so that they can access the records of the candidate's attributes and verify them. The algorithm then generates a JSS (job suitability score), a value created by analyzing the score attribute. This step is a key moment of the whole procedure because if the JSS value assigned by the algorithm will exceed a certain threshold (0.7) the process of creating the interview call between the candidate and the recruiter is started. On the contrary, if the predicted value is below the identified threshold, the selection procedure of the potential candidate will not continue. This threshold is established on the selection of the identified criteria for solving the attribute selection problem. These criteria will calculate values for every attribute. Among the various criteria, we mention two of the most important: Entropy and Information Gain. Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information; while the second one is a statistical property that measures how well a given attribute separates the training examples according to their target classification. The main goal in this case is finding an attribute that returns the highest information gain and the smallest entropy. Information about call details (such as interview date, venue, and time) is recorded in the blockchain via a timestamp procedure. Once the call has been made, the recruiter sends his assessments of the

[40] Lennart Hofeditz and others, 'Applying XAI to an AI-Based System for Candidate Management to Mitigate Bias and Discrimination in Hiring', *Electronic Markets*, 32.4 (2022), 2207–33 <https://doi.org/10.1007/s12525-022-00600-9>.

[41] Hofeditz and others.

candidate to the platform. The JSS is updated with new information. The final step concerns the creation of the employee reputation score (ERS) while "the output prediction significance is computed by Shapley additive explanations (SHAP) explainers. The XAI result along with other information is meta-recorded and updated on BC ledgers."

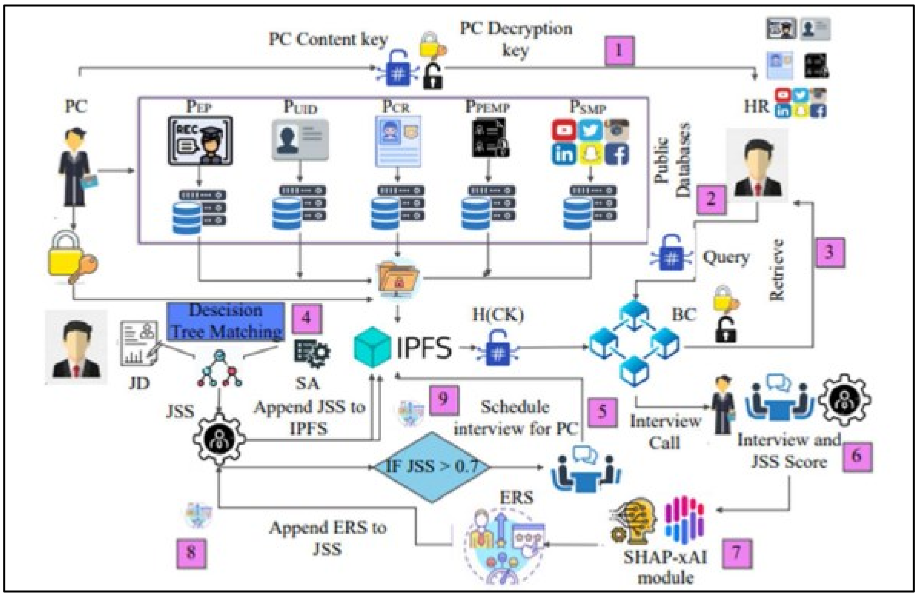


Fig. 7 Proposed AaJeeViKa model
Source: Applying XAI to an AI-Based System for Candidate Management to Mitigate Bias and Discrimination in Hiring,2022

Given this description of the model, we can note what are the most alarming criticalities: on the one hand there is the protection of privacy; on the other hand, there is the risk of making a wrong or incomplete or prejudicial assignment of the value JSS, which would exclude a subject from the possibility of taking the oral interview with the recruiter. Analyzing in detail the algorithm, we can deduce that the generation of the JSS happens through a decision tree model. Decision tree models "are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation". In this case the model is trained in order to maximize the "gain ratio" of a note in decision tree model, that is the relationship between the information gain and the cost of split information. In addition, "to model overfitting, the classificatory tree uses pruning of nodes, such that overall accuracy is not affected. For the same, data is split into train and test for validation." The creators of the AaJeeViKa model have decided to adopt this kind of algorithm in relation to the many advantages it entails. We conclude that the process of generating the JSS takes place in a fair manner and is not affected by external factors prejudicial to the candidate.

