D7.2

Preliminary description of AI systems security mechanisms and tools

Project number	830892
Project acronym	SPARTA
Project title	Strategic programs for advanced research and technology in Europe
Start date of the project	1st February, 2019
Duration	36 months
Programme	H2020-SU-ICT-2018-2020

Deliverable type	Report
Deliverable reference number	SU-ICT-03-830892 / D7.2 / V1.0
Work package contributing to the deliverable	WP7
Due date	July 2020 – M18
Actual submission date	31st July, 2020

Responsible organisation	ITTI
Editor	Marek Pawlicki
Dissemination level	PU
Revision	V1.0

Abstract	This deliverable will present the first version of the defensive and reactive mechanisms, explainability enhancing mechanisms and fairness ensuring mechanisms. The information about the AI Program contests will be provided.
Keywords	Adversarial Attacks, Machine Learning, Secure AI, Explainability, Fairness

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

Editor

Marek Pawlicki (ITTI)

Contributors (ordered according to beneficiary numbers)

Zakaria Chihani (CEA)

Jean-Marc Van Gyseghem, Manon Knockaert (UNamur)

Mohammad Reza Norouzian (TUM)

Erkuden Ríos, Mª Carmen Palacios (TEC)

Raúl Orduña, Xabier Etxeberria, Amaia Gil (VICOM)

Vincent Thouvenot, Boussad Addad, Simon Grah (TCS)

Michał Choraś, Marek Pawlicki, Damian Puchalski, Mateusz Szczepański, Rafał Kozik (ITTI)

Reviewers (ordered according to beneficiary numbers)

Vivek Nigam (FTS)

Regina Valutytė (MRU)

Executive Summary

This document is the result of the first 18 months of work of the SAFAIR program, also known as SPARTA WP7. The program, and the document which follows seeks to address two major contemporary problems caused by the proliferation of artificial intelligence (AI). The first objective is to ensure the security of AI solutions, the second objective deals with the issue of trustworthiness and fairness of AI.

To this end, the program produces formalisms, methods and tools which make current AI algorithms more robust, ensuring their security, reliability, and privacy. The preliminary descriptions of those tools and methods are offered in chapter 2 of this document. While the threat analysis conducted by the SAFAIR program made its way to the sister document, D7.1, this work holds the information regarding the defensive and reactive mechanisms designed to ensure resilience against the new, complex cyber-threats identified in D7.1. The mechanisms and tools proposed by the SAFAIR program seek to optimize resiliency without compromising the advantages provided by AI and aiming not to affect the performance of AI.

Secondly, the program seeks to extend the explainability of AI, which usually comes in the form of a black-box. This provides both the verification of functionality of AI and gives personnel relying on decisions made by AI the ability to trace back those decisions to the inputs, paving the way to better understanding of AI, increasing the transparency of AI and increasing the confidence in AI. The proposed explainability mechanisms and tools can be found in chapter 3.

The third aspect of the abovementioned objectives deals with the legal and societal aspects of trustworthiness and fairness of AI. The third task of SAFAIR seeks to provide mechanisms to reduce conscious and unconscious bias in AI decisions, ensuring functionality that does not introduce discrimination, and finding ways to grant both credibility and correct compliance with relevant legislative regulations. The current state of that work is contained in chapter 4.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Recent advances in machine learning (ML) and the surge in computational power have opened the way to the proliferation of artificial intelligence (AI) in many domains and applications.

Still, many of the ML algorithms offered by researchers, scientists and R&D departments focus only on the numerical quality of results, high efficiency and low error rates (such as low false positives or low false negatives). However, even when such goals are met, those solutions cannot (or should not) be realistically implemented in many domains, especially in critical fields or in the aspects of life that can impact whole societies, without other crucial criteria and requirements, namely: security, explainability and fairness. Moreover, the outstanding results are frequently achieved on data that is well-prepared, crafted in laboratory conditions, and are only achievable when implemented in laboratory environments.

However, when large scale applications of AI became reality, the realization came that the security of machine learning requires immediate attention. Malicious users, called 'adversaries' in the AI world, can skilfully influence the inputs fed to the AI algorithms in a way that changes the classification or regression results. Regardless of the ubiquity of machine learning, the awareness of the security threats and ML's susceptibility to adversarial attacks used to be fairly uncommon and the subject has received significant attention only recently.

Apart from security, another aspect that requires attention is the explainability of ML and ML-based decision systems. Many researchers and systems architects are now using deep-learning capabilities (and other black-box ML methods) to solve detection or prediction tasks. However, in most cases, the results are provided by algorithms without any justification. Some solutions are offered as if it was magic and the Truth provider, while for decision-makers in a realistic setting the question why (the system arrives at certain answers) is crucial and has to be answered.

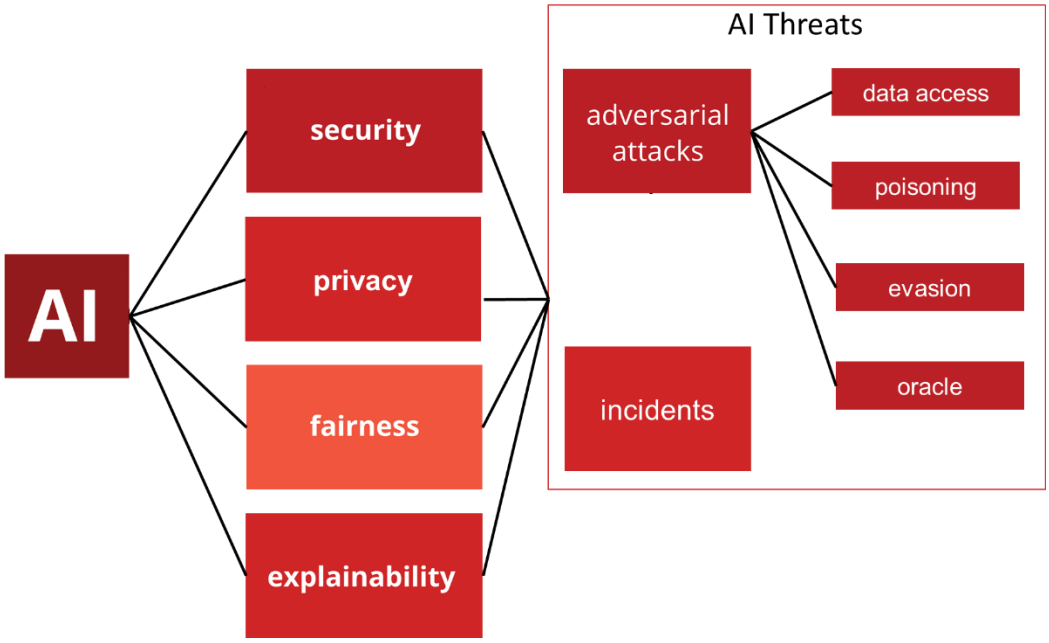


Figure 1: Contemporary artificial intelligence concerns and threats

This document delves into aspects and recent works on security protections, explainability, and fairness of AI/ML systems, and provides the expansion of those works, as well as offers new approaches. The overview of those matters can be found in Figure 1.

The descriptions of the tools and methods geared towards increasing the security of AI are offered in chapter 2. So far, with regard to designing defensive and reactive measures, the SAFAIR program tested the most widely used attacks (Fast Gradient Sign Method, Basic Iterative Method, Projected Gradient Descent, Carlini & Wagner attack), then proposed or implemented the following defensive mechanisms and tools:

- Prediction Similarity (novel approach developed under SAFAIR)
- Adversarial (re)Training (2 approaches, 1 novel approach developed under SAFAIR)
- Dimensionality Reduction (3 approaches)
- Neural Activation-based Adversarial Attack detector (novel approach developed under SAFAIR)

The threat analysis carried out by the SAFAIR program is included in the sister document, D7.1. The mechanisms and tools proposed by the SAFAIR program explore the ways to optimize resiliency without sacrificing the benefits catered by AI and aiming not to affect the effectiveness of AI.

At the same time, the program pursues the ways to broaden the explainability of AI, which usually comes in the form of a black-box. This part of the program supplies both the verification of the functionality of AI and provides personnel relying on decisions made by AI the ability to trace back those decision to the inputs, opening the doors to improved understanding of AI, increasing the transparency of AI and increasing the confidence in AI. The proposed explainability mechanisms and tools can be found in chapter 3.

With regard to explainability and fairness enhancing mechanisms proposed by the program, this document contains the descriptions of:

- Oracle-Explainer approach (developed under SAFAIR)
- Fair Adversarial Network and Explainability approach (developed under SAFAIR)

The third aspect of the program addresses the legal and societal aspects of trustworthiness and fairness of AI. SAFAIR seeks to provide mechanisms to reduce conscious and unconscious bias in AI decisions, guaranteeing functionality that does not insert discrimination into the decision-making process. This part of the program also focuses on finding the ways to provide credibility and correct compliance with relevant legislative regulations. The current state of that work is contained in chapter 4.

Chapter 2 Defensive and reactive mechanisms

2.1 Introduction

Adversarial machine learning attacks are one of the major threats against AI systems, and very particular to them, since they exploit the intrinsic nature of the AI systems' learning methods. For example, recently it has come to attention that skillfully crafted inputs can affect artificial intelligence algorithms to sway the classification results in the fashion tailored to the adversary needs. This new disturbance in the proliferation of machine learning has been a subject riveting attention of the researches very recently, and at the time of writing this document a variety of vulnerabilities have been uncovered.

With the recent spike of interest in the field of securing ML algorithms, a myriad of different attack and defense methods have been discovered; no truly secure system has been developed however, and no genuinely field-proven solution exists. The solutions known at this point seem to work for certain kinds of attacks, but do not assure protection against all kinds of adversarial attacks. Indeed, in certain situations, implementing those solutions could lead to the deterioration of ML performance. Therefore, defensive mechanisms for AI systems against such threats need to be carefully planned as part of the system design.

In this section, we propose a collection of defensive and reactive mechanisms that can be applied in AI systems to ensure secure behaviour and protection against some adversarial attacks. The selection of these protection measures is the result of the analysis and testing of the most well-known and widely used attacks including Fast Gradient Sign Method, Basic Iterative Method, Projected Gradient Descent, Carlini & Wagner attack, for which we have evaluated and designed the corresponding protections in SAFAIR. A description of such attacks can be found in the deliverable D7.1 by SAFAIR.

2.2 Measures to ensure adequate functionality of AI systems

Imagine that an adversary is attacking our artificial intelligence system. Our next step is to know what type of defences are developed and which attacks can be avoided by the defences. Therefore, in the following, we document useful defences that we have analysed within SAFAIR.

A good defence strategy involves the design and selection of defences targeted to avoid incidents and attacks exploiting system vulnerabilities. Therefore, it is necessary to first identify where these vulnerabilities may reside in the AI system and in which phase or step of the AI system life-cycle they may become an attack vector or a potential issue. A simplified architecture of an AI system is shown in Figure 2, where two major activities or steps of the AI system are outlined and depicted as rectangles, i.e. Training and Testing (Inference or Prediction), and four datasets (shown as repositories) are identified: i) *Training data* used to train the learning system, ii) *Trained model*, the model already trained and tested, iii) *Operational data* or *Input data* used in AI system operation when it is already deployed and running, and iv) *Results*, the outcome of the application of the model on top of Input data, i.e. the result of the prediction process.

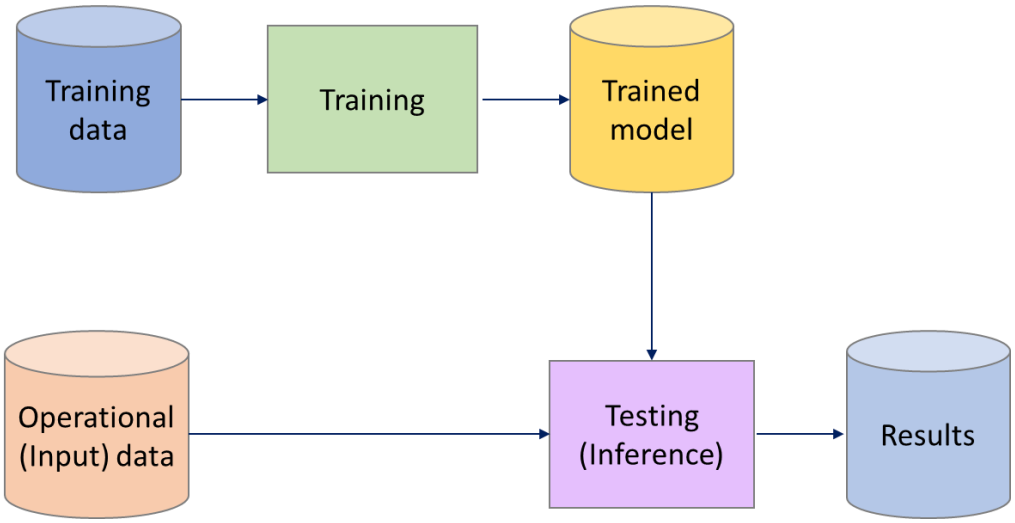


Figure 2: Simplified Machine Learning system diagram

Thus, we can classify the AI system defences by mapping them to threats in the threat model developed in deliverable D7.1 of SAFAIR. In the following, we provide the defence classification with respect to the four types of attacks identified in the SAFAIR threat model and other unintentional flaws or incidents identified therein.

2.2.1 *Protections against Data Access attacks*

These defences include all protection methods against adversaries extracting Training data. Usually, these defences are based on applying access control (identification, authentication and authorisation) mechanisms to avoid unauthorised access to the Training dataset. Data encryption can also be used as a means of protecting the data from adversaries. In general, this type of protections are not specific to AI systems.

2.2.2 *Protections against Poisoning attacks*

This type of attack takes advantage of the models' vulnerabilities generated during the training. These defences are used against vulnerabilities that can be found in Training data.

- **Accuracy:** Let us suppose that the model that we attempt to defend is being retrained when new data is available. If we want to control our model, we could calculate its accuracy constantly. In case of excessive accuracy fall, we can suspect that something is wrong.
    - **Poisoning:** By checking the accuracy of the model, this attack can be avoided. If the accuracy of the model is drastically reduced, it could be due to introduction of malicious data in the training dataset used in the retraining.

- **Loss Function:** In this case, the model continues being retrained constantly. One way to control the model is to monitor the values that the loss function is taking. In case of these values having a meaningful increase, this could imply suspicious data in the training set. This defence is similar to accuracy-based defence, since accuracy and loss function are inversely related.
    - **Poisoning:** By checking the values of the loss function of the model, this attack can be avoided. If the loss function value of the model has drastically increased, it could be due to malicious data.

2.2.3 *Protections against Evasion attacks*

Different protection measures can be used against evasion attacks that target input Operational data used in the Prediction (Inference) phase. Some defences aim to minimise vulnerabilities in the

- **Dimensionality reduction:** This defence passes the Input data through a dimensionality reduction process to remove as much noise as possible. As a result, the model is able to generalize, avoiding adversarial examples.
    - **Adversarial attack:** Adversarials are generated by an addition of noise to the original data that can be imperceptible for the humans. Hence, any process to remove as much noise as possible could be useful to avoid the success of this attack.

- **Adversarial Train:** This defense retrains the targeted model with additional training data constructed from discovered adversarial attacks (training data + adversarial examples). Once the model has been retrained, it learns to classify well the adversarial examples introduced in the training; in other words, they are not adversarial examples anymore.
    - **Adversarial attack:** As known adversarial examples are inserted in the Training data, when the model is retrained with the new Training data, it likely corrects these adversarial mistakes.

- **Distribution:** First, historical data about which input Operational data the machine learning model is using and which predictions are obtained by the model from these inputs is saved. Then, it is possible to check in this history the distribution of the wrongly predicted results and compare them with validation data. This defence is useful against:
    - **Trojaning attack:** If both distributions are compared, it is observed that in the saved history data distribution one of the outputs has the majority, while in the validation data distribution it does not.

2.2.4 *Protections against Oracle attacks*

- **Rounding:** This defence consists in modifying model parameters by rounding them. With this methodology the attacker obtains modified parameters and not the real ones.
    - **Hyperparameter Stealing:** Confidence scores of predictions are obfuscated making it more difficult for the attacker to detect real model hyperparameters. Because of that, Rounding can be an effective countermeasure against this attack.

- **Feature Importance:** Each feature has a different importance value for the machine learning model. By modifying these importance values, a new model is obtained and this can work against some attacks:
    - **Inversion attack:** As this importance may affect the accuracy of the model, then also the accuracy of the attack might be affected when the defence is applied.

- **Gradient:** This defence consists in degrading the quality or precision of the gradient information retrievable from the model. Even if this is not always possible, reducing the precision at which confidence score is reported can help with the degradation needed. This defence is useful against:

    - **Inversion attack:** This attack could be based on gradient descent. Because of that, this defence could be effective in this case.

- **Regularization:** Regularization techniques are used to overcome overfitting in machine learning. This defence is effective against the attacks that make use of overfitting. One of those attacks is:

    - **Inference attack:** This attack uses overfitting problem for its own benefit and therefore regularization can help avoiding it.

- **Backdoors:** Backdoors are not seen as a defensive mechanism, but in some specific cases can be useful in order to protect our machine learning system. The backdoors are used to create a set of new instances that later can be introduced on the training set. A random classification class from the original model is assigned to those new instances. Then, the model is retrained with this modified training set. We could apply this defence against:

    - **Robbery:** We may apply backdoors for legitimating our models so one could verify the model copyright. However, this defence is not absolutely safe. There are different possibilities to avoid this verification.

2.2.5 *Protections against Privacy threats*

The following privacy measures that avoid leaking private data are part of defences that can be adopted against Oracle inversion threats.

- **Sanitization:** This defence consists in processing sensitive and private data. For example, we may construct a blacklist with sensitive and private data and remove them from the training data. It could be interesting against some attack:

    - **Training Data Extraction attack:** It would be interesting to apply Sanitization in this attack. However, one cannot hope to guarantee that all the possible sensitive data will be found and removed through such a blacklist.

- **Differential Privacy:** This defence protects results against unintentional disclosure of potentially sensitive information related to a single record of a database. It achieves this by maximizing the accuracy of queries from statistical database and minimizing the ability to identify single records.

    - **Training Data Extraction attack:** Differential privacy uses statistics of the database to maximize the accuracy, minimizing the ability to identifying single records. Therefore, this attack will be less successful, because it takes advantage of these single records.

2.2.6 *Application of Formal Methods to AI robustness*

While engineering methods such as the ones presented in the previous section have undoubtedly shown their relevance, the field of Formal Methods, which has been at the heart of (mainly critical) software trust for many decades [1], should not be considered irrevocably extraneous to AI security, safety and robustness. In the history of computer science, one such claim was made [2] for traditional software. This effectively stalled research in an area that desperately needed it, and that ultimately proved highly beneficial to safety-critical systems, saving lives and money.

In the field of machine learning in particular, several efforts to apply Formal Methods to AI trust are already taking place in various places of the world, such as Israel [9], the US [10] and several teams in Europe [4][11][12]. Among the aspects targeted by this research, there are the following aspects of a learning-enabled system:

- satisfaction of safety properties, *i.e.,* "does my system satisfy the formal specification of its behaviour?", see for example [13]. This type of properties has the particularity of being dependent on the domain of application (*e.g.,* a car that has certain reactivity to sensors, an avoidance system that enacts the right avoidance strategy, etc.)
- increased sensitivity, *.i.e.,* "does my system behave the same knowing that the information received from the sensors or the environment may vary with a given noise factor"
- resistance to attacks, *i.e.,* "can my system resist some attack scenario"

The first two are safety properties: these relate to internal aspects of a system (as is the case for explainability, or fairness); in other words, they are used to certify that the system is able to do some specified job in nominal conditions of the environment. The third one is a security property: it aims at ensuring that the system can withstand a targeted attack from a malicious and able agent with a given level of resources. For the recent advances in machine learning, the last two properties overlap significantly in one particular aspect: they both relate to the resistance to perturbation around inputs. The difference lies in the type of perturbation:

- For safety, one would take into account the sensitivity of the sensors or the distortions coming from the environment (such as light flares or dust). Such perturbation can take familiar shapes, such as Gaussian noise, or can be derived by a deep understanding of the environment in which the AI system will evolve, and the inherent physical properties of the hardware.
- For security, the perturbations are more specific and greatly depend on the capabilities of the attackers. In particular, most of the state-of-the-art consists in proving resistance to all perturbations within a certain limit around inputs. An attacker capable of perturbing outside that limit is thus not defended against. This is a usual, probably unsurmountable, limitation of security. There is no 100% secure software, as Apple learned when Google revealed that their claimed "unhackable iPhone" was very much hackable.

Unsurprisingly, most academic tools developed with the aim of robustness or safety are susceptible to criticism on the basis that they do not scale. Here again, the history of computer science shows that dismissing these methods solely on the basis of scalability (which is an otherwise acceptable argument) seems short-sighted. After all, the solutions to the Boolean satisfiability problem did not scale before the techniques such as conflict-driven clause learning, non-chronological backtracking, two-watched literals and symmetry breaking. A problem of theoretical NP-complexity[1] does not mean any such problem is unsolvable. It is not out of the realm of the imaginable that investing research in the study of machine learning's inner workings can yield efficient heuristics that can bring considerable speed-ups for a relevant subset of trust-ensuring properties.

And, indeed, the increasing efficiency in this very young field of research is encouraging. Every year brings new tools or improved versions of tools that take us closer to their practical usage. One

[1] Very complex problems

of their main limitations, for example, is in the type of activation functions that they can handle. Indeed, with notable exceptions, they can only deal with piece-wise linear functions, such as ReLU [3][4][5]. Other popular functions, such as sigmoid, have little support. One should keep in mind such aspects when one needs to certify a learning-enabled system. Indeed, the verification and validation discipline has always involved a back and forth between the designers of the systems and its ensurers. The latter can sometimes ask that the code satisfies certain shape constraints (e.g. no jumps or no nested loops). Verification and Validation is a game of balance, of trade-offs between efficiency and verifiability. Similarly, the ML designers should be aware of the state of the art formal methods. For example, if there is no strong preference between ReLU and sigmoid, they should implement the former.

A recent research paper [6] studies extensively 17 formal methods tools, based on numerous methods such as reachability and optimisation. Benefiting from the collaboration of the authors of the respective tools, they were able to put together a detailed analysis of their assets and weaknesses. They also provide pedagogical implementations of all the methods in Julia, making them easier to use for tutorials and as a support for teaching material. Finally, some efforts consider the trust problem at an *earlier* stage. Indeed, several investigations in the robustification of the AI system training itself through formal methods have recently appeared, such as [7] using differentiable Abstract Interpretation and [8] using differentiable logic.

2.3 Designing robust AI tools resilient to adversarial attacks

2.3.1 *Defending Network Intrusion Detection Systems against Adversarial Evasion Attacks*

This section focuses on the overall approach to Evasion Attack Detection in neural networks. Firstly, the utilised dataset is disclosed, which is followed by the applied dataset pre-processing and the IDS training pipeline. Then the attacks on the IDS are performed and tested. The neural activations of those attacks, as well as the activations for clear samples are gathered; finally, the attack detector is trained and tested.