Secondly, we analyze an additional aspect of the AaJeeVika model, that is the SHAP module, necessary to establish the contribution made by the various attributes in the generation of the final ERS. "The attributes are sorted on the relative importance of attributes and the importance they would have on prediction results. As an example, the topmost attribute signifies a high impact, and vice versa." It should be noted that the attributes chosen for the assessment of the possible candidate concern mainly his academic-working, in order to establish his skills and his predisposition to the job offered. For example the AEP attribute (education profile) examines some Sub Parameters such as "educational institute reputation, education quality, degree quality, skill sets, and certificate importance;" while, the APE attribute describes "the evaluation of the performance of the previous employee, the employee satisfaction ratio, working hours, the number of companies (nature of attrition), the number of previous companies and the number of projects in the current company". However, some attributes do not directly concern the candidate’s working sphere, but his private life. For example, "For AUID (unique identifier) attribute, we consider sub attributes which denotes candidate age, liabilities, current location address, defaults, and income tax deduction status. In ACR, we find the credit history as a boolean whether the person is a defaulter or genuine." The question arises as to whether the analysis of the financial situation of the candidate contributes in an appreciable and justifiable way to his assessment. Moreover, it is noted that the prediction for the identification of gender has only two types of classes, effectively excluding the possibility of a subject to be identified differently. The Italian government and many governments around the world are debating precisely these issues. In our opinion, inclusiveness and respect for human rights is achieved when, in the personal sphere, a person is given the opportunity to think and recognize themselves in the denomination of a category of gender in an autonomous way and free from political and cultural preconceptions. We hope that this can be seen and corrected as soon as possible. Regarding the attainment of the final objective of the AaJeeVika model it is evidenced a precision of the forecasts of 84%. This number sets a very important threshold as it represents the the accuracy that the model does not label an instance positive that is negative. The model aims to avoid to produce false positive errors. Therefore, we conclude that it is more acceptable a greater number of false positives, rather than false negatives, even if the human recruiter will have to examine a higher number of candidates than would really correspond to the required job position. This possibility is an acceptable compromise for the objectives that the model wants to achieve.

4.3 Analysis FairCVtest model

Below we propose the Analysis of a third AI model used to automate the recruitment phases. This model is called FairCVtest and is a multimodal model able to understand and process information from multiple heterogeneous sources of information42. These types of models are capable of processing multiple mode inputs (text and images, text and audio, text and video, etc.) without processing them separately. In this way the data learned from the inputs can be grouped into similar categories to be evaluated together. FairCVtest model takes into input both structured data and unstructured data from images, audio, and text. The possibility to acquire a mass of data of various nature allows to contextualize and better describe the possible candidate, taking into consideration a greater number of features. It is useful to observe the nature of the data chosen for model training simulations. As in previous models, the data concerning merits of the candidate e his demographic attributes have a very high value. Five indicators have been analysed to evaluate candidate competencies, such as: education attainment, availability, previous experience, the existence of a recommendation letter and language proficiency in a set of three different and common languages. Each profile is assigned demographic attributes to describe its genre and ethnicity. Unfortunately, such encoding occurs through a binary solution between two classes only, and that of ethnicity can be described in only three classes. As already described above, providing such a limited number of options does not facilitate the position of the candidate, who rather is forced to wear a label that does not always belong to him that there is: we notice that to attribute a value correctly to the attribute of the ethnicity three options are not enough, but many more are needed. The value of this information is even more valuable when you see that "these demographic attributes determine the face image (gender and ethnicity related), name (gender related), and pronouns in the short biography (gender related)." We hope that fairer measures will soon be taken against the candidate.

The novelty introduced by the FairCVtest model is the analysis of unstructured data such as a face photo or a short biography. "A face image is rich in unstructured information such as identity, gender, ethnicity, or age. That information can be recognized in the image, but it requires a cognitive or automatic process trained previously for that task. The text is also rich in unstructured information. The language and the way we use that language, determine attributes related to your nationality, age, or gender." By combining the information extracted from the structured data analysis provided above, it is possible to outline a profile of the candidate. Each profile is matched with an image and text in specific databases that match the profile characteristics. A perfectly complete profile includes: gender and ethnicity attributes, an occupation, a face image, a name, seven attributes obtained from the analysis of candidates skills and a short biography. The evaluation phase of a candidate's profile is the final part of this long process. The candidate score predictor is designed as a multimodal neural network with three input branches: face image, text biography and candidate competencies. The first two branches are analysed separately, before fusing the

[42] Alejandro Peña and others, 'Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI-Based Recruitment', *SN Computer Science*, 4.5 (2023), 434 <https://doi.org/10.1007/s42979-023-01733-0>.

information from all three modalities. Lastly, the face and text features obtained from its respective analysys are combined with the candidate competencies to feed the multimodal network. This network is composed by two hidden layers, with 40 and 20 neurons respectively. The use of structured and unstructured data thus increases the possibility to describe better the candidate.

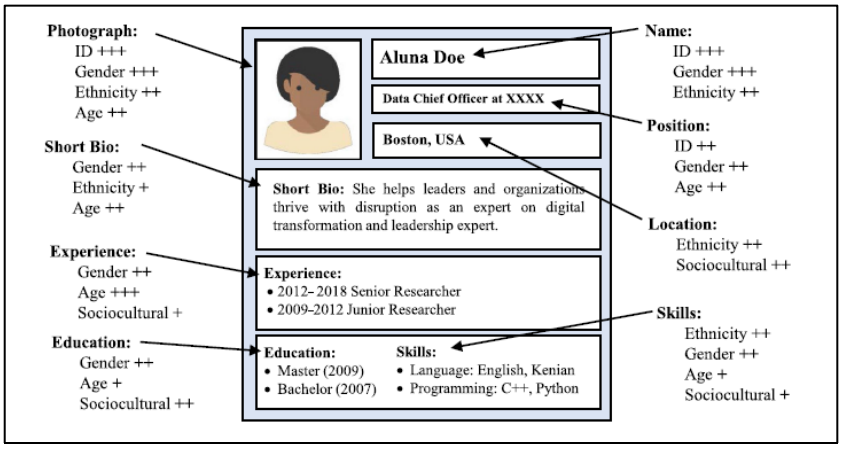


Fig. 8: Example of a candidate's profile analysed by FairCVtest model
Source: Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI-Based Recruitment

However, it also reveals a delicate critical issue: where multiple sources of both structured and unstructured data play a key role in algorithms' decisions, the task of detecting and preventing biases becomes even more relevant and difficult. For example, the correctness of an image assignment or short biography or the selection of a database that does not contain discriminatory biases of a racial or sexist nature become more complex features to complete. To know the real impact that these features have on the model were produced three different simulations, in which the model has taken into account slight different input data and the target function but which maintain an identical structure. We call the three simulations in neutral simulation, biased simulation and agnostic simulation. The first simulation provides a phase of training with unbiased scores, the original face representation extracted, and the biography with explicit gender indicators; the second simulation provides a phase of training with biased scores, the original face representation, and the biography with explicit gender indicators; and the third simulation provides a phase of training with biased scores, the gender and ethnicity agnostic representation, and the "blind" biography. For each simulation three different algorithms classifiers will be trained, it is possible to check which type of algorithms has the best accuracy. The impact of bias was by observing differences between demographic groups, especially considering the values of gender and ethnicity. From the research carried out by the developer team of this model, it turns out that in the first simulation there are substantially these differences. As expected, using the unbiased scores and a balanced training set leads us to an unbiased classifier. In the second simulation, however, there is a clear difference in treatment compared to the first simulation, in particular in the gender classification, to the disadvantage of female candidates. Finally, the third simulation is about halfway between the previous two in terms of performance offered. There are no major differences in gender distribution but there is a slight difference in the other case. "The difference observed in the behavior of gender and ethnicity agnostic cases can be explained by the fact that it's been removed almost all gender information from the input but for the ethnicity it only took measures on the face embedding, not on the competencies." Finally, comparing the data regarding the accuracy of the algorithms used, we discover on average the simulation that offers the best accuracy along all the algorithms is the second, the biased simulation that trains data containing bias. This result does not surprise us as a bias tends to divert performance in a decisive way. It is necessary to make a separate reflection for the third simulation which aims to prevent the system to inherit data biases. The agnostic simulation uses a gender blind version of the biographies, as well as a face embedding where sensitive information has been removed. This choice however has not repaid in terms of performances being the third simulation that offers medium the worst accuracy along all the algorithms. Therefore, reducing the number of personal information processed, such as gender, can imply a reduction in the level of accuracy of the algorithm that is trained on that data. We conclude by saying that the FairCVtest model has confirmed some fears that many technicians and not have against Artificial Intelligence systems applied in the workplace: discriminatory biases strongly reduce the fairness of the systems and attempt to limit them, limiting the number of information available, can adversely affect the accuracy of predictions.