2.3.1.1 Intrusion Detection based on an artificial neural network

In this work, the CICIDS2017 dataset was used [14]. The dataset was first cut into four parts in a stratified fashion to ensure full coverage of all kinds of attacks included in the dataset in all its sub-parts. This procedure results in the following setup:

- Dataset A - used to train the IDS classifier
- Dataset B - used to test the IDS classifier and to craft the adversarial attacks and test them on the original IDS ANN, then to acquire the activations of neural nodes in the IDS network of benign, attack and adversarial samples to train the Adversarial Detector
- Dataset C and D - used to craft test adversarial samples and acquire the activations for the neural nodes of benign, attack and adversarial samples

All the sub-parts were then turned into a binary classification task, leaving all the benign samples as 'BENIGN', but changing all the names of possible attacks to simply 'ATTACK'. The utilisation and partitioning of CICIDS2017 is depicted in Figure 3.

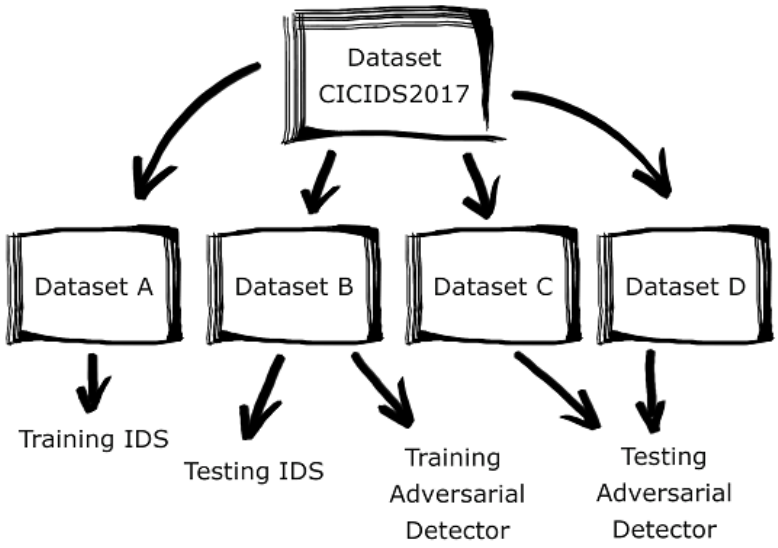


Figure 3: The utilisation and partitioning of CICIDS17 for training and testing of IDS ANN and the Adversarial Detector

The IDS setup was as follows: an artificial neural network of 3 hidden layers was compiled, with 40 neurons on the first hidden layer, 40 on the second and 20 on the third layer. The Rectified Linear Unit activation function was utilised and the optimiser selected was ADAM. With batch size of 100 and 10 epochs, the network achieved an accuracy of 0.9827 when trained with Dataset A and tested on Dataset B. The precision, recall and f1-score are showcased in Figure 4.

	precision	recall	f1-score	support
ATTACK	0.96	0.97	0.97	139675
BENIGN	0.99	0.99	0.99	405383
micro avg	0.98	0.98	0.98	545058
macro avg	0.97	0.98	0.98	545058
weighted avg	0.98	0.98	0.98	545058
samples avg	0.98	0.98	0.98	545058

Figure 4: IDS ANN trained on Dataset A and tested on Dataset B

The pipeline of the IDS training/testing process is showcased in Figure 5.

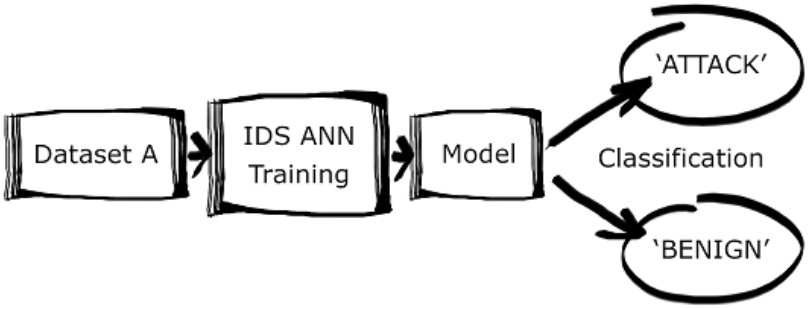


Figure 5: The IDS ANN pipeline

As seen in the figure, the binarized dataset is fed to the architecture described above, and the training procedure results in building a model capable of binary classification.

2.3.1.2 Adversarial Attacks

After testing the trained IDS (the optimisation procedure of an ANN-based IDS can be found in [15]), four different adversarial attacks were crafted based on the ATTACK class of Dataset B. The algorithms used for the creation of evasion attacks were:

- Carlini and Wagner attack (CW) [16]
- Fast Gradient Sign Method (FGM) [17]
- Basic Iterative Method (BIM) [18]
- Projected Gradient Descent (PGD) [19]

1397 samples of the 'ATTACK' class were randomly extracted from Dataset B and turned into adversarial samples with the use of those four algorithms. The IDS ANN classified those as 1353 ATTACKS and 44 BENIGNS.

Using adversarial attacks, we were able to force the IDS ANN to classify 1296 'ATTACK' as 'BENIGN' samples for BIM and PGD, 1324 for FGM and 59 for CW.

The abovementioned procedures introduce adversarial noise to the samples. This noise resulted in negative values in some of the features. Those negative values were supplanted by zeros with some loss of effectiveness of those attacks - 55 more attacks were classified as benign samples with the original BIM and PGD methods, 295 more for CW and interestingly, 21 fewer for FGM.

The samples from Dataset B not used in crafting Adversarial Attacks were annotated as 'nonadversarial', the Adversarial Attacks were labelled 'adversarial'. With 5588 adversarial attack samples, a matching number of nonadversarial records was randomly picked from unused samples of Dataset B to form the base for a balanced 'Adversarial Training Dataset' for the adversarial attack detector. The procedure is depicted in Figure 6.

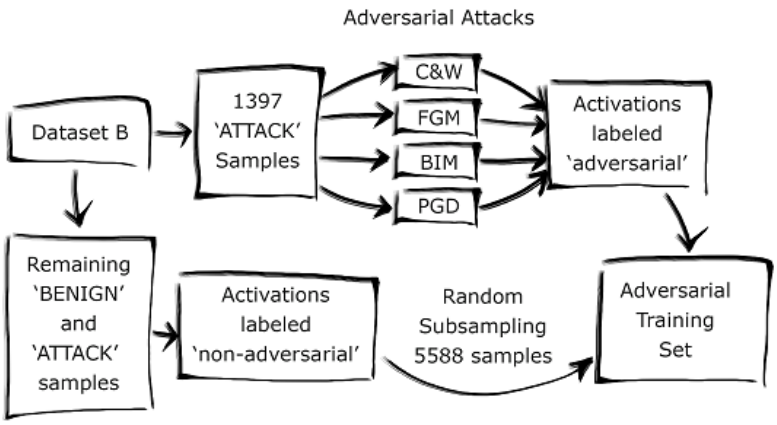


Figure 6: Forming the Adversarial Training Dataset from Dataset B

Dataset D was subjected, except for the balancing, to the exact same crafting/annotation procedure to form the base for the testing dataset for the detector.

2.3.1.3 Detection Method

The training and testing activation datasets were fed to the IDS ANN and the activations for all the 102 neurons (including the softmax layer), as shown in Figure 7, were recorded and annotated as adversarial or nonadversarial respectively.

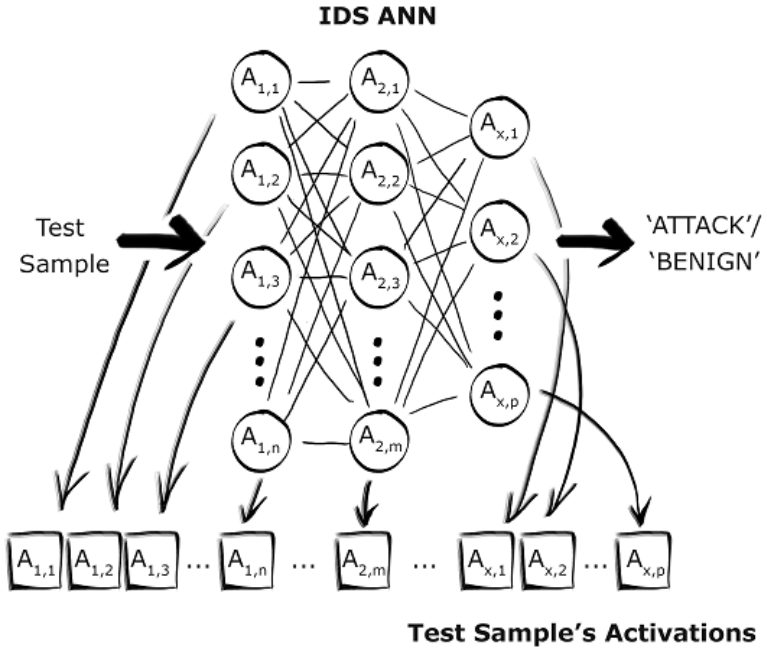


Figure 7: The acquisition of IDS ANN activations for a given test sample

The recorded activations were used to train the detector artificial neural network. The architecture of the detector is as follows: 3 hidden layers with the ReLU activation function, 51, 51 and 25 neurons respectively and the ADAM optimiser. The training / testing pipeline is presented in Figure 8.

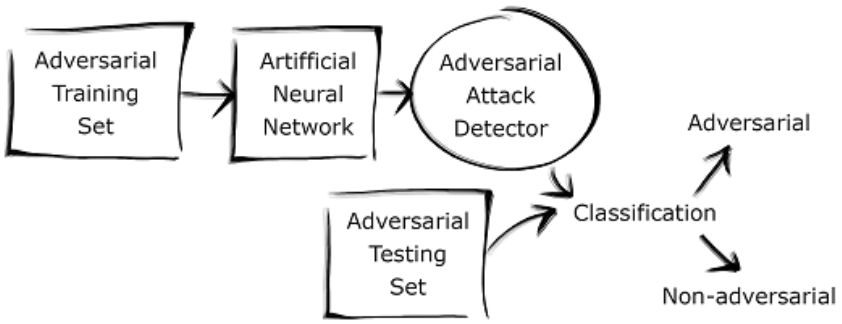


Figure 8: The Adversarial Detector Training/Testing Pipeline

Using batch size of 100 and just 10 epochs, the detector achieved an accuracy of 0.8506 on the testing set. The detailed results are assembled in Figure 9

	precision	recall	f1-score	support
adversarial	0.06	0.91	0.11	5588
non-adversarial	1.00	0.85	0.92	543661
micro avg	0.85	0.85	0.85	549249
macro avg	0.53	0.88	0.51	549249
weighted avg	0.99	0.85	0.91	549249
samples avg	0.85	0.85	0.85	549249

Figure 9: Results of ANN-based Adversarial Attack Detector over the test set activations

2.3.2 *Defending Healthcare Images against Adversarial Evasion Attacks*

In this section, several defences against adversarial attack over healthcare images will be presented. Each defence will have its pros and cons. During the whole study, the model used was formed by a VGG16 (pre-trained CNN) and DNN.

2.3.2.1 Healthcare Image Dataset

In this study, a dataset of breast cancer images was used, but could be generalized to other classification task. This dataset may be found in:

"https://www.kaggle.com/paultimothymooney/breast-histopathology-images"

This dataset is composed of the images (50X50) of Invasive Ductal Carcinoma (IDC), which is the most common subtype of all breast cancer. Each image has associated one of two classes, where the class 0 is non-IDC and the class 1 is IDC. Moreover, this dataset was divided in three subgroups: training data, validation data and test data.

2.3.2.2 Adversarial Attacks

The adversarial attack has been widely studied in deep learning. Taking advantage of the sensitivity of the models, the attacker adds noise to a specific input sample, modifying the image imperceptibly to change the original output prediction of the sample. In this case, the algorithms used to generate these adversarial examples were:

- Fast Gradient Sign Method (FGM) [17]
- Basic Iterative Method (BIM) [18]
- Projected Gradient Descent (PGD) [19]

2.3.2.3 Defence Methods

In order to avoid the success of the mentioned adversarial attacks, four different defences were implemented, namely, adversarial train, dimensionality reduction, prediction similarity and model's behaviour.

2.3.2.3.1 Adversarial Train

This defence retrains the targeted model with the training data, once the adversarial examples have been added, so it learns to classify them correctly. This idea was introduced in [20].

Adversarial training is a widely used defence against adversarial attacks and it has improved over time. However, it has not achieved competitive robustness against new adversarial examples once the model is retrained [21].

In this study, after the adversarial example dataset (adversarial dataset) was created, a basic adversarial train was implemented, retraining the mentioned model with new training data, formed by original training data and adversarial dataset.

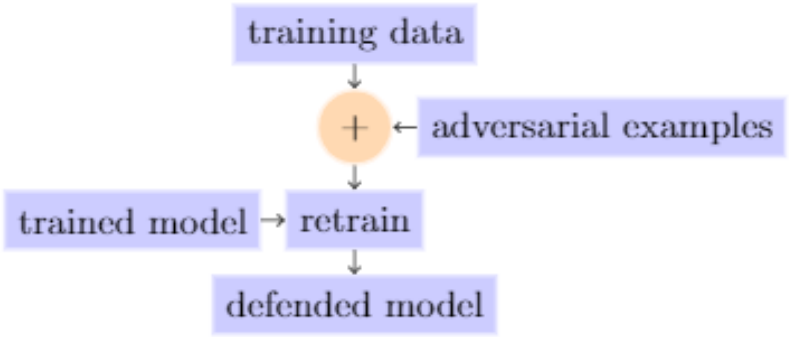


Figure 10: Standard adversarial training.

Although the new model is more robust against the added adversarial examples, it is easy to obtain new ones and stay in an inexhaustible circle of attacking and defending, as can be seen in Table 2.

2.3.2.3.2 Dimensionality Reduction:

This defence can be implemented in several ways with different effectiveness in strengthening the model. However, all the variants have the same idea behind: passing data through a dimensionality reduction layer (autoencoder and encoder layers, in our case) to remove as much noise as possible from the input image. Thus, the model is able to generalize, avoiding adversarial examples. As a machine learning model may use this initially available noise in the input as features for prediction, as a trade-off, this generalization could imply a reduction in model's accuracy. That is the reason why balance between the robustness and the accuracy must be met.

Based on reference [22], dimensionality reduction may be useful to make the targeted model more robust against adversarial examples. For the case of deep learning, CNNs and autoencoders are used to carry out the dimensionality reduction. Particularly, it is known that the autoencoders might make the model stronger against adversarial examples [23]. Among the dimensionality reduction bibliography, there are Principal Component Analysis (PCA) [22] and autoencoder implementations [23] in DNNs.

In this study, three variants of dimensionality reduction ((1) middle autoencoder, (2) encoder and (3) initial autoencoder) were covered in this subsection, which were based on the same idea, but the returned defended model was different.

The middle autoencoder variant was obtained by training an autoencoder using CNN features, that is, once the outputs of data were obtained through CNN (VGG16 in this study), an autoencoder was trained using these outputs. After the autoencoder was trained, it was inserted before the DNN. In this case, the CNN and DNN were maintained with the original structure (original weights), so they were not retrained. In short, the middle autoencoder "cleans" the noise of CNN's outputs before using them as DNN's input data.

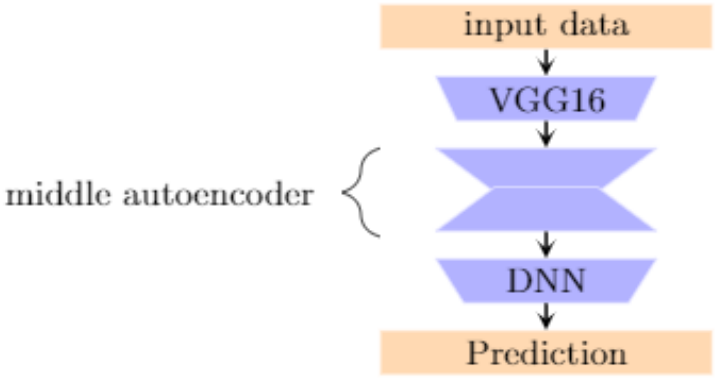


Figure 11: Middle autoencoder model.

The encoder variant was obtained by taking the encoder part of the middle autoencoder. Then, a new model was built by inserting it between the initial CNN and a new DNN. The new DNN was trained with the encoder's output as input data and outputting the initial classes. This defence differs from the others, because the encoder trains a new DNN so the structure of the model changes. As a summary, the encoder reduces the dimensionality of DNN's features, erasing the least important ones to avoid noise.

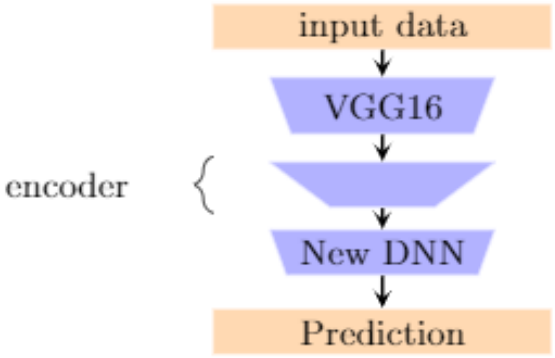


Figure 12: Encoder model.

The initial autoencoder variant trained the autoencoder using the selected dataset and inserted it before the CNN. Both the CNN and DNN keep the original weights, since they were not retrained. Again, the initial autoencoder "cleans" the image noise before making predictions with the initial model.

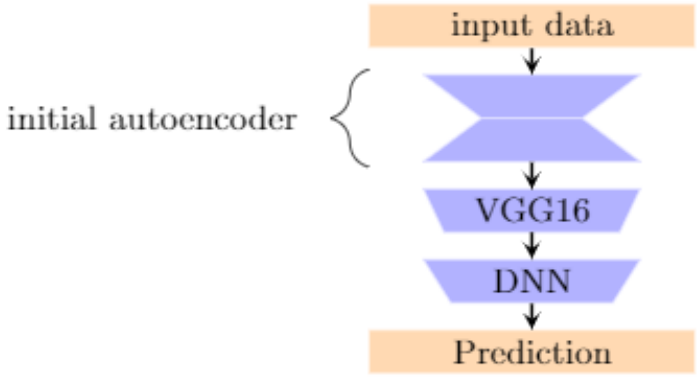


Figure 13: Initial autoencoder model.

In the case of having a trained model, the autoencoder variants could be better for their use as defence, because no parts of the original model need to be retrained. However, the encoder variant showed that a model that originally contains an encoder layer could add robustness, by using it as an adversarial detector in parallel. Therefore, two different predictions (the original one and the one created by the defended model) were obtained. In case of different pre-dictions, the model used as defence could detect a possible adversarial example.

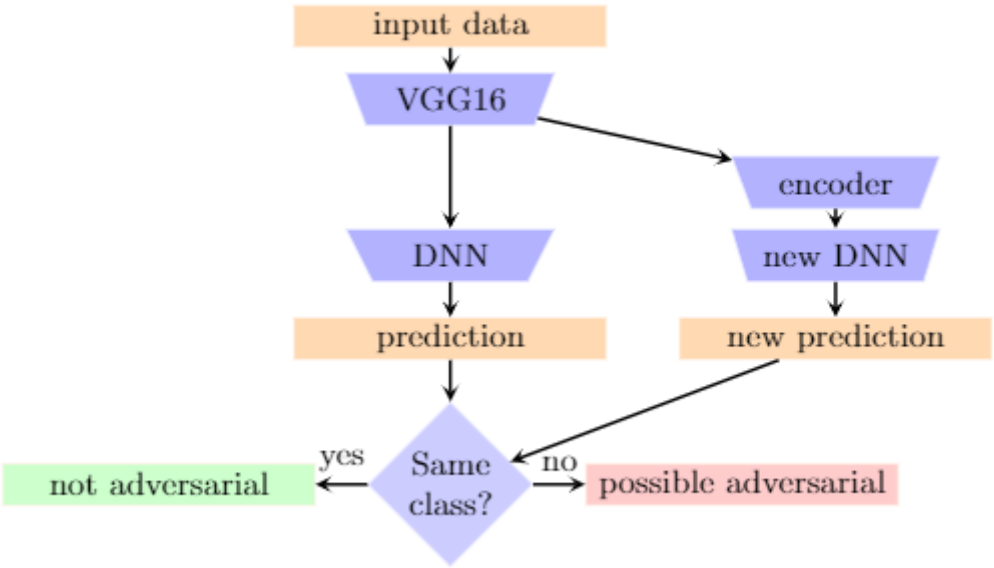


Figure 14: External detector of adverse examples using the encoder.

2.3.2.3.3 Prediction Similarity

This defence adds an external layer to the model, which saves the history of parameters obtained through the input images. Adversarial attacks need several predictions of similar images to get an adversarial example. Therefore, this layer can return an adversarial probability (likelihood that it is occurring), after computing the similarity between the input image and previous images. If this adversarial probability is high (different from case to case) this layer could take action to avoid the adversarial attack.

There are several algorithms to compute the similarity value between two images. The most widely used metrics are the mean squared error (MSE) and peak signal to noise ratio (PSNR). However, in the last three decades, different complex metrics have been developed trying to simulate the perception of human vision by comparing two images [24], i.e. structure similarity metric (SSIM) [25] and feature similarity metric (FSIM) [26].

In this study, these parameters were user, image, prediction value (the class and the probability of this class), minimum distance (to all previous images), prediction alarm (number of times the percentage of the class is smaller) and distance alarm (number of images with distance less than threshold). There were different possible actions that the output layer could take, such as blocking or predicting with a secondary model. In this case, if the layer detected something suspicious, it returned the opposite (or another) class. Thus, if the adversarial attack was detected, this action automatically avoided it, since it would return another class. This made the adversary believe that they had already achieved the adversarial example, when in fact they had not.

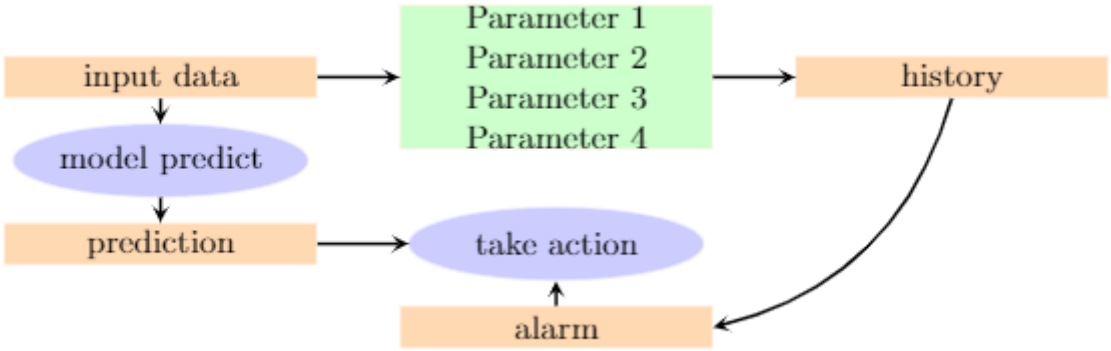


Figure 15: Generalization of the prediction similarity defence.

2.3.2.3.4 Model's behaviour

For this defence, an adversarial detector is generated behind the idea that the model's behaviour changes depending on if the input is an original image or an adversarial one.

For model's behaviour characterization, each neuron's output value is used to generate features. As actually used deep learning architectures have a very large quantity of neurons, different aggregate functions have been selected for reducing dimensionality. The selected aggregated functions are applied in each layer, obtaining a new feature per layer and aggregated function used.

The aggregated functions were selected in order to obtain distributive information on how the neurons of a specific layer act. Considering that, the selected aggregated functions are used to study the centrality (mean and median), asymmetry (skewness, kurtosis), dispersion (std and iqr), limit values (min and max) and number of activated neurons (count of non-zero output values). In our case of study, as we have 19 hidden layers and 9 aggregated functions, we end up obtaining 171 features.

After feature standardization and selection, two detectors were trained, one per original model predicted class. For both detectors, support vector classification algorithm was used, obtaining 94% accuracy in both predicted classes.

The detectors could be used as an external layer like prediction similarity defence, obtaining different defence mechanism depending on the action taken.

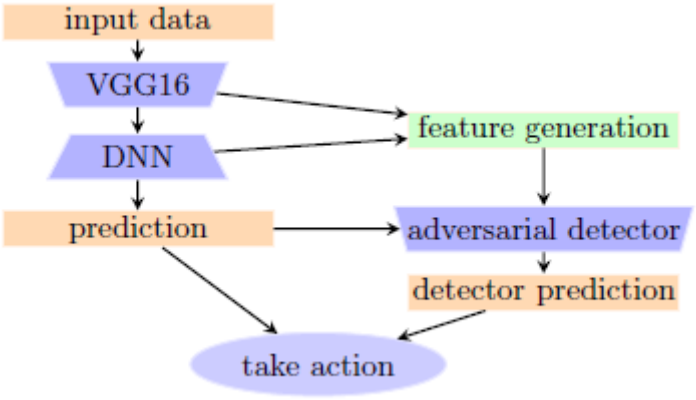


Figure 16: Detector based on model's behaviour

2.3.2.4 Proof of Concepts

Once each defended model was developed, there are two essential characteristics of each model to take in account: the accuracy and the success avoiding adversarial examples (robustness). The balance between them is necessary because a ML model must be robust without losing applicability.

On the one hand, this work studied the accuracy of each defended model and the original model. In general, defended model's accuracy is worse than the original model's accuracy (Table 1). However, among defended models the accuracy is different. For example, prediction similarity and model's behaviour have the best accuracy, which is the same the original model's accuracy (Table 1).

		Original Data	Impact on Prediction
Original Model		85.1%	
Defended Model	**Adversarial Training**	84.3%	< 1%
	Encoder	82.4%	<5%
	Middle Autoencoder	82.1%	<5%
	Initial Autoencoder	70.0%	<20%
	Prediction Similarity	85.1%	0%
	Model's behaviour	85.1%	0%

Table 1: Accuracy and impact of using defended models for classification

On the other hand, this study detailed the results obtained with each defence though two types of adversarial examples: the initial model's adversarial examples (known adversarials) and the defended model's adversarial examples (new adversarials).

Know adversarial examples: All three defences were tried on this case. Each defence was tested to calculate how many of the initial adversarial examples are no longer misclassified. As said formerly, adversarial training is the best option of the four to avoid initial adversarial examples (Table 2). Dimensionality reduction defends against this type of attack, while the prediction similarity does not, since it does not modify the structure or input data of the original model. It merely detected when an adversarial attack attempt is happening.

		Known Adversarial Examples	New Adversarial Examples
Defended Model Accuracy	**Adversarial Training**	92.0%	It does not detect new attempts of adversarial attacks
	Encoder	64.3%	They do not detect new adversarial attacks. However, they make several new adversaries distinguishable to the human eye.
	Middle Autoencoder	60.4%	
	Initial Autoencoder	70.5%	
	Prediction Similarity	0.0% (Useless against this type of adversarials)	The processes of adversarial generation are detected 99.5% of the time.
	Model's behaviour	94%	Similarly, adversarials are detected in the 94% of the cases.

Table 2: Accuracy of detection of the already known and new adversarial examples

New adversarial examples: Once the defences had been tested with initial adversarial examples, new ones were generated to attack the defended model. In this case, the adversarial training was not at all robust, as it was easy to get new adversarial examples after the retraining. However, dimensionality reduction was more robust in this case, as is visible in Figure 17, since the new adversarials became distinguishable for the human eye. Finally, prediction similarity was the one that detected the most adversarials (Table 2). The difficulty of this defence was in selecting adequate parameters and thresholds, and these can change depending on the dataset and the chosen metric. In our case, it was implemented with the parameters mentioned previously and the SSIM metric.

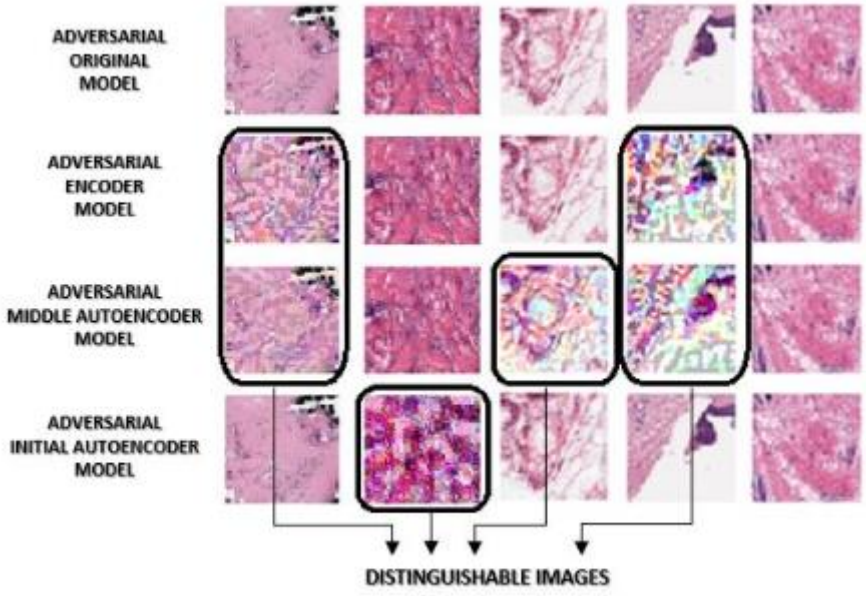


Figure 17: Image results of the different dimensionality reduction defences.

2.4 Graph Anomaly Detection

Graphs are very prevalent in a large variety of applications. From computer and social networks to movie rating services, there are a plethora of instances of graphs having influences on the daily life of the general public.

With such a widespread utilization, there are multiple problems in such environments that have emerged over the last couple of years. A key example is detecting graph anomalies. Graph anomalies can take the form of attackers requesting unusual traffic inside a computer network, bots trying to gain influence over social media, or users trying to deliberately down-vote movies on film-rating sites in order to make a political statement. Usually, when dealing with malicious intent, identifying graph anomalies is a key issue in protecting platforms and networks.

So far, numerous works have been published on this topic, proposing various methods for detecting such anomalies. However, one key issue that has not been sufficiently considered is the ability of adversaries to adapt to such methods. This robustness is a major factor in the viability of such solutions. The underlying assumption of stationary data with respect to the graph structure is not valid when dealing with adversarial attackers able to influence the base structure in a targeted way. For example, bots in social networks can modify their behaviour and connections in order to avoid being flagged.

The goal of this work is to test the robustness of methods for graph anomaly detection in the face of adversarial attacks with the goal of evading detection.

Additionally, a modular and extensible framework for attacking graph anomaly detection is created to allow for easier comparison and implementation of different attacks and different detectors. We were able to successfully evade detection for several nodes initially flagged as anomalies and significantly lower the anomaly score for others, making detection harder.

The results of our experiment show the vulnerability of the Embedding model to adversarial attacks. We tried three main different attacks (Random, Heuristic and Gradient) and all the attacks were able to successfully decrease the anomaly score in only a few steps to evade detection.

We decided to consider an evasive attack successful if and only if the score percentile decreased beyond a threshold 90%, to ensure a high enough number of other nodes to blend in with when

looking for anomalies. Under this criteria, 26 out of the 33 attacks were able to achieve evasion. Beyond that, 15 of the 33 attacks were able to achieve a final score percentage of < 60%.

2.4.1 *The procedure*

Graph anomaly detection refers to the use of anomaly detection on graph-structured data [27]. In such a case, anomalies are nodes with structural inconsistencies when compared to the entire graph dataset. For instance, anomalies could be bots trying to infiltrate social networks, or researchers using false citations to gain credibility. Generally, graph anomalies are defined in relation to clustering. Normal nodes are similar to other nodes in the network and can clearly be assigned to one or a small number of clusters. However, anomalies will be structurally different from benign nodes. This will lead to them fitting poorly into specific clusters and usually being assigned multiple ones.

For this work, we relied upon two different anomaly detection methods.

AutoPart [28] is a parameter-free algorithm that works by trying to reorder the adjacency matrix into clusters based on a minimum information principle, using an absolute clustering.

Embedding [29] uses an embedding approach to compute the best clustering based on minimizing the clustering loss and then computes the anomaly score for each node based on the variance in the embedding of its neighbourhood. This method uses a partial clustering.

2.4.1.1 AutoPart

AutoPart was proposed in 2004 by Deepayan Chakrabarti. The algorithm provides a method for parameter-free graph partitioning. This is significant since most clustering methods (METIS, k-means clustering, SVD/PCA) require the prior specification of parameters, whereas AutoPart does not rely on such input.

The method is based on the idea of using compression to guide its optimization. Specifically, it uses the MDL (Minimum Description Language) principle. This principle tries to describe the underlying structure in a compressed way. Since the compression is based on clustering nodes with similar information, it can be used to optimize the clustering.

2.4.1.2 Embedding

Published in 2016 by Hu, Aggarwal, Ma, and Huai, “An Embedding Approach to Anomaly Detection” proposed the use of network embeddings for anomaly detection. On top of providing a method for optimizing the embeddings, they introduced a novel dimension reduction technique for reducing both space and time complexity.

The computed embeddings are then used to identify the nodes that connect to a variety of diverse communities, bringing together different parts of the network. This is done by considering the variation in the neighbourhood for each node.

2.4.1.3 Adversarial Learning on Graph-Structured Data

One high-value area of attack for adversaries are classifiers operating on graph structures. These frequently occur in the real world, due to the diversity of applications where the graph information plays an important role in risk management [27]. Such areas of application include credit card fraud [30], malware/spyware detection [31], and online review deception [32].

For non-bipartite[2] graph-structure datasets, the evasion attack attempts to bypass the detection of specific anomalous nodes by introducing new outgoing edges from these nodes. Here, it is important to note that inserting edges changes the underlying graph structure, which can inform the responsible entity of possible evasion attacks. Therefore, it is important to decrease the anomalousness for the given node as much as possible, while changing the underlying structure as little as possible.

It is important to note that effectively finding an evasive variant is harder for graph structures when compared to image and text data. This is due to their discrete (while images are continuous) and combinatorial (where text data is limited by language structure) nature.

2.4.2 *Methodology*

In order to test the robustness of graph anomaly detection methods, we attempted multiple attacks on these models with varying levels of insight into the model. Running different attacks provides a broader insight into the overall robustness of the models, specific vulnerabilities, as well as the effectiveness and efficiency of different types of attacks.

2.4.2.1 Random Attack

This attack is the most generic attack of the three. It is often used as a baseline attack to evaluate the general robustness of a model and as a benchmark for other attacks.

The random attack works by randomly selecting edges to insert. To evade detection for a given node i, it will sample a node from $\{j \in V \mid (i,j) \notin E \wedge (i,j) \notin E \wedge i \neq j\}$. Then, the edge (i,j) will be added to G. After running detection again, the attack will observe whether the anomaly score for i decreased. If the score increased, it will remove (i,j) from G. If the score has dropped below the threshold, it will terminate. Otherwise, the attack will repeat the process until it reaches its iteration threshold.

This attack is a black-box attack (BBA) and relies solely on the score feedback. Thus, it requires minimal information about the model used.

2.4.2.2 Heuristic Attack

Heuristic attacks are based on the idea of using an approximate measure of how to proceed based on non-complete information extracted from the model. Therefore, heuristic attacks are grey-box attacks.

While heuristic attack refers to a general group of techniques, we will introduce the specific heuristic attack used for our experiment. This attack has been designed to work with the Embedding model and uses information specific to that detection model. However, the only information it relies on is the embeddings themselves, which occur as clusterings in the majority of other graph anomaly detection algorithms. Therefore, it is relatively easy to extend this method to such models.

Our heuristic attack is based on a fairly simple idea. The model detects anomalies by their relatively high correspondence to multiple different groups. This is measured by the distribution of $y_i^k \in \overline{NB(i)}$. Specifically, it is based on $\Sigma_{k=1}^{d} = \frac{y_i^k}{y_i^*}$, where $y_i^* = \max\{y_i^1, \dots, y_i^d\}$. Therefore, $AScore(i)$ decreases when y_i^* increases. Our strategy derived from this is to try and connect exclusively to

[2] Bipartite graph is a graph the vertices of which can be divided into two disjoint and independent sets U and V so that every edge connects a vertex in U to one in V.

nodes from one community k, thereby increasing y_i^k to become y_i^* and decrease the overall $AScore$.

The problem of choosing which community k to choose and which nodes in k to connect to, comes down to one key aspect: modifying the perceived graph structure as little as possible. In order to achieve this, our heuristic attack chooses k to be the community with the largest number of nodes. The large size of k should mask the introduction of i a lot better than smaller communities, which would be strongly affected by extra edges connecting from i. This is because of the ratio $r(k) = u / |E_k|$ of newly introduced edges to current edges of the community. Here, u is the number of new edges from i to community k, while $E_{k \subset} E$ is the set of edges in k. A smaller ratio $r(k)$ indicates a lower likelihood of affecting the overall community structure by connecting i to k. When choosing which nodes in k to connect to, the objective is similar; to modify the inherent community structure as little as possible, while still hiding i within the community. On the level of individual nodes, our method relies on the embeddings $\bar{x}_j, j \epsilon V$ and the number of intra-community edges $C(j,k) \coloneqq |\{(j,v) \in E \mid v \in v_k\}|$. Our hypothesis is that nodes j with larger values for x_j^k and $c(j,k)$ are less susceptible to being reassigned to different communities when a foreign node i is connected. This is because community correspondence is determined based on x_j^k, where larger values have more room to be decreased without affecting the overall classification. Also, $AScore(i)$ is determined based on $\overline{NB(i)}$, where a higher $c(j,k)$ indicates greater stability. Therefore, our algorithm works by looking at all nodes $\{j \in V | x_j^* = x_j^k \wedge (i,j) \notin E \wedge (j,i) \notin E \wedge i \neq j\}$. These prospective nodes are then ordered discerningly based on x_j^k and $c(j,k)$. Then, one by one, (i,j) is inserted into E and the model is re-run. If $AScore(i)$ increased, (i,j) is removed from E. If $AScore(i)$ drops below the specified threshold, the algorithm terminates. Otherwise, the next iteration is run with the next node in the previously specified order.

2.4.2.3 Gradient Heuristic Attack

The final attack we implemented was the gradient heuristic attack. This is technically a heuristic attack, though; it requires more information than the standard heuristic attack introduced above. Therefore, this attack can be considered a white-box attack.

It is important not to confuse the gradient heuristic attack with a gradient attack, which works by exploiting the gradient of the anomaly score w.r.t. the malicious sample. Implementing gradient attacks for graph-based models is considerably harder, due to their discrete structure as well as the non-linear steps taken by the detection models.

The key difference of our gradient heuristic attack from our standard heuristic attack is the method used to determine which nodes to connect to within the selected (in our case, largest) community. Namely, this attack does not rely directly on the embeddings determined by the detection model. Instead of using x_j^k to determine the likelihood of j being affected by connecting to it, this attack utilizes ∇x_j^k, which represents the gradient of $O\,w.r.t$ the k-*th* value of embedding $\bar{X}_j$. This is because this loss indicates the stability of the value to which x_j^k has been, which is also a measure of the susceptibility of x_j^k to being decreased when connecting a new node i. Here the stability increased the smaller $|\nabla x_j^k|$ is. Therefore, the nodes in k are sorted based on $|\nabla x_j^k|$ in ascending order.

2.4.3 *System Description*

In order to implement the attacks and models, we introduce the following system. This is aimed to be a framework for attacking graph anomaly detection.

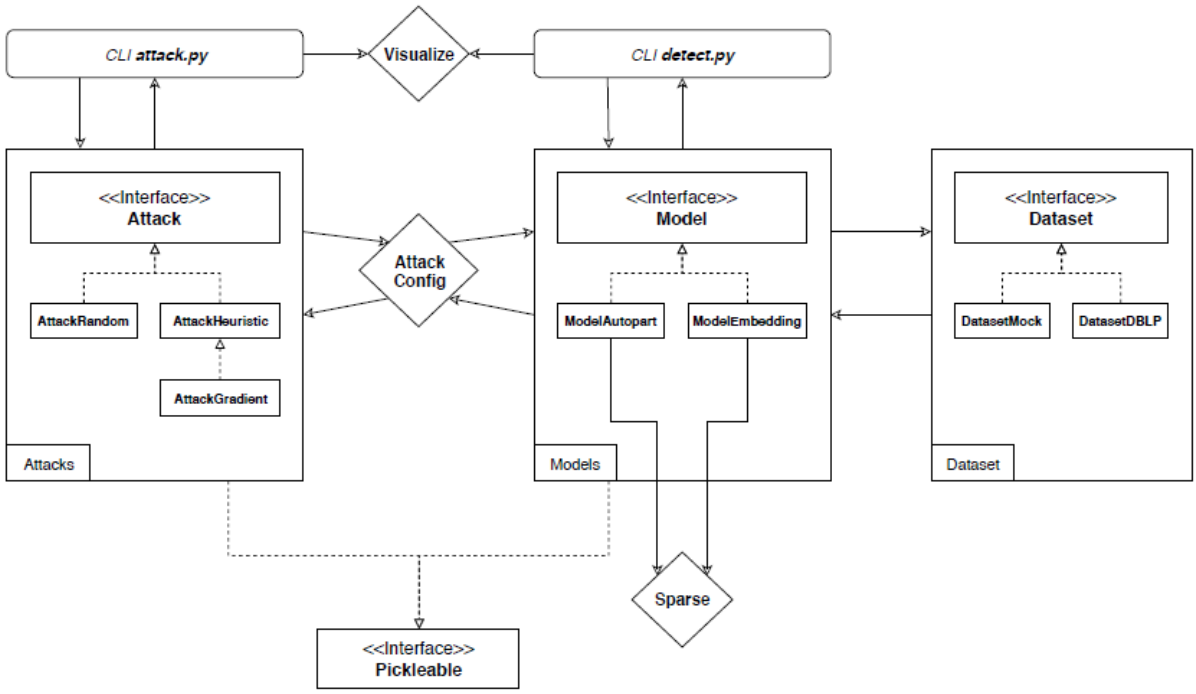


Figure 18: System Description

It provides a way to both run detection and attacks on given models, attacks, and datasets. The key idea here is extensibility. Each model instance is provided with an instance of a dataset (which contains a graph) to run detection on. Then, the detection is run iteratively, and optionally each step is saved as a checkpoint. The final results are then evaluated in order to identify anomalous nodes within the graph.

For attacks, each attack instance is provided with a model instance and the corresponding dataset. The model is then run for a specified number of iterations, according to the process specified above. Afterward, the attack modifies the underlying graph by inserting an edge with the goal of minimizing the likelihood of detection. At this point, it might also remove the edge inserted in the previous step, should it not have decreased the anomaly score. The next edge is chosen based on the information derived from the model iterations such as loss, gradient, et cetera.

Then, the model iteration is run again. This process is repeated until the likelihood of detection is sufficiently low. Optionally, each step is saved as a checkpoint. Finally, the difference in the anomaly score distributions is computed and displayed, as well as the change in the underlying graph structure.

Checkpoint saving for models and attacks is implemented through the use of the $Pickleable$ interface. This interface provides the functionality for storing and loading dictionaries using the Python serialization framework $pickle$.

Both detection models rely on the external $Sparse$ component for implementing their desired behaviour. This component provides two classes, $SparseColIndexer$ and $SparseCol$. The first maintains a $numpy$ array of the latter. $SparseColIndexer$ can be used as a fast sparse matrix implementation for slicing arrays into direct references, instead of newly created objects. The detailed implementation will be discussed later on.

2.4.3.1 Implementation

Our system provides an implementation of both introduced models using different datasets ($mock0\ and\ dblp$) as well as a mechanism for attacking those models. This mechanism provides access to different scopes of information about the model, based on the chosen attack type (white-box, grey-box, black-box). All previously introduced attacks have been implemented. The entire

software model is modular and extensible, providing room for additional attacks, models, and datasets. Both models and attacks provide automatic checkpoint support in order to avoid data loss when the scripts are interrupted and supports longer experiments. The framework is implemented using Python3.

In order to enable quick and easy user interactions, a $command\ line\ interface\ (CLI)$ is provided for both detection and attack. Beyond that, a visualization class provides a clear and informative overview of the distribution of anomaly scores in the dataset based on the model result. This can also be used to display the change in anomaly score for a given node after running an attack. Furthermore, the initial setup is simplified through a setup script, which installs all the necessary libraries and datasets and sets all configurations. Alternatively, a docker script is provided which will install all the necessary requirements inside of a docker container and launch it. This enables cross-platform support and avoids overwriting any existing configurations and setups.

2.4.4 *Evaluation*

2.4.4.1 Datasets

All the experiments were done using the $dblp$ dataset. Specifically, the version of $dblp$ mined by the *Stanford Network Analysis Project (SNAP)* [33] was used for all the experiments discussed below. The $dblp$ dataset is a citation dataset of major computer science publications. It consists of bibliographic information about all the included publications and authors. As of January 2019, the $dblp$ dataset contains over 4.4 million publications and over 2.2 million authors.

The $SNAP$ version of $dblp$ has mined the bibliographic information and compiled an undirected citation graph based on author relationships. Here, an edge between author A and B exists if A has cited B or B has cited A. The *SNAP* graph consists of 317,080 nodes and 1,049,866 edges.

Still, the framework supplies multiple datasets, which can be specified when running attacks or detection. The provided datasets are mock0 and $dblp$.

2.4.4.2 Detection Models

In order to validate the model implementations and gain insights into which models to attack, both models were run on the $dblp$ dataset until convergence and the results were analyzed.

2.4.4.3 Results

The AutoPart model converged after 58 steps. Its clustering algorithm identified 58 clusters with an average size of 4945 nodes. The distribution of its detected anomaly scores is showcased in Figure 20. Using 0.9 as an anomaly threshold, the AutoPart model detected 72615 anomalous nodes (17.05% of nodes).

The Embedding model converged after 138 steps. Its clustering algorithm identified 3170 clusters with an average size of 76 nodes. The distribution of its anomaly scores is showcased in Figure 19. Using 16.0 as an anomaly threshold, the Embedding model detected 135 anomalous nodes (0.04% of nodes).

The distributions of anomaly scores for the AutoPart and Embedding models are shown in Figure 19 and Figure 20. These figures plot the anomaly scores for each node, using the ordered indices of the dataset (so that index 140115 corresponds to the node with the 140115th highest anomaly score). All nodes above the anomaly threshold have been highlighted in red, while all the other nodes are shown in blue.