4.4 Discussion

In this chapter we presented the technical analysis of three different AI models used in recruitment. The first model, named FAT-CAT model, aims to ensure the integrity of some attributes considered fundamental by the team of developers, such as fairness, accountability and transparency. In our opinion, among the three models described, it is presented as the most careful to ensure these attributes. Not only that, this model also focuses on the theme of augmentation between human and AI. It is the only model that explicitly aims to improve the level of trust and reduce the level of anxiety and complexity of the subjects with whom it interacts. These features underscore the importance and attention the developer team has had to make the model as fair and understandable as possible. The second model, AaJeeViKa model, has many strengths. It allows to establish a clear and direct match between the human-working skills offered by the candidate and the skills required by the company by publishing the job offer. The evaluation of each candidate takes place through information collected through publications. This aspect is a great strength, as it aims to protect the privacy of the subject and ensures greater transparency on the dynamics that govern the entire decision-making process. Moreover, according to the simulations produced by the team of developers, the model offers a level of performance of good level, evaluating the accuracy reached

84%. Finally, we observe that a method has been skillfully implemented to provide an explanation about the attributes that have had a greater impact in the evaluation. This model was developed in India and therefore to adapt to the legal limitations required by the Indian country. As already specified in the previous chapter, the European Commission identifies a series of mandatory minimum measures for Artificial Intelligence systems that pose a high risk for the security of the person and for the protection of his fundamental rights. This category includes systems deployed in employment, worker management and access to self-employment, such as the AaJeeVika model. We believe that, at the moment, the proposed model does not fully meet the minimum measures required. However, we hope that with some changes this model can be approved and introduced into the lives of EU citizens. As already mentioned above, the third model, called FairCVtest, appears to be the most complex model among those proposed. The use of structured and unstructured data significantly increases the degree of complexity of the model. We note, however, that the model has been designed taking into account the forecasts imposed by Regulation 679/2019 promoted by the European Commission, called General Data Protection Regulation (GDPR) and therefore lends itself very well to being part of the European legislative context, even though we have noticed some small problems that we think should be resolved as soon as possible. Please note that the FairCVtest model does not integrate any XAI system. This strongly limits the level of transparency and trust of the model, not implementing the explainability attribute like the other two models. Below we propose a summary table of the models analyzed, taking into account their main characteristics both at a technical and legal level.

MODELS		
FAT-CAT	AjeeViKa	FairCVtest
Model that describes the path from explainability to AI system adoption considering augmentation, assuming that the capability of the AI decision maker to explain the basis of its decision. It focuses on implementing fairness, accountability, and transparency in order to increase trust and reducing the complexity and anxiety of the model. This study promotes the using of XAI, proving its empirical contribution on supporting a fair model. In the future this model potentially can become an important instrument for human-resource management.	This model integrates blockchain and XAI technology in order to product trusted analytics in recruitment process. It allows to establish a clear and direct match between the human-working skills offered by the candidate and the skills required by the company by publishing the job offer. It integrates two different explanation methods: a Decision Tree method and SHAP method. This model was developed according to the legal limitations required by the Indian country. At the moment, the proposed model does not fully meet the minimum measures required by the European Commission for AI systems.	Automatic recruitment algorithm which using a set of multimodal synthetic profiles including image, text, and structured data, such as education attainment, availability, previous experience, the existence of a recommendation letter and language proficiency in a set of three different and common languages. Detecting and preventing biases becomes even more relevant and difficult. This model does not integrate any XAI system, limiting the level of transparency and trust of the model. However, it's been projected following GDPR principles and the European approach.