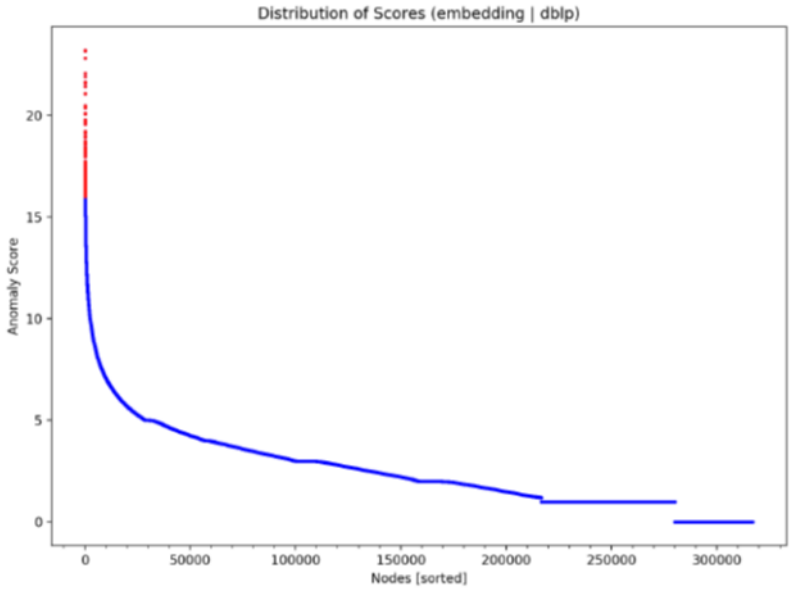


Figure 19: Anomaly Score Distribution for Embedding

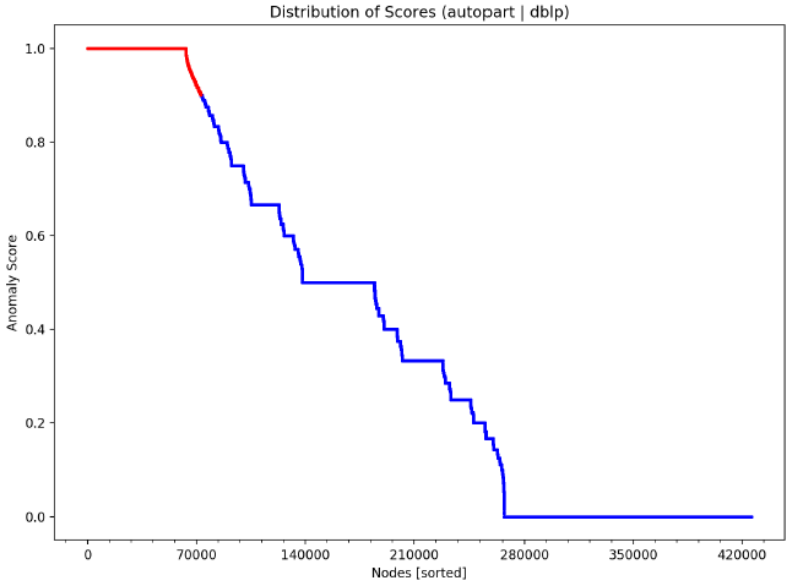


Figure 20: Anomaly Score Distribution for AutoPart

When interpreting these distributions, two distinct patterns emerge. The score distribution for the Embedding model in Figure 20 resembles that of a Pareto distribution, with a very small number of nodes with very high anomaly scores and an increasingly larger number of nodes with smaller and smaller scores. This closely resembles the expectation for anomaly detection, since their very definition makes anomalies rare occurrences. At the same time, there are different degrees of anomalies. The less severe the degree, the more anomalies are likely to be classified at that level.

The distribution of anomaly scores decreases fairly rapidly and smoothly, with the exception of two jumps to the levels of 1.0 and 0.0. These can be explained by two phenomena. The first jump occurs at from slightly above 1 to 1.0. When looking at the definition of $AScore$, this can be recognized as the cutoff between nodes the weighted neighborhood of which belongs to exactly one embedding community (indicating $AScore$: 1.0) and all the other nodes. The jump from 1.0 to 0.0 is the result of nodes without any connections in the network. These are largely discarded for the purpose of anomaly detection. Overall, the embedding model produced intuitive and probable results when tasked with detecting anomalies.

On the other hand, the distribution of anomaly scores for the AutoPart model (see Figure 20) diverges significantly from the expectation. First, 72615 anomalous nodes are detected when using

a threshold of 0.9. This represents about 17.05% of the entire dataset. Even when increasing the threshold to 1.0, 63204 anomalous nodes (14.84% of the dataset) are detected. In the context of the AutoPart model, this means 14.84% of nodes have 100% inter-cluster edges. This implies the node has 0 connections to any other node in its cluster, which would indicate an incorrect cluster assignment for said node. Beyond that, this high degree makes it virtually impossible in practice to detect any anomalies in the dataset. This is because such a high percentage of nodes have maximum or close to maximum anomaly scores that a comparative approach is infeasible in practice. In the example of $dblp$, it would be quite unfeasible to manually inspect 72615 researchers in order to confirm anomalous patterns in the citations.

Furthermore, the distribution is approximately symmetric to the score of 0.5, with about half of nodes classified above and half classified below. Such a high percentage of high anomaly scores seems unlikely in reality, given that the uniqueness property of anomalies is undermined if they occurred very frequently.

One possible explanation for these results is that an imprecise clustering was achieved, which undermined anomaly detection. The evidence for this would be the comparatively small number of clusters and their large size. For perspective, AutoPart found 58 clusters with an average size of 4945 nodes, compared to the 3170 clusters of average size 76 found by the Embedding model.

2.4.4.4 Evasion Attack

The robustness of the Embedding model against evasion attacks was evaluated using the following setup. Of the 135 detected anomalies, 11 nodes were uniformly sampled to represent different nodes with varying distances to the anomaly threshold. For each of the sampled nodes, all the three attacks were run. Each attack was run for a maximum of 18 steps until being stopped. If a score of below 2.5 was reached, it was stopped immediately. The step size was set to 100, meaning the Embedding model would run for 100 iterations between each attack step. This number was chosen as a trade-off between attack time and model $convergence$, since the rate of descent of the model loss decreased strongly around iteration 100. After each attack step, the recomputed anomaly score of the evasive node as well as the inserted edge was printed and saved as a checkpoint.

2.4.4.5 Results

As Table 3 shows, all attacks were able to successfully decrease the anomaly score in only a few steps to evade detection for the anomaly threshold of 16.0 they had been detected with. In fact, the highest post-attack anomaly score of 6.21 was less than half the threshold. However, due to the distribution of the underlying anomaly scores, an anomaly score of 6.21 would still be within the 97th percentile of anomalousness, making it feasible to still be detected when lowering the anomaly threshold.

Node	Initial	Random [Edges-C]	Heuristic [Edges-C]	Gradient [Edges-C]
69460	25.14	2.48 (59%) [2-0.08]	3.01 (46%) [3-0.09]	5.03 (91%) [5-0.08]
42656	20.15	3.68 (79%) [2-0.12]	3.04 (71%) [2-0.08]	2.73 (66%) [6-0.15]
27513	18.91	2.31 (49%) [3-0.12]	2.52 (61%) [7-0.11]	1.48 (38%) [7-0.12]
106237	18.19	3.56 (80%) [3-0.09]	3.99 (82%) [4-0.09]	5.97 (95%) [6-0.13]
92904	17.75	4.91 (92%) [3-0.09]	1.98 (50%) [4-0.09]	5.10 (93%) [2-0.08]
163051	17.00	1.60 (39%) [2-0.11]	2.69 (54%) [6-0.18]	2.81 (64%) [5-0.08]
26641	16.42	1.94 (51%) [2-0.08]	6.21 (97%) [2-0.09]	6.21 (97%) [2-0.05]
29960	16.78	1.41 (42%) [6-0.18]	1.47 (43%) [2-0.09]	2.95 (69%) [1-0.07]
170972	16.65	4.82 (91%) [4-0.19]	1.97 (49%) [5-0.11]	4.22 (88%) [5-0.09]
107629	16.22	2.20 (58%) [1-0.06]	2.03 (52%) [2-0.09]	2.03 (52%) [2-0.08]
149041	16.00	3.64 (80%) [4-0.08]	1.79 (48%) [3-0.12]	3.83 (84%) [3-0.12]

Table 3: Evasion attack results on different nodes against random, heuristic and gradient attacks

While this would be an improvement over the 99th percentile all anomalies started out in, it could still not be enough to achieve the stated objective. Therefore, we decided to consider an evasive attack successful if and only if the score percentile decreased beyond a threshold 90%, to ensure a high enough number of other nodes to blend in with when looking for anomalies. Under this criteria, 26 out of the 33 attacks were able to achieve evasion. Beyond that, 15 of the 33 attacks were able to achieve a final score percentage of < 60%. About half of all the attacks were able to achieve evasion by inserting only three edges or fewer.

When analyzing the factors leading to easing the evasion, distance to the anomaly threshold did not seem to play a significant role. There seems to be some correlation between the evasive score and the methods used. For example, the random and heuristic attacks managed to consistently achieve a low evasive score regardless of the node, while the gradient attack was more volatile and had higher variation between node scores. One somewhat indicative factor was the node itself, with nodes generally achieving somewhat correlated scores across the different attacks (such as generally lower, generally higher, et cetera).

In order to fully evaluate the impact of an evasive attack, it is important to consider the amount of change introduced to the model due to the modified data. To quantify this, a new measure C was introduced specifically to the Embedding model. It is defined as the average Euclidean distance between new and old embeddings:

$$C = \frac{1}{n}\sum_{i=0}^{n}\left|X_i^{(old)} - X_i^{(new)}\right|$$

It is important to note that due to their norm, the maximum Euclidean distance between two embeddings is 1. When evaluating the average change of embedding for each attack, the C-value ranged from 0.05 to 0.19 with the majority of attacks inducing a change of roughly 0.10. While this level of change is not drastic, it is not insignificant. A small average change for the embeddings could indicate substantial changes for specific parts of the network. Alternatively, it could be the effect of a not entirely stable embeddings mapping which induces small updates through the randomness of restarting the embedding with different edges being sampled at each step. Overall, these results show small changes in the embeddings, which is unlikely to cause the detection of the evasive attacks. However, it is not impossible.

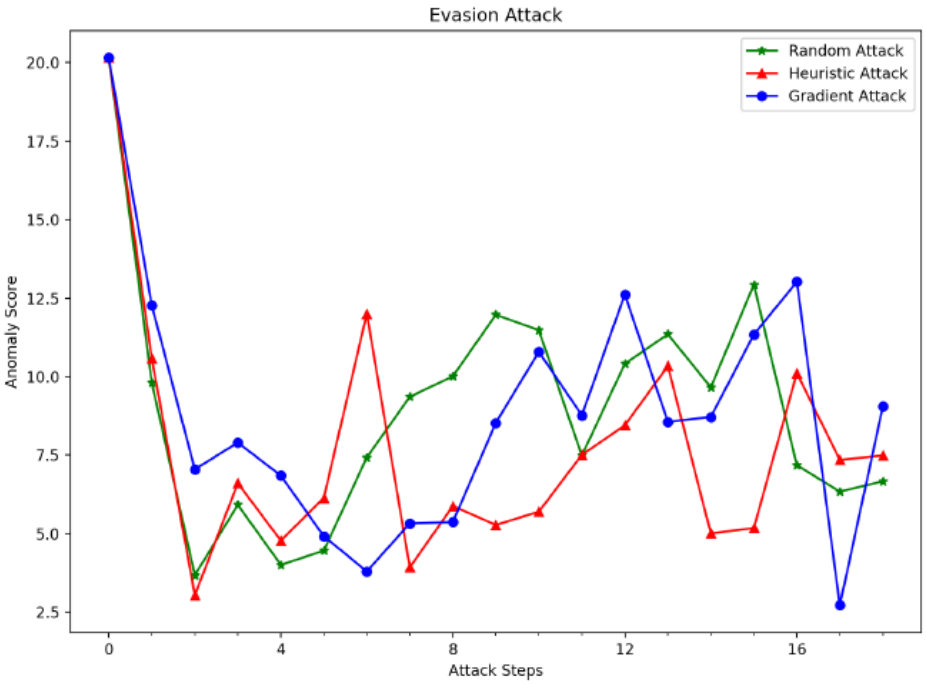


Figure 21: Step-by-Step AScore for attacks on Node 42656

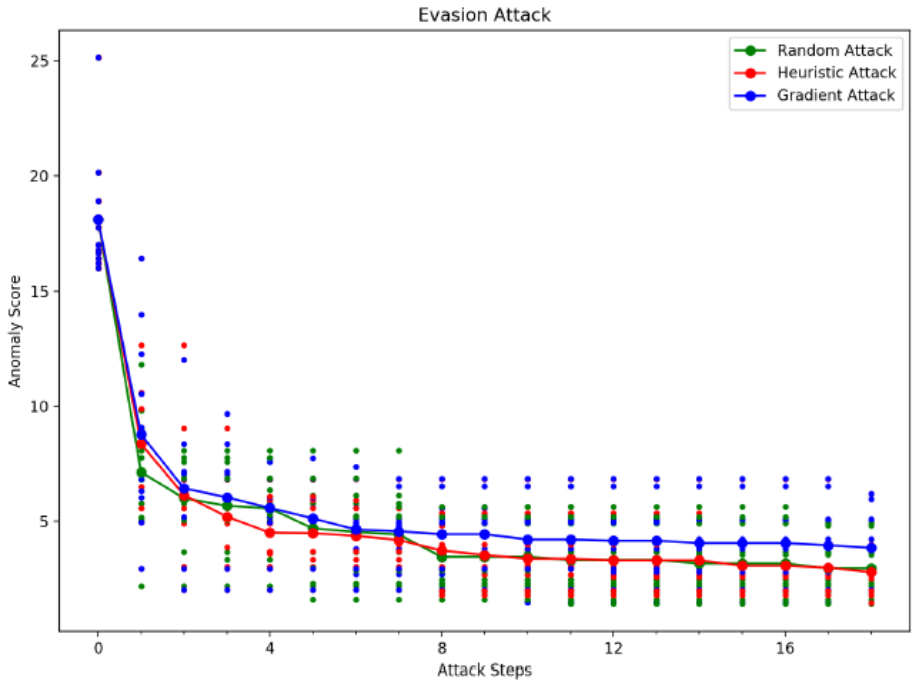


Figure 22: Step-wise Lowest AScore (averaged by attack type)

Figure 22 compares the performance of the different attack types. At each step, the smallest anomaly score reached so far is plotted for each attack on each node. These values are then averaged by attack type and summarized in the three graphs. One interesting factor to observe is the effectiveness of each step. By far, the largest reduction in scores is seen after the first step, then the second, third and fourth keep decreasing the scores significantly. After that, the scores slowly converge to their final values over the next 12 steps. When comparing performance, it seems like the random and heuristic attacks were able to achieve slightly better evasion on average than the gradient attack. This difference manifests mostly in the two largest nodes for which the gradient attack was not able to achieve evasive scores as low as the other attack types. One possible reason for the relatively poor performance of the gradient attack could be an unstable underlying gradient, since this method relies directly on it. Meanwhile, the heuristic attack only loosely relies on the embeddings as well as balancing it with a graph-based factor (the number of connections within the cluster). Overall, the surprisingly good performance of the random attack would support the idea of a relatively volatile underlying embedding, which can be tricked easily through a couple of random changes.

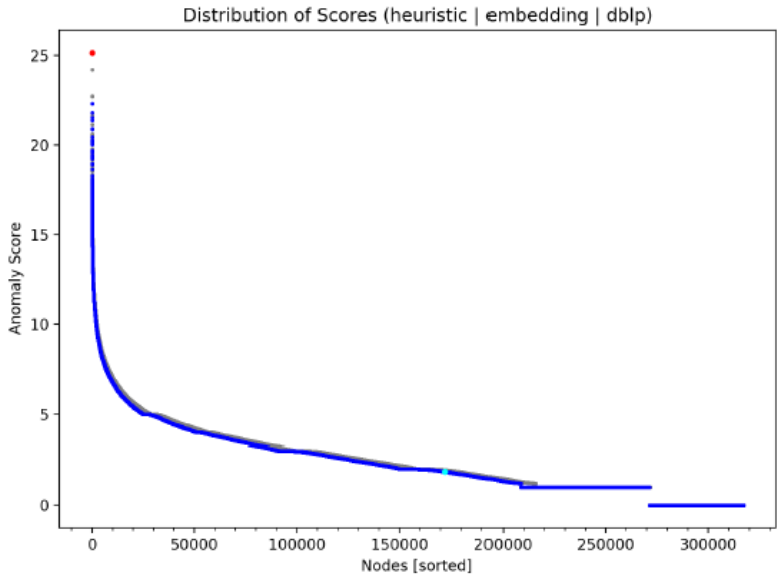


Figure 23: Distributions for Heuristic Attack on Node 69460

While Figure 22 showcases the scores as continuously decreasing, that is only true for the minimum score achieved so far. The actual step-by-step scores for the different attacks for select nodes are highlighted in Figure 21. As can be seen here, while the scores initially decrease quite quickly, the score periodically increases. This is due to the process of the attacks, where connections to different nodes are tested and kept only if the anomaly score decreases. Therefore, each drop in the graphs in Figure 21 corresponds to the addition of a new edge, while each rise represents the unsuccessful testing of one.

The change in the overall embedding has been quantified using C; however, it is important to investigate how the underlying distribution changed for successful evasion. Figure 23 shows this for one example attack. It plots the initial score distribution in grey, with the initial anomaly score of the selected node highlighted in red. On top of that, the final score distribution is plotted in blue with the final anomaly score for the selected node being emphasized in cyan. This figure shows a very small change in the overall distribution of the anomaly scores, which increases the odds of successful evasion and supports the idea of the attack having altered the underlying structure only minimally.

2.4.4.6 Transferability

In order to evaluate the degree of specificity of the attack to the model used, a transferability attack is run with the goal of evading another method without having any knowledge of it.

For each node, the final edges determined by the evasion attacks were inserted into the graph. Then, the AutoPart model was run on the modified graph, with the score for the evasive node compared to its initial score.

2.4.4.6.1 Results

The initial anomaly score, as well as the modified score, after transferring the results of the random, heuristic and gradient attacks, are listed in Table 4.

There does not seem to be a clear decrease in anomaly scores when using the transfer attacks. Overall, the scores seem to stay fairly similar to the initial score, regardless of the attacks. There are multiple possible explanations for this result. On the one hand, the accuracy of AutoPart for detecting anomalies has not been sufficiently established. However, the scores for the selected anomalies tend to be comparatively high, which would indicate an overlap in anomaly detection between the two models. Another option is the possibility that the underlying features vary between

the two models and that the attacks have adjusted closely to the specific details of the Embedding implementation, such as the detailed nature of the embeddings.

Node	Initial	Random	Heuristic	Gradient
69460	0.86	0.87	0.87	0.88
42656	0.90	1.00	0.87	0.96
27513	0.97	0.97	0.95	0.88
106237	0.39	1.00	0.97	1.00
92904	0.99	0.97	0.96	1.00
163051	0.50	1.00	0.79	0.77
107629	1.00	1.00	1.00	1.00
149041	0.69	0.73	0.69	0.79
170972	0.86	0.89	0.89	0.79

Table 4: Transferability Attack Results

Papernot, McDaniel, and Goodfellow [34] showed that transferability tends to be higher when transferring attacks from complex models to the less complex ones or to the models that are based on similar principles. This could be a possible explanation for the lack of transferability, given the difference in model structure between the AutoPart and Embedding models as well as their comparative complexity.

2.4.4.7 Discussion

The results of our experiment show the vulnerability of the Embedding model to adversarial attacks. Since most anomaly detection techniques are similar in nature, it is possible for other graph anomaly detection techniques to share this vulnerability. Given the wide-spread applications of graph anomaly detection in industry, the consequences of anomalies being able to evade detection are quite concerning. In order to prevent such vulnerabilities, robustness against evasion attacks should be given more consideration when designing models with the purpose of detecting anomalies.

2.4.4.8 Limitations

There are multiple factors to take into consideration when evaluating the completeness of the results obtained through the experiments. First, all the experiments were run using only the $dblp$ dataset. This reliance on only one dataset introduces the risk of the behaviour of models and attacks having been learned from the specific dataset features, which would prevent the results from translating to general applications. It is important to note that this limitation is hypothetical in nature. There is so far no evidence that such overfitting of attacks and models to the data has occurred. However, it is a noteworthy limitation to consider.

One similar factor is the relatively small number of models that were considered. Given the evasive attacks were only run on the Embedding model, it is possible the attack performance is conditional on taking advantage of model-specific properties that may not generalize to other models. As with the first consideration, there is no specific evidence suggesting this happened; however, it remains a realistic possibility to entertain.

2.5 Study of limitations of AI in the cybersecurity context

As always, while creating commercial or research-based products, several risks, threats to validity and human factors should be discussed and be taken into account. In this section, we address such aspects as the quality of data, the security of the system itself and the cognitive understanding of the system results/decisions.

2.5.1.1 Quality of the Data

The proposed machine learning based IDS needs to be trained on real or realistic and reliable data. Such data has to come from certified devices (probes, software-based parsers etc.) and needs to represent real and non-biased network traffic. Yet another problem lies in dataset characteristics: it should be well balanced in terms of the percentage of malicious data vs. normal not-attacked/ benign traffic. Another challenge that falls roughly in the quality of data category comes with the fact that even the best, most clever and well optimised algorithms are optimised at the time of deployment, and their performance deteriorates in time, even more so with the fast-paced advancement of network technologies. This unfortunate predicament could be countered with the life-long learning approach [35].

2.5.1.2 Security of the system and detecting adversarial attacks

One of the crucial (even though often forgotten) aspects of each security system is that such a system should be secure itself. Therefore, machine learning based intrusion detection systems have to be well protected against attacks (both physical and cyber). Another emerging aspect is that the internal machine learning algorithms have to be secure as well, which mostly means being resilient to adversarial attacks such as poisoning, model extraction etc., as it was demonstrated in [36].

2.5.1.3 Explainability and Cognitive Understanding

Many well-designed and developed software products (especially coming from R&D advances) were not successful due to human factors, lack of usability or badly designed GUI. All the results, indication of anomalies and/or cyber-attacks and the suggested decisions have to be well presented to end-users (e.g. a network security officer or IDS operator). The user cannot be overwhelmed by the information from the system, but on the other hand should be able to clearly see the context and the situation in order to increase the situational awareness. Yet another threat lies in the fact that quite often machine learning solutions are used as black-box single truth providers, while in fact, operators should understand the algorithms to make informed decisions. Therefore, the aspect of so called explainability of ML solutions should be addressed and taken into account to ensure wide adoption of this class of intrusion detection systems.

2.6 Summary of Chapter 2

In this chapter, a wide variety of defences against adversarial attacks was proposed across multiple domains. The solutions were tested and their results were presented, as well as their impact (or lack thereof) on the performance results of the defended classifier. The limitations of AI in cybersecurity were also considered.

Chapter 3 Explainability enhancing mechanisms

The aspects of explainability and interpretability are trending topics in the area of machine learning and artificial intelligence in general. The two terms – explainability and interpretability tend to be used interchangeably; however, despite the fact that they are related concepts, there are some minor differences in their meanings. Interpretability addresses the aspects related to observation of AI system outputs. Interpretability of AI system is higher if the changes of the systems outputs in result of changing algorithmic parameters are more predictable. In other words, system interpretability is related to the extent to which a human can predict the results of AI systems based on different inputs. On the contrary, explainability is related to the extent to which a human can understand and explain (literally) the internal mechanics of an AI/machine learning system. In its simplest form, the definition of explainability refers to an attempt to provide insights into the predictor's behaviour. Nowadays, the attempts to define these concepts are not enough to form a common and monolithic definition of explainability and interpretability, and to enable their formalization.

3.1 Why must we explain Artificial Intelligence?

The explainability and interpretability of different AI models can help to answer different questions that can be asked by the human operator/beholder. It is also required for safety cases when AI is being used in safety-critical systems.

Increase acceptance and trust of the designer: Before inserting the AI algorithm in the system, it shall be validated. The designer wishes to answer many questions: Why not apply another model? When does the model succeed and fail? Where and why can we trust the model? As far as learning algorithms are concerned, is it possible to provide guarantees on the outcome for unseen situations?

Increase acceptance and trust of user: In Man-Machine Interaction, once the AI algorithm has been integrated in the system, the user wishes to understand its outcomes. Why does the model decide this? It is important to note that not all algorithms need the same level of explanation. The acceptance and trust of the user is related to validation, verification, qualification and certification of a system. The more explainable a system is, the easier one expects to conduct these four steps. As a result, explainability is an important component of Trustable AI.

Bring knowledge on the domain. Having an interpretable model can bring knowledge on the domain. In medicine, data scientists aim at helping physicians to construct rules from field studies and experiments.

Interpretable model could be executed by humans. In medicine, doctors like very simple models as they can be executed by hand. This is especially true for scoring methods.

Defend the decision to stakeholders. In a decision support system, if the recommendation ought to be defended to some stakeholders, a convincing explanation becomes necessary.

Debugging mode (called inspection in ML). In knowledge representation and reasoning (KRR), when an inconsistency is raised in a knowledgebase, the explanation of the inconsistency can be turned in funding one or several possible repairs, such as constraints relaxation in Constraint Satisfaction and Search (CSS). In ML, inspection is widely used to understand for instance what is the meaning of each layer in a deep neural network, or which pixels in an image mostly explain a given classification. The explanation of models outputs makes it possible to understand their "reasoning" for an operator but also for people who create these models. Thanks to a better understanding of how a model works, the data scientist / data analyst / statistician is able to improve and debug it more easily. Indeed, performance criteria do not always guarantee that an algorithm will behave well. For instance, a model could give a right prediction but for wrong

reasons. This can happen for example in object detection tasks with images when a model classifies an object based on contextual elements rather than the object itself. This ability can also be beneficial for an operator who knows the job and wants to improve or debug a model iteratively. [37] present a case study in which non-experts (in machine learning) perform variable selection based on the explanations related to model outputs. More generally, this approach enforces connection between a human and computer. Furthermore there are some links with active learning field.

Abduction/Diagnostic. One would like to identify the root causes in a fault tree. In a diagnostic, one needs to identify the most probable hypothesis that can explain a given symptom. Diagnostic is also coined in UAI, where there is uncertainty on what is exactly the cause of what is observed.

Causality. Is it possible to make the right inference and find the right root cause? If we consider BN (Bayesian Network), there are many equivalent rewriting of a given BN yielding exactly the same outcome. What distinguishes these BNs is whether causality is correctly represented. Causality is about the correct inference. For instance, we can observe that we can often see together a fire and smoke. However, the fire is the cause of the smoke and not the other way around. Guessing the correct causality among concepts is key in many domains and is essential to generate convincing explanations.

Fairness and unbiasedness. As already discussed in the previous examples, explainability can help to analyse the potential biases of AI. In particular, the AI algorithm shall not be influenced by some racial/gender factors. More generally, is it possible to ensure that the AI function has not a biased perception of the world due to a bias in the dataset, the model or the objective function to optimize?

3.2 Explainability of AI – state of the art

Explainability and interpretability have been moving into the spotlight in the ML and AI research community over the last few years. As discussed in [38] and [39], these two terms – explainability and interpretability - tend to be used (also in literature) interchangeably, however despite the fact that they are related concepts, there are some minor differences in their meanings. Interpretability addresses aspects related to observation of AI system outputs. Interpretability of AI system is higher, if the changes of the systems outputs in result of changing algorithmic parameters are more predictable. In other words, system interpretability is related to the extent to which a human can predict the results of AI systems based on different inputs. In contrary, explainability is related to the extent to which a human can understand and explain (literally) the internal mechanics of an AI/machine learning system. In its simplest form, the definition of explainability refers to an attempt to provide insights into the predictor's behaviour [41].

According to [39], nowadays, attempts to define these concepts are not enough to form a common and monolithic definition of explainability and interpretability and to enable their formalization. It is also worth mentioning, that the "right to explanation" in the context of AI systems directly affecting individuals by their decisions, especially legally and financially is one of the subjects of the GDPR [40].

Different scientific and literary sources focus on surveying and categorization of methods and techniques addressing explainability and interpretability of the decisions resulting from AI systems use.

[42] discusses the most common practical approaches, techniques and methods used to improve ML interpretability and to enable more explainable AI. They include, among others, algorithmic generalization, i.e. shifting an attitude from case-specific models to more general ones. Another approach is paying attention to feature importance, described also in [41] as the most popular technique addressing ML explainability, also known as feature-level interpretations, feature attributions or saliency maps. Some of feature importance-based methods found in the literature are perturbation-based methods based on Shapley values adapted from the cooperative

game theory. In the explainability case, Shapley values are used to attain fair distribution of gains between players, where a cooperative game is defined between the features. In addition, some recent works [42][44] show that adversarially trained models can be characterized by increased robustness but also provide clearer feature importance scores, contributing to improved prediction explainability. Similar to the feature importance way, counterfactual explanations [43] is a technique applied in the financial and healthcare domains. Explanations using this method are based on providing point(s) and values that are close to the input values for which the decision of the classifier possibly changes (case-specific threshold values). Another method used for increasing explainability of AI-based predictions is LIME (Local Interpretable Model-Agnostic Explanations) based on approximation of the model by testing it, then applying changes to the model and analysis of the output. DeepLIFT (Deep Learning Important Features) model is used for the challenge-based analysis of deep learning/neural networks. As described in [42], DeepLift method is based on backpropagation, i.e. digging back into the feature selection inside the algorithm and “reading” neurons at subsequent layers of network.

In the literature, one can find different attempts of categorization of the methods aimed at increased explainability of AI. Integrated/Intrinsic and post-hoc explainability methods [39][43] is one of such categorization. Intrinsic explainability in its simplest form is applicable to the low complexity models (linear ones, decision trees, rule-based) where the explanation of a simple model is the model itself. On the other hand, more complex models are explainable in a post-hoc way, providing explanations after the decision and using techniques such as feature importance, layer-wise relevance propagation, or mentioned Shapley values. Other forms of post-hoc explanations include also textual and visual justification of the decision.

Similar categorization is given in [43] where in-model (integrated/intrinsic) and post-model (post-hoc) methods exist alongside additional pre-model interpretability methods. Pre-model methods are applicable before building (or selection) of the ML model and are strictly related to the input data interpretability. They mainly use the classic descriptive statistical methods, such as PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and clustering methods such as k-means. Another criterion described in [43] is the differentiation into model-specific and model-agnostic explanation methods. In the majority of cases, model-specific explanation methods are applicable to the intrinsically interpretable models (for example analysis and interpretation of weights in a linear model), while model-agnostic methods can be applied after the model and include all post-hoc methods relying on the analysis of pairs of feature input and output.

Alternative criterion based on explanation methods is described in [45]. In such differentiation, methods are categorized based on type of explanation that a given method provides, including: feature summary (providing statistic summary for each feature with their possible visualization), model internals (for intrinsic explainable or self-explainable models), data point (example-based models) and surrogate intrinsically interpretable model - that is trained to approximate the predictions of a black-box model.

According to [43] and [46], explanation models can be evaluated and compared using qualitative and quantitative metrics, as well as by comparison of the explanation method’s properties, including its expressive power, translucency (model-specific vs. model-agnostic), portability (range of applications) and computational complexity. On the other hand, individual explanations can be characterized by accuracy, fidelity, consistency (similarity of explanations provided by different models), stability, comprehensibility, certainty, to list most relevant ones. According to the literature we can also distinguish qualitative and quantitative indicators to assess the explanation models. Factors related to quality of explainability are: form of the explanation, number of the basic units of explanation that it contains, compositionality (organization and structure of the explanation), interactions between the basic explanation units (i.e. intuitiveness of relation between them), uncertainty and stochasticity. Quantitative indicators are presented in some works (e.g. [43][47][48]). The most common metrics used to quantify the interpretation of ML models are identity, separability and stability. These three factors provide the information on to what extent identical,

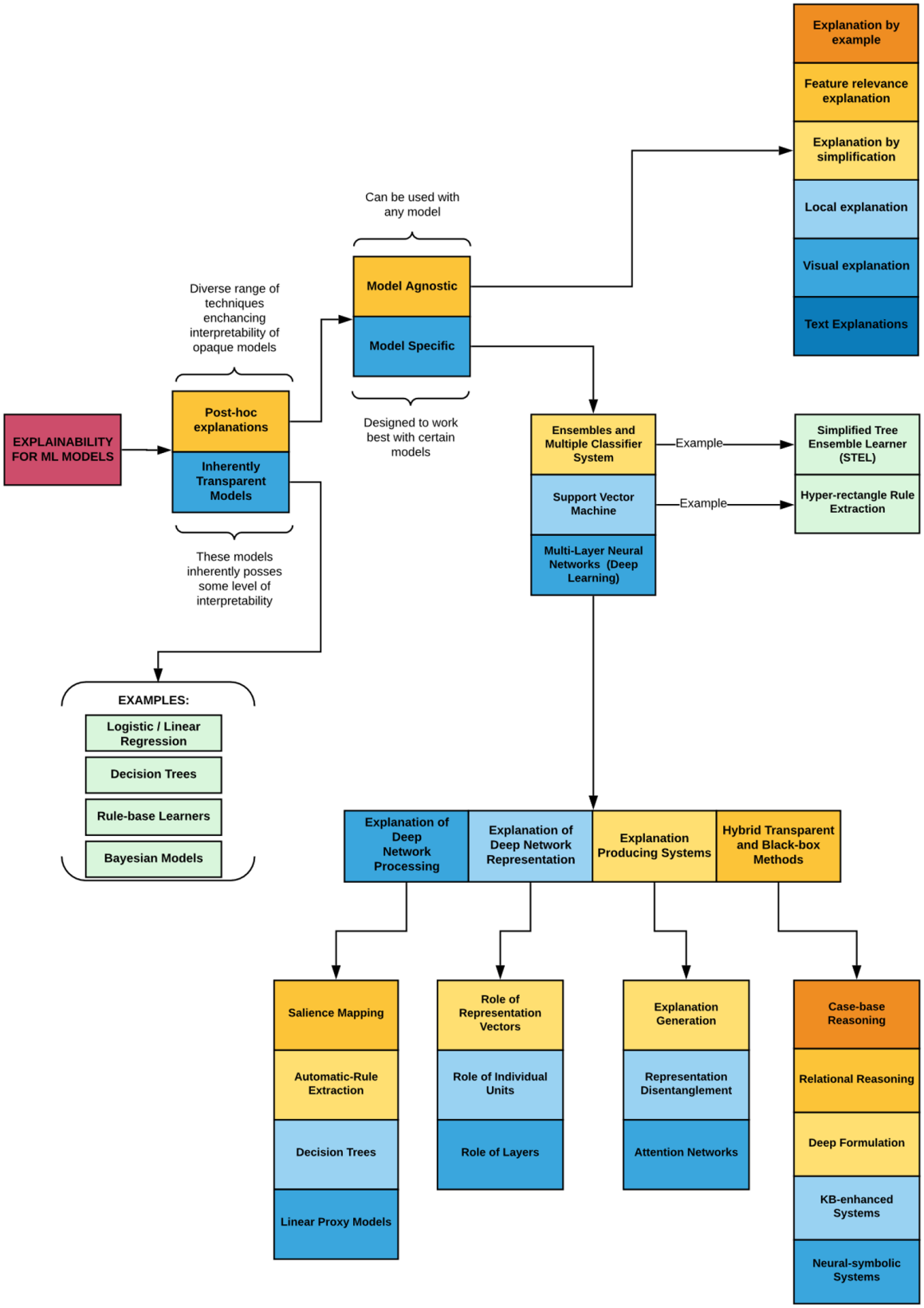


Figure 24: A comprehensive categorization of xAI approaches

non-identical and similar instances of predictions are explained in identical, non-identical and similar way, respectively. In addition, according to [47], the explanation should be characterized by high completeness (coverage of the explanation), correctness and compactness. However, these indicators are applicable only to simple models (rule-based, example-based).

3.2.1 *Existing explainability solutions*

Since AI models have become sophisticated enough to outclass many competing approaches in their respective fields (in certain applications even surpassing human capabilities), their popularity is on the rise. With initiatives such as autonomous vehicles, different recommendation systems (e.g. used by Netflix or Google's Sybil), personal assistants and many more, intelligent systems are being naturally ingrained in contemporary lifestyle.

This increasing ubiquity along with black-box nature of the best performing solutions has led to a few important realisations. Questions about model fairness and correctness were raised. How does one know that the reasons behind its decision are right or even acceptable? Should the society trust the system if it is not well understood? To provide valid answers for those question the notion of Explainable Artificial Intelligence (xAI) has emerged. Its main concern is not to develop the best models, but instead deliver tools and methods that allow human operators to understand the process and the reasoning for AI-based decisions.

Currently explainability is a very dynamic and quickly expanding discipline, drawing attention of the biggest corporations like Google and IBM. Although the need for xAI products is clear and there are many ongoing studies testing and honing explanation-oriented methods, there is currently a small recognition for existing, user friendly, enterprise solutions. Though the field of explainability has already received acknowledgement within the academic world, it just begins the process of becoming a part of the wider awareness. First, the landscape of xAI solutions on the market will be presented, to setup the stage as to the current situation in that sector.

3.2.2 *xAI Through Libraries and Frameworks.*

A step above single files and code fragments there are modules, libraries and frameworks. Those offer practitioners whole collections of methods in a single package. The iNNvestigate library is a good example. This library can be imported using Pythons package manager pip. It allows a developer to quickly use algorithms such as PatternNet, PatternAttribution and different variants of Layer-wise Relevance Propagation (LRP). Another representative of this category is Skater. It provides completely different methods from iNNvestigate, like PartialDepen-denceortextbf Local Interpretable Model-agnostic Explanations (LIME). The last example for this category is AI Explainability 360 Open SourceToolkit from IBM[1]. It presents itself as one of the best frameworks currently available for the practitioners. It offers a diverse array of algorithms from different categories like ProtoDash or Contrastive Explanation Method (CEM), even improving on some of them [49].

Additionally, in contrast to the libraries mentioned earlier, it also provides some metrics to evaluate the quality of provided explanations, though it is still quite limited. All of this is backed up by an extensive amount of materials, tutorials, and guidelines, which makes it easier to start working with xAI.

3.2.3 *xAI as Part of the System*

A popular alternative for frameworks is designing and implementing solutions integrated into a specific system. The main benefit of this approach is a full customisation allowing to cater for specific needs of stakeholders and their product.

Explainability is therefore a natural part of the whole and seamlessly (at least in theory) integrates with the remaining parts of the solution. On the other hand, the main disadvantage would be the need for additional resources necessary to develop the xAI module from scratch. Additionally, it is necessary to have someone with expert knowledge about the subject. Not all companies have

personnel of this kind at their disposal, rising the probability of failure significantly. There are some interesting examples of such systems presented to the wider audience.

In [50], the authors conceived a methodology allowing to develop explainable decision-support-system automating loan underwriting. It was designed to help money lending companies decide whether to accept or reject a loan application. To achieve that, they automated the whole process using belief-rule-base (BRB) expert system. This is basically extension of traditional IF-THEN rule based systems. It can for example use hierarchical knowledge structure. Factual rules engine can make up a first level and contains rigid, expert made rules that are not adjustable. Second level is then taken by heuristic rules engine, that is train-able and simulates behaviour of experienced loan underwriter. The whole decision process is traceable. As it was shown in case study conducted for Together Financial Services, system presents the clear impact of different rules on each decision in form of weights, combined probabilities, and textual explanations to support all of those. Additionally, it achieves quite good performance, with test accuracy around 95%. Unfortunately, the authors point it out that development of such system is very time consuming and needs deep expert knowledge for rule engines. Also, there is demanding data preparation process for the training step. Finally, it was unable to cover all loan application rejection scenarios.

Another interesting example from the financial technology market is Flowcast [51]. They offer machine learning products for money lending companies. Smart-credit is one of them and is supposed to help in decisions on financing small and medium-size enterprises (SMEs), that is, companies with small amount of traditional financial data used by banks in classic loan application process. This often leads to the rejection of such applicants, although some of them are potentially good clients. Authors claim that this market offers 540 billion dollars' worth financeable opportunities. Therefore, their system was designed to collect information from non-traditional sources like transaction data, to help a lender get a better picture of an SME company and more correctly asses risks. Their platform supports a selection of ML algorithms, one of them being a variant of the boosted trees algorithm. Because explainability is crucial in the finance sector and the mentioned algorithm is naturally opaque, they had to find a way to clearly explain the system assessments. They declare that they did. In their white paper, they describe the usage of SHapley Additive exPlanations (SHAP), along with Natural Language Processing (NLP), to generate plain text sentences explaining the output in layman's terms. This explanation is supposed to provide a description of why the system made such a decision, what must be done to change it and finally the confidence in it. They highlight, that risk professionals employing their platform can access up to then top reasons why each decision was made. The quality of those explanations is tested by focus groups comprised of risk management professionals and consumers.

The concluding example of a system with an integrated xAI module comes from the area of cybersecurity. To protect the network environments from an unwanted, malevolent activity, intrusion detection systems (IDS) are deployed. Lately, systems that employ some form of ML has become very popular. The ones with the best performance usually utilise some form of deep learning. However, because the artificial neural networks' (ANN) black-box nature, cybersecurity experts have difficulties in making decision based on their output. As in the first example, this opaqueness raises concerns and fosters lack of trust. This is a serious issue in the field of cybersecurity, where a wrong decision can have dramatic consequences. The experts need clear understanding to make right decisions. Authors of [52] present a way to help achieve that. On a benchmark dataset, they have trained two deep neural networks to act as IDS. Then, they attached an explainability module that uses the SHAP algorithm. It is important that it can work with any model, and present impact of each attributes locally i.e. for a specific sample, and globally i.e. for the whole model. Explanation is provided using simple charts that clearly show features and their contributions. Additionally, the paper introduces a new way to show the global relations among feature values and classes.

Approaches to xAI	
xAI Through Libraries and Frameworks	**xAI as Part of the System**
iNNvestigate	"An explainable ai decision-support-system to automate loan underwriting" [13]
Skater	Flowcast[14]
AI Explainability 360 Open SourceToolkit[1]	"An explainable machine learning framework for intrusion detection systems," [15]

Table 5: Approaches to the explainability of artificial intelligence (xAI)

3.3 Taxonomy of Explainability techniques

The explainability domain (XAI) is a wide area covering all AI topics. Very similar techniques are used in different fields. This is for example the case of the use of influence indices or counter-factual examples. A general taxonomy of XAI will be helpful to put in the same picture the techniques designed in very different domains. A taxonomy of XAI techniques is proposed in [53] for a black-box model for machine learning and data mining.

The XAI approach is operating on a model that has already been constructed. Often, this model is complex and not transparent, in which case it is called a black-box model. A classical XAI approach is to construct another model -- called the explanator -- from which the explanation will be generated. The fidelity of the explanator measures how close the predictor is to the original black-box model.

Applicable Models: Interpretable models are divided into two categories. One option is to use model-specific interpretability (or intrinsic interpretability) methods. The drawback here is that it ties you to this one algorithm and it will be hard to switch to something else. It refers to the ML models that are considered interpretable due to their simple structure (e.g. regression tree, linear regression, etc.).

Another alternative is to exploit model-agnostic methods (or post hoc interpretability). A model agnostic method is anything that is built on top of an interpretation of a ML model, like a graphic or some user interface, which becomes independent of the underlying ML model. Here the explanation is built after the ML model training, and can be applied whatever the ML model was.

The big advantage of the model-agnostic interpretability methods over the model-specific ones is their flexibility. The ML developer is free to use any ML model they like, when the interpretability methods can be applied to any model.

Input Data: This is the type of input data that is used for the model to explain. It can be of three types: Tabular (classical dataset representation through features which can be a number; a label, etc…), Image, or Text.

Scope: We speak of global interpretability if the aim is to explain the complete behaviour of the black-box model, that is for any possible instance. On the contrary, local interpretability only aims at constructing an explanator that is deemed to represent the black-box model only locally around a particular instance.

Local interpretability can be further sub-divided. The explanation of a given instance can be performed intrinsically, without the comparison with any other instance. This is Absolute Explanation. Another classical approach (called contrastive) is to compare the instance with other situations. This is motivated by the fact that people tend to explain their actions in comparison to another situation. The comparison can be made with a reference situation (One Instance vs Reference) that is always used, with what the user would have done in the same circumstance

(One Instance vs Human Expectation), and with other instances (Comparison between 2 or more instances) when the models allow to compare instances. Both global and local explanations answer different questions:

– How the models make their prediction? This global interpretation of the model, explaining its general functioning, can be obtained by learning a simple model on its outputs.

– How one part of the model influences the prediction? This global level of interpretation concerns the effect of a particular factor/feature/covariate on the outputs of the model

– Why does a model make its predictions for a group of records or for one instance? This local level tries to explain the model's prediction made for an observation or a group of observations

The first two concern global explanations and enhance models behaviours. The last ones are local and try to explain why a model produces some predictions for an (group of) instance.

XAI technique: At first approximation, the many XAI techniques can be organized as follows.

In order to avoid trying to explain a black-box model, a first approach – called Intrinsically Interpretable Model – aims at constructing the initial model, a model that is intrinsically interpretable rather than a black box. This can be obtained by restricting the class of model to directly interpretable models. This covers the use of Simple Models. In Machine Learning, one generally considers the following three models simple: a linear model taking the form of a weighted sum of the input feature variables, the rule-based systems and decision trees. Interpretability can also be obtained by enforcing specific properties to the model. The most well-known property is monotonicity between the inputs and the output of the model. It is used in machine learning, decision and other fields. Other more specific properties also exist. These properties are provided by the expert and are essential to ensure good behaviour of the model. If monotonicity of the system is essential for the end-user but is locally not satisfied by the model, the end-user will completely lose confidence in the model if they discover the situations of violation of such property. We restrict here to models that mathematically necessarily satisfy these properties. The last possibility to get intrinsically interpretable model consists in encouraging sparsity of a complex model by penalizing the complexity of the model in the objective function. The complexity can be the number of non-zero terms in the model.

The other big class of XAI techniques consists then in constructing an efficient model (regardless of any explainability concern) and then trying to explain this model. This is called reverse engineering. This can be split into two approaches: Construct Explanator and Inspection.

In the first one (Construct Explanator), an explanator (explanatory model) is constructed. It is deemed to be simple, so that the models used here are similar to what has been described in the category 'Intrinsically Interpretable Model'.

In the second one (Inspection), no specific explanator is constructed. The explanation is obtained by analysing the behaviour of the model. This can be done by sensitivity analysis, which analyses how the uncertainty in the output of the model can be explained by the uncertainty in the inputs. Feature Contribution summarizes the feature effect by some statistics. Some return a single number per feature, such as feature importance or Shapley Value, or a more complex result, such as pairwise feature interaction strengths. Data Point techniques include all the methods that return a set of data points to allow the model interpretation. These points can already exist in the dataset or can be simulated or generated.

Output of XAI: The output of the XAI technique can be a Graphical Visualization, such as the use of the visualization of feature statistics or Partial Dependence Plot. The XAI output can also take the form of a text that is automatically generated by NLP, or numerical values (such as feature statistics). It can also be directly a simple model (the explanator or intrinsically interpretable model).

3.4 Determination of the contributions of each attribute and counterfactual examples techniques

We focus here on two techniques which we think are interesting to exploit for the future of the project.