4.5 Remarks

In conclusion, we believe that the use of AI systems in recruitment can be an excellent tool to support decision-makers. However, there must be fundamental attributes to the design without which an adequate level of trust cannot be guaranteed, such as fairness, transparency and accountability. Moreover, the implementation of AI systems that take account these factors together also reduces the level of anxiety in the subjects with which you must interface, reducing the level of technical skills required to use them. Finally, forecasting an appropriate form of explanation helps the trust creation process, as it allows a faster and more complete understanding of the decisions made by AI systems.

In the next chapter, we will draw the conclusions of the considerations developed within the thesis

CHAPTER 5: CONCLUSIONS AND FURTHER STUDIES

SECTION 5.1 SUMMARY

The aim of this thesis is to provide an overview of the AI systems used in the field of job recruitment and determine the effect that the XAI has on the confidence of the systems. The use of artificial intelligence systems has become very popular in recent years and new applications are discovered every day. We have identified in the field of job recruitment, an area in which these systems can be of great help, automating the screening processes of personnel. The impact these systems have on people's lives can potentially be very important. We therefore asked ourselves what measures were planned to protect the candidate. We find that all systems are based on three fundamental attributes: fairness, accountability and transparency. The implementation of the XAI allows to increase these attributes, significantly improving the confidence of the model.

SECTION 5.2 FUTURE WORKS

We hope that the literature on the topic proposed in this thesis will grow, as we believe that the use of AI systems in this area can increase significantly in the coming years. The European Commission is also moving in this direction, and in June 2023 it issued an act aimed at regulating the use of AI systems. The aim of the Commission's timely intervention is to guide the design of systems, not to obstruct their use. We hope that on the subject both the technicians of the sector and the common population can take note of the technology behind these systems to be able to interface with them in a safe, free and informed.

BIBLIOGRAPHY

Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli, ‘Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda’, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18 (New York, NY, USA: Association for Computing Machinery, 2018), pp. 1–18 <https://doi.org/10.1145/3173574.3174156>

Adadi, Amina, and Mohammed Berrada, ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, PP (2018), 1–1 <https://doi.org/10.1109/ACCESS.2018.2870052>

Akram, Muhammad Shakaib, Aneela Malik, Mahmud Akhter Shareef, and M. Awais Shakir Goraya, ‘Exploring the Interrelationships between Technological Predictors and Behavioral Mediators in Online Tax Filing: The Moderating Role of Perceived Risk’, *Government Information Quarterly*, 36.2 (2019), 237–51 <https://doi.org/10.1016/j.giq.2018.12.007>

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, and others, ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’, *Information Fusion*, 58 (2020), 82–115 <https://doi.org/10.1016/j.inffus.2019.12.012>

Berente, Nicholas, Bin Gu, Jan Recker, and Radhika Santhanam, ‘Managing Artificial Intelligence’, *MIS Quarterly*, 45 (2021), 1433–50 <https://doi.org/10.25300/MISQ/2021/16274>

Bodria, Francesco, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo, ‘Benchmarking and Survey of Explanation Methods for Black Box Models’, *Data Mining and Knowledge Discovery*, 37.5 (2023), 1719–78 <https://doi.org/10.1007/s10618-023-00933-9>

Chui, Michael, James Manyika, and Mehdi Miremadi, ‘Where Machines Could Replace Humans—and Where They Can’t (Yet)’, 2016

Diakopoulos, Nicholas, ‘Accountability in Algorithmic Decision Making’, *Communications of the ACM*, 59.2 (2016), 56–62 <https://doi.org/10.1145/2844110>

Duan, Yanqing, John S. Edwards, and Yogesh K Dwivedi, ‘Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges and Research Agenda’, *International Journal of Information Management*, 48 (2019), 63–71 <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>

European Commission, ‘Artificial Intelligence Act’, 2023

‘Evaluating Structural Equation Models with Unobservable Variables and Measurement Error - Claes Fornell, David F. Larcker, 1981’ <https://journals.sagepub.com/doi/abs/10.1177/002224378101800104>.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, ‘A Survey of Methods for Explaining Black Box Models’, *ACM Computing Surveys*, 51.5 (2018), 93:1-93:42 <https://doi.org/10.1145/3236009>