3.4.1 *Determination of the contributions of each attribute in the prediction for an instance*

Shapley's values, derived from the game theory, offer a solution of local explanation from additive feature importance measure class ensuring desirable theoretical properties [55]. A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. The objective is to fairly distribute the payout around all the features to obtain the prediction. We can make the following correspondence between coalitional game theory and model interpretability:

- The features are the players
- The model is the game
- The feature attribution is the gain attribution

Following the links with collaborative game theory, Shapley's values are the only indicators verifying local accuracy, i.e. that the sum of the feature contributions is equal to the prediction, missingness, i.e. if a feature has not had an impact on the output of the model, then its contribution will be zero and symmetry, i.e. that if two features have an identical effect when observed in any situation, then the Shapley values for the features must be the same.

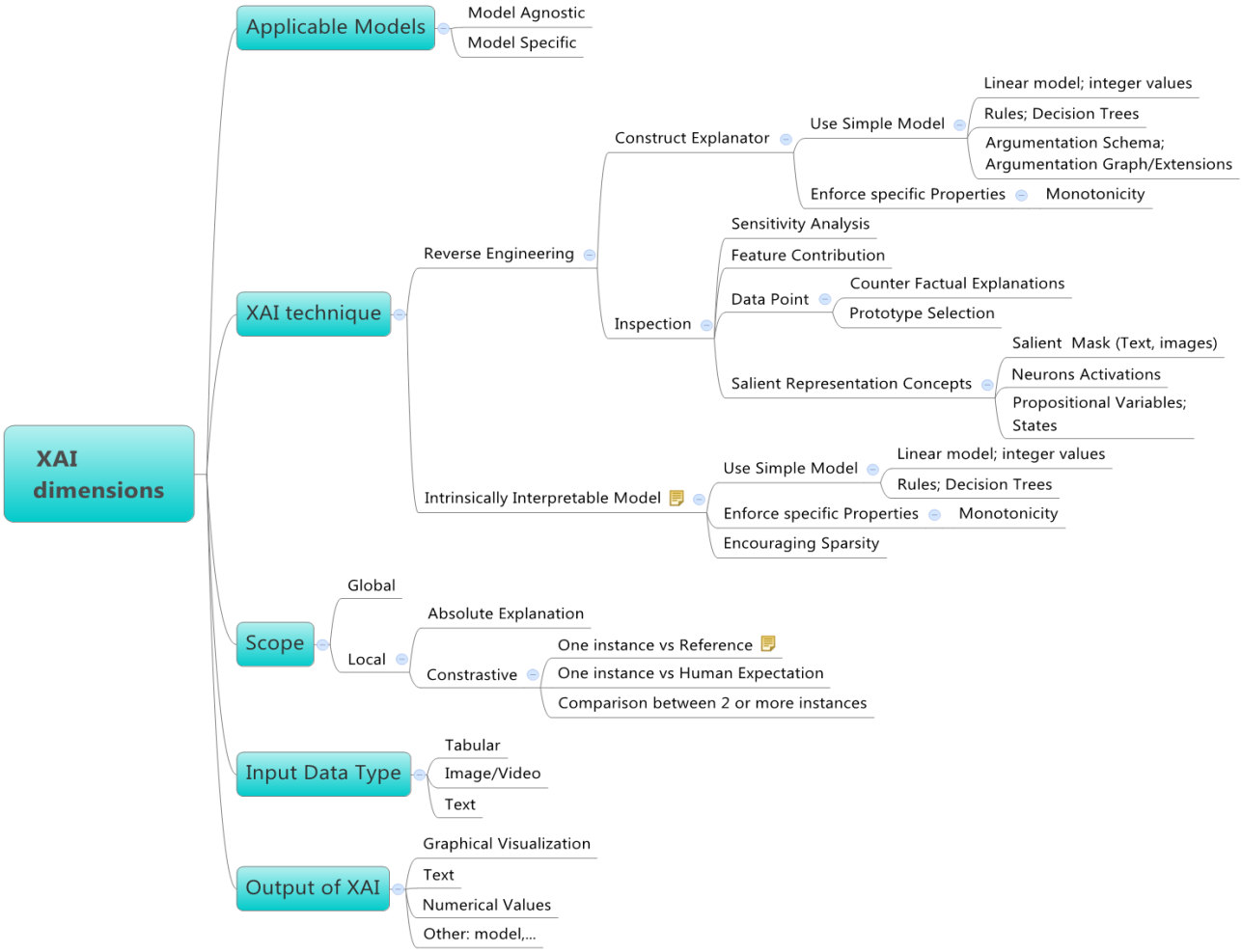


Figure 25: xAI Taxonomy

Because of all the coalitions, Shapley values are very time-consuming to compute. There exist some efficient tricks to compute Shapley value in the case of some particular models (Linear model, tree ensemble [56]). [55] proposes Kernel Shap, a model-agnostic algorithm to approximate Shapley values. It is an adaptation of LIME with given weight, loss and penalty which allows efficient Shapley value estimation. Another technique is based on a Monte Carto approximation [34]. We begin to implement these techniques in a Python library (https://github.com/ThalesGroup/shapkit).

Moreover, basically, Shapley Values are computed when the baseline level is considered by average model prediction. As in [60], we work on Shapley Value computation when the baseline level is not the average prediction, but is obtained by considering a reference sub-population. For instance, in a cyber security use case of detection of abnormal activities using log proxy information, for an abnormal IP address, the classical computation will compute the Shapley Value in function of the average score. It could be more interesting to compute it versus the normal IP address. It allows contrastive explanation, which could be clearer.

3.4.2 *Development of techniques allowing to find efficient counterfactual examples in cybersecurity*

The counterfactual examples are about what could have been, but is not. This is the description of the smallest change to change the forecast to reach a target value. A good counterfactual explanation is a realistic individual with a forecast close to the target value and close to the observation to be explained, with as few variables as possible modified.

The difficulty of the method is to generate a relevant, realistic and understandable counterfactual example. [57] proposes to solve the optimization problem given below:

$$\arg\min_{x'} \max_{\lambda} \lambda \left(\hat{f}(x') - y'\right)^2 + d(x, x'),$$

where

- x' the counter-example
- y' the target value (chose by the user)
- $\hat{f}$ the estimator to explain and $\hat{f}(x')$ the prediction of this estimator for the counter-example
- d(x, x') a distance between the observation to explain and the counter-example

The first term of the expression measures the distance the prediction of the counter-example with the target value (here measured by the Euclidean distance). The second term measures the distance from the counter-example with the observation, which the authors propose to calculate by a Manhattan distance weighted by the inverse of the medians of absolute deviations (thus introducing sparsity). The term lambda is a compromise term between the two terms of the equation. The closer it is to 0, the more important is the proximity from the counter-example at the expense of the first term.

There is no guarantee that optimization problem has a solution, nor that this solution is realistic.

[58] proposes a "growing" sphere approach to finding counterfactual examples in classification cases:

- Build a sphere around the observation
- If the sphere contains a single observation respecting the target prediction, keep it as a counter-example
- If the sphere does not contain any observations that do not meet the target prediction, enlarge the sphere and repeat steps 2-3-4
- If the sphere contains several observations respecting the target prediction, either keep them or narrow the sphere and repeat steps 2-3-4

There are two main approaches for generating counterfactual examples. The first one is a post-hoc method. The other concerns a model which is trained to produce good examples given an input and its prediction to explain.

This second approach could deal either with white-box and black-box models. Given an input, the model learns the perturbation it has to apply for creating a new individual sample the prediction of which is as close as possible to a desired and defined value. The challenge is to apply a small perturbation and guarantee that the new instance generated is realistic (i.e. could be generated by the same distribution of the training sample).

An example of such models is described by [59]. It is related to adversarial networks for training a generator model of adversarial examples.

Previous methods look for counter-examples. [54] seek the broadest possible rules to maintain the same classification, i.e. the set of rules for which the prediction of the model is (almost) always the same, regardless of the values of the features not involved in the rule. The authors propose two approaches to construct these rules.

The counterfactual example is one of the possibilities that can be explored during the project.

3.5 Achieving Explainability of an Intrusion Detection System by the Hybrid Oracle-Explainer Approach in the Cybersecurity Domain

This section describes a method for achieving explainability of Intrusion Detection System using the hybrid Oracle-Explainer Approach. The proposed solution is called Hybrid Oracle-Explainer Intrusion Detection System. It uses two separate modules to deliver human interpretable answers about system decisions, at the same time allowing for highest possible accuracy.

This section shows its fundamental assumptions, scheme and detailed description. To support all of that, an early prototype has been delivered and tested.

3.5.1 *Explainable Artificial Intelligence in the Context of Intrusion Detection Systems*

There are a few concerns about xAI that must be stressed in the context of Intrusion Detection Systems (IDS) and cybersecurity in general (which are our domains/application of interest in this work).

During the design of an AI (or machine learning based detection) system for cybersecurity there are a lot of aspects that must be taken into consideration. A developer should know the answers to the "Six Ws" [61] (Who? What? Where? When? Why? How?) in order to deliver reliable, secure and useful solutions (e.g. explanation for alarms, detected anomalies and so called IoC (Indicator of Compromise)) for all the stakeholders (e.g. security operators in SOCs (Security Operations Centres)).

As for xAI in cybersecurity context, we agree that the use of interpretability should not, under any circumstances, lead to any decrease in model performance, i.e. introduce vulnerability. As stated in [62], there are possible dangers to transparency delivered by an incorrectly designed model.

For example, there is a difference between target audience and system beneficiaries [62], as it is possible that by gaining insights into model learning functions, we can gain the means to manipulate it. While in the context of recommendation system it does not really matter, it can compromise the whole IDS.

Besides, there is an issue of accuracy and/or efficiency ahead, that the xAI methods can have [62]. Of course, with an IDS it is crucial to have as accurate a model as possible in order to deliver protection and threat mitigation. Therefore, xAI in the context of cybersecurity should be treated more as a means of reaching the end [62], which is to foster trust and reduce risk of unwanted,

unknown behaviour, rather than a goal on its own. This idea is the foundation and motivation for the solution proposed in this work.

Therefore, in the context of IDS and cybersecurity, there is a need for a system that fulfils the following conditions:

- Delivers reliable predictions about potential threats,
- Delivers easy to understand explanations about its decisions,
- Keeps flexibility necessary to adapt program towards new challenges,
- Meets all of the above without detrimental effect on the performance.

3.5.2 *Three Principles*

The model proposed in the further part of this section is based upon three important assumptions:

1) In the context of IDS, the accuracy and reliability of a system are the top priority.
2) One phenomenon can have more than one explanation, a.k.a the Rashomon effect [64].
3) The delivered explanation should be simple and help to develop trust [65].

Because of those principles it was decided that a surrogate type system with local explanations may be the best solution. It has low overhead and no impact on accuracy, therefore it realises the principle number one. The Rashomon effects makes an approach valid. Though the derived explanation is not a faithful representation of the opaque classifier function in general, it is a potentially possible approximation of it. Therefore, it still provides useful insights into the data and helps to develop trust. Finally, because of its model agnostic and modular approach, it allows to freely use a wide range of explanatory methods and as a consequence to tailor the explanation to any potential user.

In other words, this proposed method sacrifices, to some degree, the first point of "xAI Desiderata", i.e. fidelity (the explanation must be a reasonable representation of what the system actually does) presented in [63], to better realise the rest of them (understandability, sufficiency, low construction overhead and efficiency) and to fully solve the problem described in the previous subsection.

3.5.3 *Model Overview*

Figure 26 reveals the general scheme of Hybrid Oracle-Explainer IDS solution.

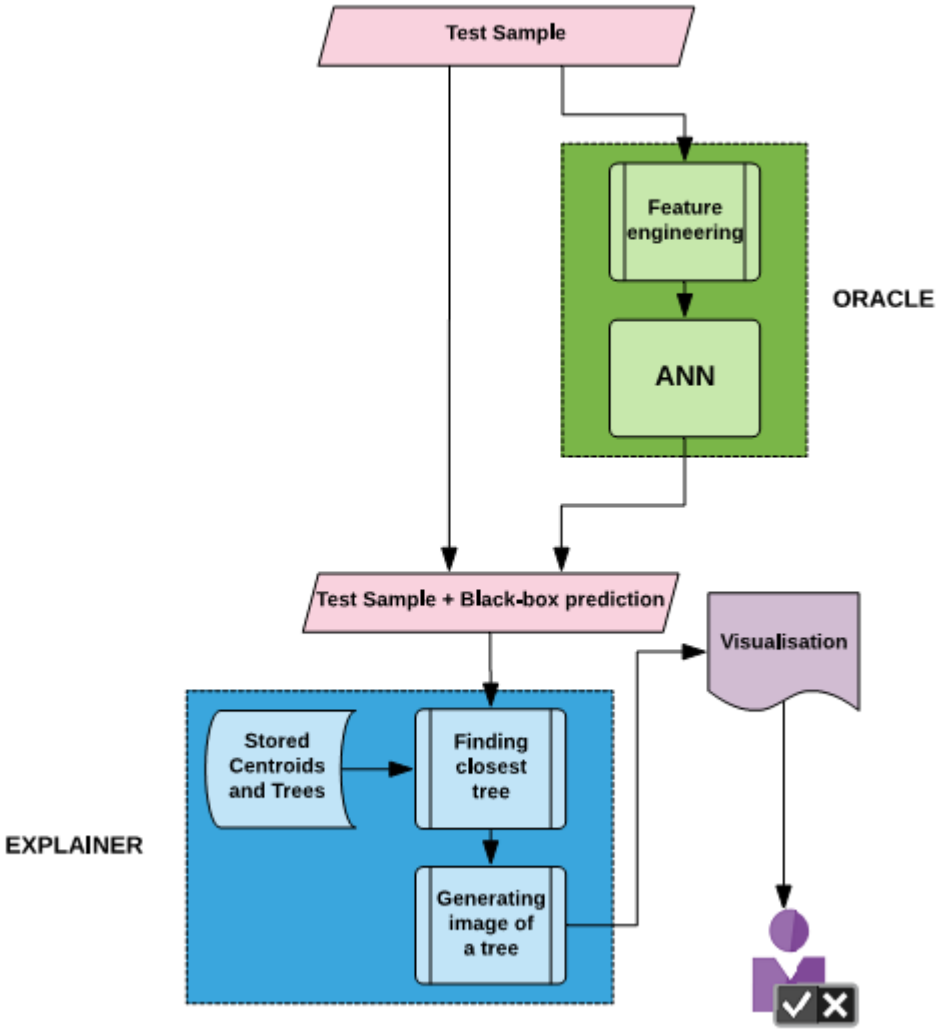


Figure 26: Proposed system overview

The chosen sample is first being transformed to the form used by the opaque classifier during training. In this case, the role of the black-box machine learning algorithm is fulfilled by a Feed Forward Artificial Neural Network (ANN).

Then, after obtaining a prediction, the sample in its original form, along with the Oracle output, is being passed to Explainer module. There it is compared with the saved centroid of each cluster made during the training process in order to find n closest (most similar) in terms of l^2 (Euclidean) norm.

Following that, starting with the closest centroid, Decision Tree trained on the according cluster is being retrieved. If its prediction matches that of the Oracle, the search stops and the local explainer is returned. Otherwise, the algorithm continues until it finds a supporting Tree or runs out of centroids. In that case the Tree linked to the closest centroid is returned.

This introduces a divergence in some cases and development of a strategy to minimise and properly handle this is a part of the future work. Next, the scheme of the decision tree is being drawn, resembling the one in Figure 27, but with a highlighted path to prediction made by the chosen explainer.

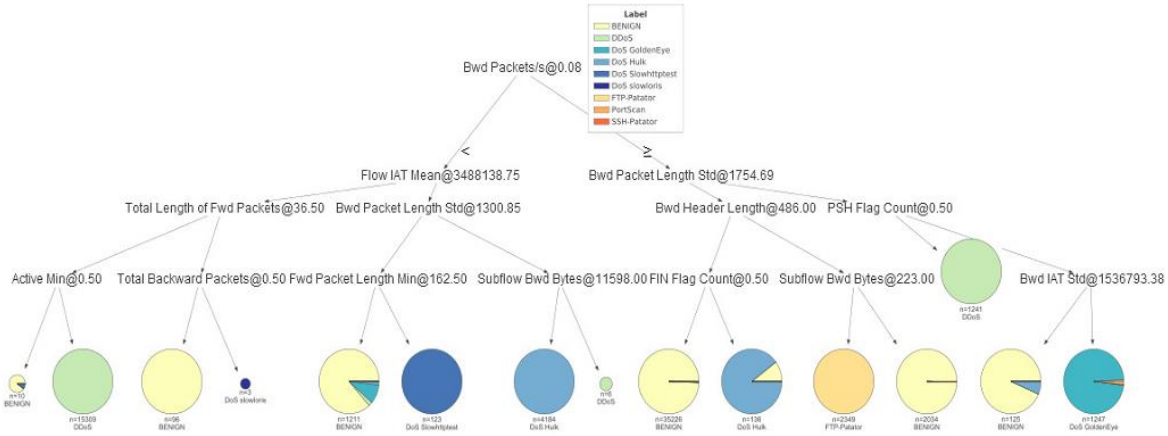


Figure 27: An example of a Decision Tree trained on CICIDS2017 dataset using microaggregation method

The created visualisation is then presented to the security analyst, who uses it to understand why the chosen sample could be classified in such a way and/or to obtain a better understanding of the potential threat's characteristics.

3.5.3.1 Data Preparation

Because the training data for both main modules must be the same, some standard parts of machine learning pipeline must be carried beforehand. It includes data cleaning, formatting, balancing samples and feature selection. Afterwards, the dataset is split to a training set and a testing set, which are saved as files accessed by both modules.

3.5.3.2 Oracle Module

This part of the solution is relatively straightforward, being a standard machine learning pipeline oriented toward maximised precision. It means that most feature engineering methods and transformations can be used, along any classifier. In our prototype, an ANN with Principal Component Analysis (PCA) is being adopted as an example (since we have a running IDS/cybersecurity system based on ANN).

3.5.3.3 Explainer Module

It should be reminded that because of both the modular and the agnostic nature of the whole system, the presented implementation is not the only valid one. It can be, like Oracle, changed to another or even expanded upon with additional algorithms; of course, as long as they are model agnostic and with a local scope. Experimentation with different explainers and their potential compositions is part of the future work.

The training procedure strictly follows the structure presented in [66].

For the readers' convenience it is presented here as Figure 28. The number and the size of clusters is controlled by the parameter *k*, which indicates the level of representativity. The higher its value, the bigger the clusters, and therefore, there are fewer of them there.

Algorithm 1 Generation of cluster-based explanations

procedure CLUSTER(Training set **X**)
    Compute a clustering $C(X)$ for X based on all attributes except the class attribute
    for *each cluster* $C_i \in C(X)$ **do**
        Compute a representative, e.g. the centroid of average record $\tilde{c}_i$
    end for
    for *each cluster* $C_i \in C(X)$ **do**
        Train an interpretable model, such as a decision tree DT_i
    end for
end procedure

Figure 28: Generation of cluster-based explanations

To compute the clusters, the method uses a microaggregation heuristic named the Mean Distance to Average Vector (MDAV). A detailed description is available in [67], while the algorithm can be found in [66].

The prepared train set and test set are imported. No additional transformations are performed, so clusters are generated directly on the training set. Having centroids, clusters and trees saved, the procedure of finding explanation for chosen sample follows the algorithm [66] presented in Figure 29.

Algorithm 2 Guided provision of explanation

Require: list of centroids C, list of interpretable models DT
procedure GUIDED EXPLANATION($sample, prediction, n$)
    for *each centroid* $C_i \in C$ **do**
        calculate Euclidean distance $dist(sample, C_i)$ and add result to the dictionary $dict(C_i, dist(sample, C_i))$
    end for
    using dictionary sort C, where C_1 is the closest representative
    define iterator $i = 0$
    while $i < n$ **do**
        take interpretable model DT_i corresponding to the C_i
        if decision $(d = DT_i(sample)) == predictions$ **then**
            return d, C_i, DT_i
        else
            $i = i + 1$
        end if
    end while
    return d, C_1, DT_1
end procedure

Figure 29: Guided provision of explanation

Next, as mentioned before, the samples with the retrieved tree are handled to the function of the library dtreeviz [68], which is responsible for generating the visualisation.

Chapter 4 Fairness ensuring mechanisms

4.1 Technical viewpoint

4.1.1 *Fairness in AI - state of the art*

As the fairness in ML/AI is a trending topic, there are many algorithms focusing on improving fairness in ML described in the literature. Most of them fall into three categories: preprocessing, optimization at training time, and post-processing [69]. In general, algorithms belonging to the same category are characterized by common advantages and flaws.

4.1.2 *Preprocessing:*

The idea is based on building a new representation of input data by removing the information correlated to the sensitive attribute and at the same time preserve the remaining input information as much as possible. The downstream task (e.g. classification, regression, ranking) can thus use the "cleaned" data representation and produce results that preserve demographic parity and individual fairness.

In [72], authors use the optimal transport theory to remove disparate impact of input data. They also provide numerical analysis of the database fair correction. In [73], authors propose a learning algorithm for fair classification addressing both group fairness and individual fairness by obfuscating information about membership in the protected group. Authors of [74] propose a model based on a variational autoencoding architecture with priors that encourage independence between sensitive and latent factors of data variation. To remove any remaining dependencies, an additional penalty term based on the "Maximum Mean Discrepancy" (MMD) measure is additionally introduced. A statistical framework for removing information about a protected variable from a dataset is presented in [75], along with the practical application to a real-life dataset of recidivism, proving successful predictions independent of the protected variable, with the predictive accuracy preserved. [76] proposes a convex optimization for learning a data transformation with three goals: controlling discrimination, limiting distortion in individual data samples, and preserving utility.

The advantages that are common for fair pre-processing algorithms include the possibility to use preprocessed data for any downstream task and no need to modify the classifier. There is also no need to access the sensitive attributes at testing time. In contrast, preprocessing algorithms can be used only to preserve statistical parity and individual fairness, and their performance in terms of accuracy and fairness trade-off are not as promising as in the case of two other groups of algorithms.

4.1.3 *Optimization at training time:*

Data processing at training time provides good performance on accuracy and fairness measure and ensures higher flexibility in optimizing the trade-off between these factors. The author of [1] describes that the common idea that can be found in the state-of-the-art works falling into this category of algorithms is to add a constraint or a regularization term to the existing optimization objective. Recent works considering algorithms to ensure ML fairness applied at the training time include: [77], where the problem of learning a non-discriminatory predictor from a finite training set is studied to preserve "equalized odds" fairness; [78] and [71], where a flexible mechanism to design fair classifiers by leveraging a novel intuitive measure of the decision boundary (un)fairness is introduced, and [79] that addresses the problem of reducing the fair classification to a sequence of cost-sensitive classification problems, the solutions of which provide a randomized classifier with the lowest (empirical) error subject to the desired constraints.

The disadvantages of the abovementioned approaches include the fact that these methods are highly task-specific and they require a modification of the classifier, which can be problematic in most applications/cases.

4.1.4 *Post-processing:*

The post-processing algorithms are focused on editing the posteriors to satisfy the fairness constraints and can be applied to optimize most of fairness definitions except the counterfactual fairness. The basic idea is to find a proper threshold using the original score function for each group. Some sample recent work that falls into this category is publication [70], in which the authors show how to optimally adjust any learned predictor to remove the discrimination according to the "equal opportunity" definition of fairness, with the assumption that the data about the predictor, target, and membership in the protected group are available.

The advantage of post-processing mechanisms is that retraining/changes are not needed for the classifier (the algorithm can be applied after any classifier). Another benefit comes in the form of good performance in the terms of fairness measures. The disadvantages include a need for test-time access to the protected attribute and the lack of flexibility in picking the accuracy–fairness trade-off [69].

4.1.5 *Summary:*

The author of [79] provides an experimental comparison of the selected algorithms applied to reduce unfairness using four real-life datasets with one or two protected sensitive attributes (gender or/and race). The selected methods include preprocessing, optimization at training time and post-processing approaches. The methods that achieve the best trade-off between accuracy and fairness are those falling into the optimization at training time category, while the advantage related to the implementation of preprocessing and post-processing methods is the preservation of fairness without modifying the classifiers. On the whole, experimental results prove the ability to significantly reduce or remove the disparity, in general not impacting the classifier's accuracy for all the methods. The reduction methods (optimization at training time) used to preserve the demographic parity achieve the lowest constraint violations, outperforming or matching the baselines. The post-processing algorithm performs well for small violations. Pre-processing algorithms (based on reweighting and relabelling) achieve the worst fairness measures [69][79].

4.1.6 *Development of Machine Learning techniques which integrate bias correction*

4.1.6.1 Penalized Random Forest

Random Forest is a very popular classical machine learning techniques. [81] propose an extension applied to fairness. The main idea is to penalize direct and indirect effect of the protected attribute. At each division, the choice of the covariate where the division is made depends on the improvement for the target prediction, as in the classical Random Forest (the improvement has to be maximal) and the separation quality that this leads to the protected attributes (which must be minimal). The covariate which achieves the best trade-off between the two previous objective is selected.

4.1.6.2 Adversarial Network

[80] develop an Adversarial Network to limit the effect of nuisance parameters on an output of a neural network. This architecture, shown in the next Figure, has been reused in the context of Fairness by some authors (see https://blog.godatadriven.com/fairness-in-ml). Two networks are combined. The first one achieves a classification task (e.g. obtaining a loan). The second one uses the probability predicted by the first neural network and tries to predict the attribute the impact of which (on the classifier output, e.g. the gender) we want to remove.

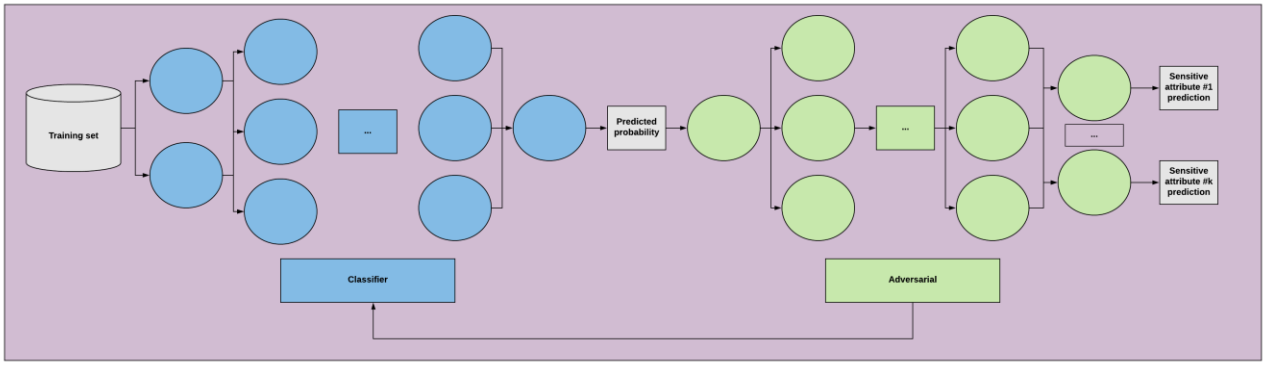


Figure 30: Fair Adversarial Neural Network

Note x the classifier prediction, based on input X, y the true value and Z the protected attribute (the attribute impacting x, the impact of which we want to remove). Note **a** and **b** the parameters of respectively the classifier (first network of previous Figure) and the adversarial network (the second). Note Ly(**a**) and Lz(**a**, **b**) the loss of pre-trained classifier and adversarial network. During the iteration, the following objective functions are considered:

$$\boldsymbol{a}, \boldsymbol{b} = \arg min_{\boldsymbol{a}} max_{b} (Ly(\boldsymbol{a}) - \alpha Lz(\boldsymbol{a}, \boldsymbol{b})$$

a and **b** are updated iteratively, one at a time.

4.2 Law, regulatory and ethical viewpoint*Methodology*

The objective of the Task 7.4 is the development of a systematic method to prevent machine learning and the decisions based on AI systems from being used to support discrimination.

We can see that artificial intelligence and machine learning have generated a lot of debate and lively discussion, both at the level of the European institutions. One of the objectives of this section is to analyse this EU "package" regarding artificial intelligence and how the notion of fairness is understood at the European level.

Moreover, artificial intelligence calls for a multidisciplinary approach in that it involves legal, political, economic, technical and social issues. One of the main challenges will be to collect major political statements and scientific papers regarding the notion of fairness in order to provide a basis for evaluating the performance of an AI system.

However, it also seems important not only to focus on what is specific to artificial intelligence. The notion of fairness seems closely linked to the compliance with legal standards. In this regard, regulatory norms have an important impact on the development of artificial intelligence mechanisms. We begin by taking the example of the GDPR.

4.2.2 *Fairness and artificial intelligence: State of the Art*

4.2.2.1 EU "Package"

In January 2017, the European Parliament called upon the Commission to assess the impact of AI. Alongside, it released recommendations on civil law rules on robotics[3]. The European Parliament also drew up a code of ethics for robotics engineers and, in the meantime, asked for the creation of a European agency in charge of thinking about robotics and AI in the European Union. This agency has to include technical, ethical and regulatory experts[4].

In 2018, the European Commission released its first communication regarding AI matters: "Artificial Intelligence for Europe"[5]. It was followed by a second communication, which focused on a coordinated plan on artificial intelligence[6].

In April 2019, the European Union issued its artificial intelligence "package". This includes[7]:

- A study by the European Parliamentary Research Service untitled "A governance framework for algorithmic accountability and transparency"
- The results of the work from the High-level Expert Group on Artificial Intelligence:
  - "A definition of AI: main capabilities and disciplines"
  - "Ethics Guidelines for Trustworthy AI"
  - "Policy and Investment Recommendations for Trustworthy AI"
- A communication from the European Commission "Building Trust in Human-Centric Artificial Intelligence"

In addition to that, very recently, the European Consumer Organization (BEUC) released a position paper called "AI Rights for Consumers"[8]. This paper claims new rights for consumers in the European Union:

[3] European Parliament, "Recommendations to the Commission on Civil Law Rules on Robotics", 27.01.2017, available at: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect.
[4] EPRS, "EU guidelines on ethics in artificial intelligence: Context and implementation », p. 2.
[5] European Commission, "Artificial Intelligence for Europe", 25.04.2018, COM(2018) 237 final.
[6] European Commission, "Coordinated Plan on Artificial Intelligence", 07.12.2018, COM(2018) 795 final.
[7] All the documentation is available at: https://ec.europa.eu/digital-single-market/en/artificial-intelligence.
[8] BEUC, "AI Rights for Consumers", available at: https://www.beuc.eu/publications/beuc-x-2019-063_ai_rights_for_consumers.pdf.

- Right to transparency, explanation and objection: "*Consumers should have a right to get a clear picture of how decisions that affect them are made and be able to oppose wrong or unfair decisions and request human intervention*"
- Right to Accountability and control: "*Consumers should have a right that appropriate technical and organisational systems as well as measures are put in place that ensure legal compliance and regulatory oversight*"
- Right to fairness: "*Consumers should have a right that algorithmic decision making is done in a fair and responsible way*"
- Right to non-discrimination: "*Consumers should have a right to be protected from illegal discrimination and unfair differentiation*"
- Right to safety and security: "*Consumers should have a right that AI-powered products are safe and secure throughout their lifecycle*"
- Right to access to justice: "*Consumers should have a right to redress and public enforcement if risks associated with artificial intelligence materialise*"
- Right to reliability and robustness: "*Consumers should have a right that AI powered products are technically reliable and robust by design*"

Furthermore, in September 2019, the European Parliamentary Research Service (EPRS) sorted out a document called "EU guidelines on ethics in artificial intelligence: Context and implementation"[9].

Finally, Ursula von der Leyen, the new president of the European Commission, engaged herself to the following: "In my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence. This should also look at how we can use big data for innovations that create wealth for our societies and our businesses. I will make sure that we prioritise investments in Artificial Intelligence, both through the Multiannual Financial Framework and through the increased of public-private partnerships"[10].

Although the President von der Leyen did not provide policy proposals within the first hundred days of her mandate, the Commission since issued its white book on AI which proposes policy options, and was followed by a broad stakeholders consultation[11].

We can note that this enthusiasm at the European Union level is matched by an important interest in the doctrine for the issue of "fairness and artificial intelligence"[12].

4.2.2.2 An attempting definition of Artificial Intelligence

[9] Available at: http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163

[10] https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

[11] European Commission, "White Paper On Artificial Intelligence -A European approach to excellence and trust", 19.02.2020, COM(2020) 65, final.

[12] See notably KROLL, J., HUEY, J., BAROCAS, S., FELTEN, E., REIDENBERG, J., ROBINSON, D., and YU, H., "Accountable Algorithms", *University of Pennsylvania Law Review*, 2017 vol. 165, pp. 633 – 706 ; AI Now, "AI Now Report 2018", December 2018, HACKER, P., "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law", *Common Market Law Review*, 2018, vol. 55, pp. 1143 – 1186 ; Data & Society, "Algorithmic Accountability: a Primer", 18 April 2018 ; SELBST, A., and BAROCAS, S., "Big Data's Disparate impact", *California Law Review*, 2016, vol. 104, pp. 671-732 ; PASQUALE, F., CITRON, D., "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, 2014, vol. 91, pp. 1-33 ; F. DOSHI-VELEZ and K. MASON, "Accountability of AI Under the Law: The Role of Explanation", *Berkman Klein Centre Working Group on Explanation and the Law, Berkman Klein Centre for Internet & Society*, 2017, pp. 1-15 ; EDWARDS, L., and VEALE, M., "Enslaving the algorithm: from a 'right to an explanation' to a 'right to better decisions'?", *IEEE Security and Privacy Magazine*, 2018, n.° 16, pp. 46-54 ; Naudts, L, Towards Accountability: The Articulation and Formalization of Fairness in Machine Learning (August 5, 2018). IFIP Summer School on Privacy and Identity Management "Fairness, Accountability and Transparency in the Age of Big Data" (20-24 August 2018) (submitted for pre-proceedings). Available at SSRN: https://ssrn.com/abstract=3298847 or http://dx.doi.org/10.2139/ssrn.3298847 ; Naudts, Laurens, How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them (2018). R. Leenes, R. van Brakel, S. Gutwirth & P. De Hert (Authors), Data Protection and Privacy: The Internet of Bodies (Computers, Privacy and Data Protection) 2019. Available at SSRN: https://ssrn.com/abstract=3468121.

The European Commission defines Artificial Intelligence (AI) as: *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)"*[13].

The European Parliamentary Research Service defines AI as: *"AI commonly refers to a combination of: machine learning techniques used for searching and analysing large volumes of data; robotics dealing with the conception, design, manufacture and operation of programmable machines; and algorithms and automated decisions-making systems able to predict human and machine behaviour and to make autonomous decisions"*[14].

The independent High-level Expert Group on Artificial Intelligence (set up by the European Commission) defines AI as: *"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)"*[15]

4.2.2.3 An attempting definition of fairness: the study from the European Parliamentary Research Service

4.2.2.3.1 An attempting definition

In its communication of 8 April 2019, the European Commission fears that artificial intelligence might support discrimination: "*Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to (in)direct discrimination. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition. Moreover, the way in which AI systems are developed (e.g. the way in which the programming code of an algorithm is written) may also suffer from bias. Such concerns should be tackled from the beginning of the system' development*"[16].

Therefore, we could assume, as a starting point, that fairness is equivalent to the absence of bias in algorithmic processing. In fact, it appears that the concept of fairness is richer than that.

Indeed, a study makes by the European Parliamentary Research Service called "A governance framework for algorithmic accountability and transparency", states the following observation: "*Fairness turns out to be a multi-faceted, and inherently complex concept. Given this, it is difficult to articulate in a single definition and may also be subject to competing definitions. Fairness reflects the appreciation of a situation based on a set of social values, such as promoting equality in society. The assessment of fairness depends on facts, events, and goals, and therefore has to be understood as situation or task-specific and necessarily addressed within the scope of a*

[13] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final.
[14] http://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf
[15] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
[16] European Commission, "Building Trust in Human-Centric Artificial Intelligence", 08.04.2019, COM(2019) 168 final, p. 6.

practice (…). The concept of fairness in the context of algorithmic implementations appears as a balance between the mutual interests, needs and values of different stakeholders affected by the algorithmic decisions"[17].

It therefore seems that the concept of fairness refers to the development of artificial intelligence that supports European values, in particular through the absence of bias. Various studies explore the role of algorithms to support European values, such as pre-processing algorithms, algorithms to optimize at training time and post-processing algorithms. For example, the European principle of equity could be enhanced by the use of pre-processing algorithm. Furthermore, the optimization at training time algorithms or post-processing algorithms could be used to test the equality of opportunity in AI system. Finally, the use of these algorithms, jointly or separately, can also increase user confidence in AI system[18].

In particular, based on case studies, the European Parliamentary Research Service identifies 8 European values that can be impacted by the use of algorithms[19]:

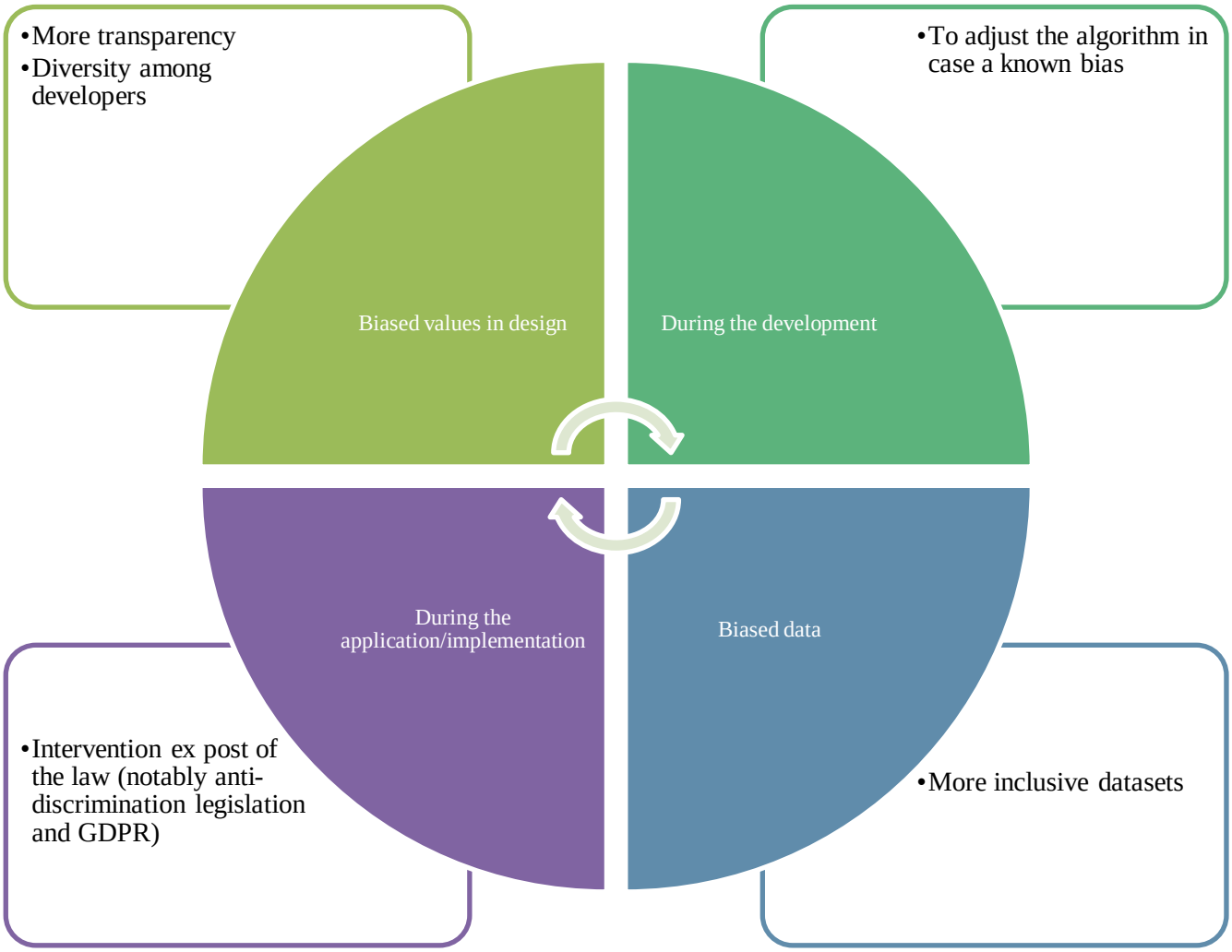
- Equality of opportunity/equality of outcome
- Equity
- Freedom of choice
- Justice
- Truth
- Trust
- Autonomy
- Privacy

The European Parliamentary Research Service states that: "*Each of these values is closely entwined with understandings of fairness and social justice. Therefore, where one or more of the values is undermined, it is possible that protests will arise stating that the algorithmic process connected to it is in some way unfair*"[20].

A first approach to the notion of fairness thus seems to be emerging.

[17] EPRS, "A governance framework for algorithmic accountability and transparency", p. 10.
[18] See in this chapter, part 4.1 Technical viewpoint.
[19] EPRS, "A governance framework for algorithmic accountability and transparency", p. 19.
[20] EPRS, "A governance framework for algorithmic accountability and transparency", p. 19.

4.2.2.3.2 Sources of unfairness and solutions[21]



4.2.3 *Guidelines from the independent High-level Expert Group on Artificial Intelligence*[22]

4.2.3.1 Overview

After having identified the European values that could be impacted by the development of artificial intelligence, the High-level Expert Group on Artificial Intelligence developed guidelines for a trustworthy AI. These guidelines are grouped into three themes/principles. Compliance with these principles would mean supporting European values and therefore considering the algorithms as *fair*. These characteristics are[23]:

- the lawfulness of the system, the artificial intelligence mechanisms must comply with all applicable regulations
- the ethics of artificial intelligence
- technical and social robustness

[21] See EPRS, "A governance framework for algorithmic accountability and transparency", p. 20 and following.
[22] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top
[23] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 6-7.

Focus: the human-centric approach

In its explanatory document of the guidelines from the High-level Expert Group on Artificial Intelligence called "EU guidelines on ethics in artificial intelligence: Context and implementation", the European Parliament insists on the fact that the human-centric approach is the core principle of the guidelines and explains:

"*The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come*"[24].

In this regard, all AI stakeholders have to adopt this approach. Therefore, all the stakeholders must be concerned about the respect of fundamental rights at all stages of the development of an artificial intelligence tool.

This view seems to be supported by the European Consumer Organisation (BEUC). Indeed, according to the BEUC, fairness must be a combination of compliance with applicable laws and ethics concerns. According to the document called "AI Rights for Consumers", the right to fairness is: "*decision-making processes must be fair from the perspective of the data that is processed, the means used in the decision process and the intention of what do to with the result. The outcome should be fair too, hence the result should not lead to an unjust treatment or behaviour. The latter aspect is not fully addressed by EU data protection law, which focuses on the fair processing of personal data but not on the consequences resulting from predicting analysis. Modern rules should therefore focus on processing results, to prevent unfairness, deception and manipulation stemming from algorithmic inferences and mathematical-statistical methods. Businesses practices must be fair too: the use of algorithms should never lead to consumers being deceived or impaired in their freedom of choice. Their expectations should be protected and their weak position with regards to the business be respected. However, current rules on unfair market practices do not sufficiently consider consumer detriment associated with algorithmic decision making. Questions of fairness should also be seen under the aspect of general welfare considerations. A lack of fairness can foster greater societal asymmetries, lead to unequal benefits for citizens or could even lead to certain groups of people being exposed to higher risks of poverty. The deployment of AI systems must thus consider their impact on the well-being of citizens*".

4.2.3.2 Ethics in AI

In order to determine the ethical principles necessary to ensure the development of trustworthy artificial intelligence in the European Union, the High-Level Expert Group on Artificial Intelligence is based on respect for fundamental rights (European Union treaties, the Charter of Fundamental Rights of the Union and international law). According to the High-level Expert Group on Artificial Intelligence: *"Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI"*[25].

[24] http://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf
[25] AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf%20(1).pdf, p. 9.

Four "ethical imperatives" are recognised by the High-level Expert Group on Artificial Intelligence[26]:

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

4.2.3.3 Robustness and AI

Four criteria are used to ensure and evaluate the robustness of an artificial intelligence system[27]:

- Resilience to attack and security; "*AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking. Attacks may target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both software and hardware*"[28]
- Fallback plan and general safety; *"AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action. It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimisation of unintended consequences and errors. In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established"*[29]
- Accuracy; "*Accuracy pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions"*[30]
- Reliability and Reproducibility; "*A reliable AI system is one that works properly with a range of inputs and in a range of situations. This is needed to scrutinise an AI system and to prevent unintended harms. Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions*".[31]

4.2.3.4 Law and AI

In a rather brief way, the High-level Expert Group on Artificial Intelligence emphasises the fundamental importance of respecting privacy and the protection of personal data[32]. Thus, the High-level Expert Group on Artificial Intelligence calls for:

- Personal data protection throughout the entire life cycle of the system
- Concern and assurance for data quality and integrity
- The implementation of a protocol on access to personal data.

However, in its report, the High-level Expert Group on Artificial Intelligence emphasizes that it does not intend to focus on the first characteristic for a trustworthy AI, namely the legality of the system. In this introductory report to the notion of fairness, we therefore intend to supplement its guidelines by including an analysis of the legal framework for the protection of personal data. There are two reasons for this choice. First, it is not uncommon for artificial intelligence systems to use and

[26] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 11 and following.
[27] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 16 and following.
[28] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 16.
[29] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 16-17.
[30] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 17.
[31] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 17.
[32] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 17.

process a significant amount of personal data. Indeed, in its study called "A governance framework for algorithmic accountability and transparency"[33], the European Parliamentary Research Service underlines that: "*A key concern in the expansion of AI 's scope is its effects on personal privacy, which is a fundamental human right*"[34] Secondly, we will see that the GDPR also refers to the notion of "fairness".

4.2.3.5 Trustworthy AI: Assessment list

The High-level Expert Group on Artificial Intelligence has established a trustworthy AI assessment list to help all the stakeholders. The list is available on their website[35]. The criteria are:

- Human agency and oversight; "*AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight*"[36].
- Technical robustness and safety: "*A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm. Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured"*[37].
- Privacy and data governance; Privacy is a fundamental right affected by AI systems. "*Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy*"[38].
- Transparency; The transparency of elements relevant to an AI system (the data, the system and the business models) refers to traceability, explainability and communication[39].
- Diversity, non-discrimination and fairness; *"In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment*"[40].
- Societal and environmental well-being; *"the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern,*

[33] Available at: http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262
[34] EPRS, "A governance framework for algorithmic accountability and transparency", p.19.
[35] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. See pp. 26 and following.
[36] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 15-16.
[37] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 16-17.
[38] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 17.
[39] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 18.
[40] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 18-19.

such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations"[41].