Hofeditz, Lennart, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz, ‘Applying XAI to an AI-Based System for Candidate Management to Mitigate Bias and Discrimination in Hiring’, *Electronic Markets*, 32.4 (2022), 2207–33 <https://doi.org/10.1007/s12525-022-00600-9>

Holzinger, Andreas, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, and others, ‘Information Fusion as an Integrative Cross-Cutting Enabler to Achieve Robust, Explainable, and Trustworthy Medical Artificial Intelligence’, *Information Fusion*, 79 (2022), 263–78 <https://doi.org/10.1016/j.inffus.2021.10.007>

Hu, James, ‘Report: 98% of Fortune 500 Companies Use ATS’, *Jobscan*, 2018 <https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/>.

‘Intelligenza Artificiale, Nel 2022 Crescita Del 22%: 700 Milioni Nel 2025 - Il Sole 24 ORE’ <https://www.ilsole24ore.com/art/intelligenza-artificiale-2022-crescita-22percento-700-milioni-2025-AEzvvYkC?refresh_ce=1>.

Jarrahi, Mohammad Hossein, ‘Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making’, *Business Horizons*, 61.4 (2018), 577–86 <https://doi.org/10.1016/j.bushor.2018.03.007>

Lee, ChangHyun, and KyungJin Cha, ‘FAT-CAT—Explainability and Augmentation for an AI System: A Case Study on AI Recruitment-System Adoption’, *International Journal of Human-Computer Studies*, 171 (2023), 102976 <https://doi.org/10.1016/j.ijhcs.2022.102976>

London, Alex John, ‘Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability’, *Hastings Center Report*, 49.1 (2019), 15–21 <https://doi.org/10.1002/hast.973>

Lundberg, Scott M, and Su-In Lee, ‘A Unified Approach to Interpreting Model Predictions’, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), <https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>

Meske, Christian, Enrico Bunde, Johannes Schneider, and Martin Gersch, ‘Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities’, *Information Systems Management*, 39.1 (2022), 53–63 <https://doi.org/10.1080/10580530.2020.1849465>

Meuter, Matthew L, Amy L Ostrom, Mary Jo Bitner, and Robert Roundtree, ‘The Influence of Technology Anxiety on Consumer Use and Experiences with Self-Service Technologies’, *Journal of Business Research*, Strategy in e-marketing, 56.11 (2003), 899–906 <https://doi.org/10.1016/S0148-2963(01)00276-4>

Miller, Tim, ‘Explanation in Artificial Intelligence: Insights from the Social Sciences’, *Artificial Intelligence*, 267 (2017) <https://doi.org/10.1016/j.artint.2018.07.007>

Mitchell, Tom M., *Machine Learning*, McGraw-Hill Series in Computer Science (New York: McGraw-Hill, 1997)

Mujtaba, Dena F., and Nihar R. Mahapatra, ‘Ethical Considerations in AI-Based Recruitment’, in *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 2019, pp. 1–7 <https://doi.org/10.1109/ISTAS48451.2019.8937920>

Peña, Alejandro, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, and others, ‘Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI-Based Recruitment’, *SN Computer Science*, 4.5 (2023), 434 <https://doi.org/10.1007/s42979-023-01733-0>

Rouse, William B., ‘AI as Systems Engineering Augmented Intelligence for Systems Engineers’, *INSIGHT*, 23.1 (2020), 52–54 <https://doi.org/10.1002/inst.12286>

Sendak, Mark, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, and others, ‘“The Human Body Is a Black Box”: Supporting Clinical Decision-Making with Deep Learning’, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20 (New York, NY, USA: Association for Computing Machinery, 2020), pp. 99–109 <https://doi.org/10.1145/3351095.3372827>

Simon, Herbert A., *Models of Man; Social and Rational*, Models of Man; Social and Rational (Oxford, England: Wiley, 1957), pp. xiv, 287

Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education Limited. (Pearson, 2016)

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, Fermín Moscoso del Prado Martín, *Using ontologies to enhance human understandability of global post-hoc explanations of black-box models*, “Artificial Intelligence”, 2021.