- Accountability; "*It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use*"[42]. It includes auditability, minimisation and reporting of negative impacts, trade-offs and redress.

For each criterion listed above, the High-level Expert Group on Artificial Intelligence sets out different questions to be answered by, mainly, developers and deployers of AI systems in order to evaluate their systems[43].

4.2.3.6 To summarize:

As explained by the European Parliament, "*The guidelines are addressed to all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI in the EU, including companies, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers. Stakeholders can voluntarily opt to use these guidelines and follow the seven key requirements (see box on the right) when they are developing, deploying or using AI systems in the EU*"[44].

Source: *Independent High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*[45]

[41] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 19.
[42] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", pp. 19-20.
[43] An update of this assessment list is expected for mid-2020; European Commission, White Paper on Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final, p. 9.
[44] http://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf
[45] See also the European Commission's reference to the work of the High-level Expert Group on Artificial intelligence; Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Building Trust in Human-Centric Artificial Intelligence, Brussels, 8.4.2019, COM(2019), 168 final.

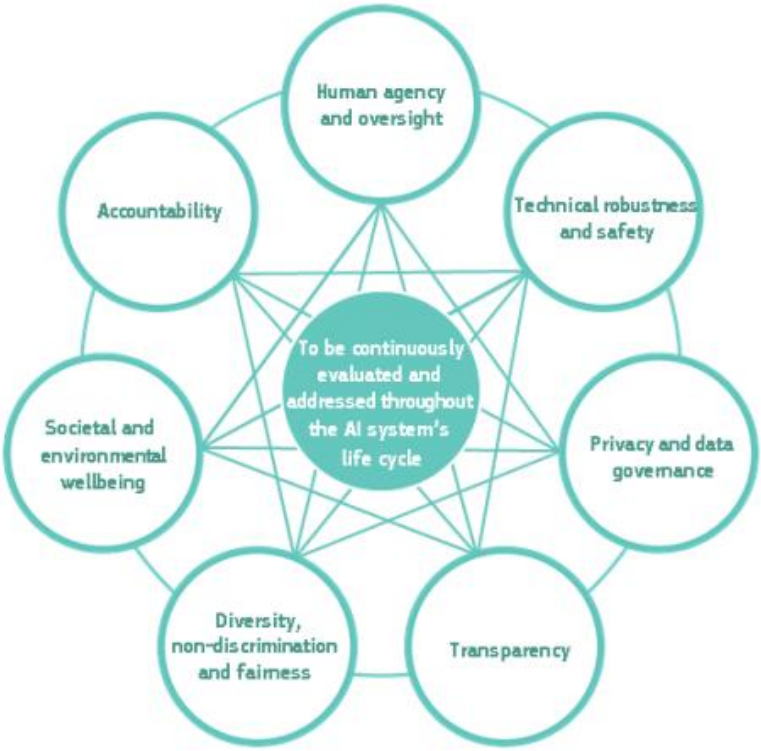


Figure 31: Interrelationship of the seven requirements

4.2.4 *Fairness in the General Data Protection Regulation*

Recital 39 of the General Data Protection Regulation begins with "Any processing of personal data should be lawful and fair". The legislator follows as such: « *It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Every reasonable step should be taken to ensure that personal data which are inaccurate are rectified or deleted. Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including for preventing unauthorised access to or use of personal data and the equipment used for the processing"*.

As highlighted by the European Data Protection Board (EDPB), "*Fairness is an overarching principle which requires that personal data shall not be processed in a way that is detrimental, discriminatory, unexpected or misleading to the data subject. Measures and safeguards implementing the principle of fairness also support the rights and freedoms of data subjects, specifically the right to information (transparency), the right to intervene (access, erasure, data*

portability, rectify) and the right to limit the processing (right not to be subject to an automated individual decision-making and non-discrimination of data subjects in such processes)"[46].

Following this statement, the EDPB, in its Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, gives two examples of a violation of the fairness principle due to the non-compliance with GDPR requirements[47]:

> Example 1
>
> A controller operates a search engine that processes mostly user-generated personal data. The controller benefits from having large amounts of personal data and being able to use that personal data for targeted advertisements. The controller therefore wishes to influence data subjects to allow extensive collection and use of their personal data.
>
> When implementing the fairness principle, taking into account the nature, scope, context and purpose of the processing, the controller realizes that they cannot present the options in a way that nudges the data subject in the direction of allowing the controller to collect more personal data than if the options were presented in an equal and neutral way. This means that they cannot present the processing options in such a manner that makes it difficult for data subjects to abstain from sharing their data, or make it difficult for the data subjects to adjust their privacy settings and limit the processing. The default options for the processing must be the least invasive, and the choice for further processing must be presented in a manner that does not deter the data subject from abstaining.

> Example 2
>
> Another controller processes personal data for the provision of a streaming service where users may choose between a regular subscription of standard quality and a premium subscription with higher quality. As part of the premium subscription, subscribers get prioritized customer service. With regard to the fairness principle, the prioritized customer service granted to premium subscribers cannot discriminate other data subjects' rights according to the GDPR Article 12. This means that although the premium subscribers get prioritized service, such prioritization cannot result in a lack of appropriate measures to respond to request from regular subscribers without undue delay and in any event within one month of receipt of the requests.
>
> Prioritized customers may pay to get better service, but all data subjects shall have equal and indiscriminate access to enforce their rights and freedoms according to the GDPR.

Therefore, the GDPR links the principle of a "fair" processing of personal data to compliance with several requirements.

4.2.5 *Criteria for fairness: a GDPR perspective*

In the deliverable D7.1, the Article 5 of the General Data Protection Regulation has already been considered. However, in this deliverable, the conditions of Article 5 to legalise the processing of personal data[48] are analysed and detailed from a specific perspective. The issue is to determine

[46] EDPB, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, 13 November 2019, available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf, p. 16.
[47] EDPB, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, 13 November 2019, available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf, p. 17.

[48] Personal data means any information relating to an identified or identifiable natural person. An identifiable natural person is the person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification

which conditions contribute to the development of a definition of fairness when personal data are at stake.

Fairness in the GDPR consists of two approaches. First, there is a formal approach with respect to procedures and legal requirements in the Regulation. Secondly, the data controller shall adopt an effect-based approach for the data subject to avoid vulnerability, discrimination and to prevent adverse effects[49]. Therefore, fairness includes to take into consideration both the effects on the data subjects and their expectations[50].

4.2.5.1 Transparency

> Article 5 GDPR: "Personal data shall be:
>
> (a) processed lawfully, fairly and in a transparent manner in relation to the data subject"

Transparency has a special status in data protection regulations. This is both an obligation for the controller and a right for the data subject. In addition, transparency is now a real challenge in the field of artificial intelligence given the complexity of its operation, making a whole technology and its implications opaque for citizens. We can distinguish a *procedural transparency* and a *fair transparency*. The procedural transparency occurs when the data controller respects the obligations set in the Regulation. On the contrary, to adopt a fair transparency, the data controller needs to consider the "reasonable expectations" for the data subjects in each concrete situation[51].

The additional question is about the deepness of transparency towards algorithms. Indeed, does the data controller have to be transparent about the algorithm itself (kind of open source) or about the processing using the algorithms? In other words, do we put the transparency at the level of the means (algorithms) or at the purpose one? Articles 13 and 14 about information specify that the information must be given about various elements. Amongst them, we find back the purpose but no word about the content of the means themselves including algorithms.

4.2.5.2 An obligation for the data controller

The GDPR requires the controller to provide different information to the data subject. This is a minimum requirement, as the GDPR imposes to the data controllers to provide other information relating to the processing of personal data to ensure the best possible understanding on the part of the data subjects[52].

number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (article 4.1 of the GDPR).

[49] Malgieri, Gianclaudio, The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation (January 10, 2020). Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, 14 pages. DOI: 10.1145/3351095.3372868. Available at SSRN: https://ssrn.com/abstract=3517264, pp. 155 and 156.

[50] See Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection*, 2017, 19-22, https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-dataprotection.pdf; Michael Butterworth, "The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework", *Computer Law & Security Review*, 34, no. 2 (1 April 20187): 257-68, https://www.sciencedirect.com/science/article/abs/pii/S026736491830044X?via%3Dihub, cited by Malgieri, Gianclaudio, The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation (January 10, 2020). Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, 14 pages. DOI: 10.1145/3351095.3372868. Available at SSRN: https://ssrn.com/abstract=3517264, p. 159.

[51] Malgieri, Gianclaudio, The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation (January 10, 2020). Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, 14 pages. DOI: 10.1145/3351095.3372868. Available at SSRN: https://ssrn.com/abstract=3517264, p. 157.

[52] Articles 13 and 14 of the GDPR.

The principle of transparency is attached to the loyalty expected of a controller. The collection and processing of personal data may not take place in an unfair, obscure or misleading manner[53].

The former Article 29 Working Party (now called the European Data Protection Board), which is the independent European working party that dealt with issues relating to the protection of privacy and personal data, elaborated several Guidelines to help the data controller to respect the transparency requirements[54]:

Characteristics[55]:

- user-centric. The data controller has to identify the "audience" and ensure that the information is understandable by an average member of this audience.
  In this regard, it is interesting to note that the European Commission and the High-level Expert Group on Artificial Intelligence plead for putting the human in the centre of artificial intelligence development[56]
- clear and plain language
- free of charge
- clearly separated from other non-privacy related information
- indication of where and how additional information can be obtained
- providing actively the information to the data subject

As best practice, the Article 29 Working Party encourages the data access to provide an easy access to the information related to the processing of personal data. The data controller has also to take into consideration the possibility to provide express reminders to data subject about the information notice[57].

What information?[58]

In any case, the data controller has to give the following details:

- The identity and the contact details of the controller and, where applicable, of the controller's representative
- The contact details of the data protection officer, where applicable
- The purposes of the processing for which the personal data is intended as well as the legal basis for the processing
- Where the processing is based on point (f) of Article 6(1), the legitimate interests pursued by the controller or by a third party
- The recipients or categories of recipients of the personal data, if any
- Where applicable, the fact that the controller intends to transfer personal data to a third country or international organization and the existence or absence of an adequacy decision by the Commission, or in the case of transfers referred to in Articles 46 or 47, or the second

[53] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 90.
[54] Art. 29 Working Party, Guidelines on transparency under Regulation 2016/679, 29.11.2017 (Revised and adopted on 11.04.2018), available at: https://ec.europa.eu/newsroom/article29/news.cfm?item_type=1360.
[55] Art. 29 Working Party, Guidelines on transparency under Regulation 2016/679, 29.11.2017 (Revised and adopted on 11.04.2018), pp.5 and following.
[56] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Building Trust in Human-Centric Artificial Intelligence, Brussels, 8.4.2019, COM(2019), 168 final.
[57] Art. 29 Working Party, Guidelines on transparency under Regulation 2016/679, 29.11.2017 (Revised and adopted on 11.04.2018), p. 18.
[58] Articles 12-14 GDPR.

subparagraph of Article 49(1), reference to the appropriate or suitable safeguards and the means by which to obtain a copy of them or where they have been made available.

In addition, the controller shall, at the time when personal data is obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:

- the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
- the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject or to object to processing as well as the right to data portability;
- where the processing is based on point (a) of Article 6(1) or point (a) of Article 9(2), the existence of the right to withdraw consent at any time, without affecting the lawfulness of processing based on consent before its withdrawal;
- the right to lodge a complaint with a supervisory authority;
- the existence of automated decision-making, including profiling, referred to in Articles 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

As explained by G. MALGIERI, "(…) *fair transparency seems to require additional efforts if compared to merely formal transparency, since it takes into account also 'reasonable expectations' of data subjects.(…) Actually, some scholars argued that fairness at Articles 5 and 6 GDPR is an 'ex ante" assessment on the average data subjects, while data subjects rights such as right to object and erasure (Articles 17 and 21) are based on an 'ex post' idea of fairness, tailored on specific circumstances* "[59].

4.2.5.3 A right for the data subject

The data subject has a right to transparency. Indeed, any data subject has the right to be informed of the existence of processing operations concerning him/her. This right is of particular importance in that the European Court of Justice has repeatedly recognised that the right to transparency is an essential prerequisite for the persons concerned to exercise their other rights. Indeed, it is only by being aware of the use of their personal data that the person can ask to have access to them, check what is being done, request a correction, etc[60].

In the field of artificial intelligence, where a lot of data is used and where the functioning may still seem obscure to citizens, transparency is all the more necessary. As stated by the GDPR, "*The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used. Such information could be provided in electronic form, for example, when addressed to the public, through a website. This is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what*

[59] MALGIERI, G., The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation (January 10, 2020). Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, p. 157 and p. 158. DOI: 10.1145/3351095.3372868. Available at SSRN: https://ssrn.com/abstract=3517264; See also Recitals 60 and 71 of the GDPR and D. CLIFFORD and J. AUSLOOS, 'Data Protection and the Role of Fairness', Yearbook of European Law 37 (1 January 2018): 130–87, https://doi.org/10.1093/yel/yey004 (cited by G. Malgieri).

[60] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 92 ; EUCJ, 1 October 2015, *Smaranda Bara*, C-201/14.

purpose personal data relating to him or her are being collected, such as in the case of online advertising"[61].

a) The right of access itself is also an implementation of transparency. It is therefore necessary that the data subjects not only have access to the information concerning them (this is a right recognised by the GDPR) but also that they can understand them and understand how artificial intelligence tools work[62]. **Transparency in AI**

It is interesting to note that the concept of transparency within the meaning of the GDPR and the legal obligations it imposes are similar to the transparency requirements identified by the High-level Expert Group on Artificial Intelligence. In the following table, we highlight the identical criteria while also noting their differences in their implementation:

GDPR	**Ethical Guidelines from the High-level Expert Group on Artificial Intelligence**
Accountability	Traceability
Right not to be subject to an automated decision However, - the right to explanation only applies if there is an automated decision-making process - is less strong in the degree of explanation than what the High-level Expert Group on Artificial Intelligence offers	Explainability
Article 12-14: Right on information for the data subjects Article 21: Right to object	Communication

4.2.5.4 Confidentiality

Article 5 Personal data shall be: (f)processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').

While transparency is an important part of privacy regulation, so is confidentiality. In fact, it is important to note that the transparency requirement only applies to the data subject.

[61] Recital 58 of the GDPR.
[62] High-level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI", p. 18.

In the same article requiring transparency on the part of the controller, the GDPR also insists on the integrity and confidentiality of data. Both the data controller and data processor are responsible for the security/confidentiality of the processing [63].

In addition, the contract concluded between the data controller and the data processor have to stipulate that the data processor ensures that persons authorised to process the personal data have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality[64].

Security and confidentiality are an integral part of a *fair* processing of personal data. Indeed, these two obligations ensure to the data subject that his or her information has not been manipulated by anyone.

Therefore, the notion of control also has an important place in the GDPR and is transversal. First, it is crucial that the data control keeps the control of the processing throughout the entire process. Indeed, the data controller has a predominant place throughout the use of personal data:

- Upstream,
    - When determining the means and purposes of the processing operation[65]
    - in the choice of a data processor. Indeed, the GDPR requires the controller to choose only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of the GDPR and ensure the protection of the rights of the data subject[66].
- Throughout the lifetime of the personal data
    - by complying with safety requirements[67]
    - by the organisation of regular audits of the data processor's activities[68]

Secondly, the controller must be able to control access to and circulation of personal information in order to guarantee the confidentiality of the data[69].

Thirdly, the control must also be exercised by the persons concerned. Once again, through the transparency requirements (as information, access) imposed on the controller, the GDPR ensures that the data subjects have all the necessary information to exercise the rights granted to them by the regulations[70].

4.2.5.5 Misuse of data

The purpose principle is the starting point for any processing of personal data. The purpose, i.e. the purpose for which the information is collected, must be determined from the outset by the controller, before any actual collection or operation on personal data takes place[71]. It seems that the definition of a purpose could serve the apparition of "fair" algorithm. As stated by the European Parliamentary Research Service, "*it is important to have a stated purpose of a given deployment of algorithmic decision-making. The articulated purpose serves both as a benchmark of the*

[63] Article 5.1, f) of the GDPR.
[64] Article 28 of the GDPR.
[65] Article 4.7 of the GDPR.
[66] Article 28 of the GDPR.
[67] See article 5.1 f) and article 82 of the GDPR.
[68] Article 28.3 (h) of the GDPR.
[69] See article 5 of the GDPR.
[70] It includes the right of access, the right to rectification, the right to erasure ("right to be forgotten"), right to restriction of processing, right to data portability, the right to object and the right not to be subject to an automated decision-making, including profiling.
[71] Article 5.1 a) et b) of the GDPR; C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, pp. 94 and following.

algorithmic performance and a legitimizing force, since the relationship between means and ends becomes verifiable"[72].

In addition, the GDPR prohibits further processing in a manner that is incompatible with the original purpose[73]. To determine the legality of a further processing, the GDPR establishes a list of factors that should be considered[74]. Here are some examples of pertinent factors (this list is not exhaustive[75]):

- the existence of a link between the original purpose and the new one,
- the context in which the data was collected,
- the nature of the data
- the possible consequences of the processing,
- and the existence of safeguards such as encryption or pseudonymisation[76].

Once again, the principle of purpose determined upstream is a challenge in terms of artificial intelligence. It is important that the controller maintains control over the algorithms and the use that will be made of personal data.

Misuse of data could also occur through a failure to comply with the requirements of integrity and confidentiality of information (e.g. unauthorised modification, manipulation, data theft, etc.) called, in the GDPR, data breach. It is therefore essential that the data controller and the data processor ensure:

- Preventing unauthorised access to or use of personal data
- Authenticity of the personal data and has not been maliciously or accidentally altered during processing, storage or transmission
- Traceability of who had access to what personal data (log files).

4.2.5.6 Data quality

Article 5

Personal data shall be:

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('**accuracy**');

The issue of the quality and updating of personal data is of fundamental importance in artificial intelligence systems. In particular, in a context of automated decision-making through the use of artificial intelligence algorithms, the consequences can be seriously negative for the data subjects if the controller does not ensure the quality of the information regularly.

[72] EPRS, "A governance framework for algorithmic accountability and transparency", p. 10.
[73]. Article 6.4 of the GDPR. Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered incompatible with the initial purposes.
[74] Article 6.4 of the GDPR.
[75]. Recital 50 of the GDPR.
[76] There is controversy about the applicability of these criteria for particular categories of personal data (sensitive data). See the Opinion of the EDPB 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection Regulation (GDPR) (art. 70.1.b)). The position of CRIDS is to be careful because, in principle, these personal data cannot be processed. On this aspect, see J.-M. VAN GYSEGHEM, « Les catégories particulières de données à caractère personnel », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier.

4.2.5.7 The right not to be subject to an automated decision

As seen above, the GDPR increases the duty of transparency from the data controller, in particular by reinforcing the duty of information to the data subjects[77]. In addition, the Regulation reinforces the right not to be subject to an automated individual decision-making[78]. The data controller has to implement the appropriate measures for safeguarding the rights and freedoms of the data subject to an automated decision. Furthermore, the data subject has the right to receive the information necessary to contest the decision[79].

Article 22.3 of the GDPR provides guarantees for data subjects who would be the subject of an automated individual decision. Thus, the data controller must put in place technical and organisational measures to safeguard their rights and freedoms and their legitimate interests. The European legislator enshrines as minimum guarantees the right to obtain human intervention on the part of the data controller, thus preventing a total submission of the human being to software and algorithms, the right to express his/her point of view and the right to contest the decision. In addition, Articles 13 and 14 on the right to information provide that in addition to the information referred to in paragraph 1, the controller shall provide the data subject, at the time the personal data are obtained[80] or within one month at the latest if the personal data have not been obtained from the data subject[81], with such additional information as would be necessary to ensure fair and transparent processing. These include information about the existence of automated decision-making, including profiling. In addition, meaningful information about the logic involved, the significance and the envisaged consequences of such processing for the data subject must be communicated by the controller to the data subject[82].

This is particularly relevant in AI context where one of the major challenges is to avoid “black-boxes phenomenon”.

4.2.5.7.1 Automated decision-making

Article 22 of the GDPR states that the data subject has the right not to be subject to a decision based solely on automated processing, which produces legal effects concerning him or her or similarly significantly affects them. The principle in the Regulation is the prohibition on fully automated decision-making.

We remind that all the fundamental principles of personal data protection apply in the context of automated decision-making such as the principle of a lawful, fair and transparent processing, data minimisation, accuracy and storage limitation.
Furthermore, the data subject has the right to be informed, right of access, right to rectification, right to erasure, right to restriction of the processing and the right to object[83].

The Article 29 Working Party defines the concept of automated decision-making as "*the ability to make decisions by technological means without human involvement. Automated decisions can be made with or without profiling*"[84].

The Article 29 Working Party is sensitive to a decision that has "*the potential to significantly influence the circumstances, behaviour or choices if the individuals concerned*"[85]. It implies a case by case analysis. The following criterion are determinatives:

[77] See articles 11-14 of the GDPR.
[78] Article 22 of the GDPR.
[79] Recital 71, article 13.2 f) and article 14.2 g) of the GDPR.
[80] Article 13.1 of the GDPR.
[81] Article 14.1 of the GDPR.
[82] Article 13.2 f) and article 14.2 g) of the GDPR.
[83] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017.
[84] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017, p. 8.

- The type and volume of personal data processed
- The purpose of the processing
- Whether the processing:
    - has, or is likely to have, an actual effect in terms of limiting rights or denying an opportunity;
    - causes, or is likely to cause, damage, loss or distress to individuals;
    - affects, or is likely to affect individuals' health, well-being or peace of mind;
    - affects, or is likely to affect, individuals' financial or economic status or circumstances;
    - leaves individuals open to discrimination or unfair treatment;
    - involves the analysis of the special categories of personal or other intrusive data
    - causes, or is likely to cause individuals to change their behaviour in a significant way;
    - has unlikely, unanticipated or unwanted consequences for individuals;
    - creates embarrassment or other negative outcomes, including reputational damage;
    - involves the processing of a wide range of personal data[86].

4.2.5.7.2 Exception to the interdiction: Consent of the data subject

The consent is an exception that permits the use of automated decision-making. There are several safeguards. Firstly, the data controller has to inform properly the data subject about the existence of an automated decision-making in order to ensure a real consent from the data subject[87]. They have the right to be informed about the logic involved, the explanation of the mechanism and how the system works. It is not mandatory to enter into details on the functioning of algorithms[88]. This information must be provided by the data controller when the data are collected. Consequently, and at this stage, it does not concern information about how the decision was reached by the system in a concrete situation[89].

Secondly, the data subject has the right to obtain a human intervention, to express their opinion and to contest the decision[90]. Recital 71 of the GDPR specifies that they need to have the possibility to obtain an explanation of the decision reached. It goes beyond a simple right of information as it requires to give an *ex post* information on the concrete decision. It implies for the data controller to understand and to be able to explain in a comprehensive way the algorithmic functioning of the system[91].

The possibility to obtain an explanation of the decision reached is not mandatory as it is in a recital and not in the article itself, but permits a real and meaningful possibility to express an opinion and

[85] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017, p. 11.
[86] Article 29 Working Party, Guidelines for identifying a controller or processor's lead supervisory authority, 05.04.2017, pp. 3-4.
[87] See articles 13 and 14 GDPR
[88] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017, p. 14.
[89] S. WACHTER, B. MITTELSTADT and L. FLORIDI, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). *International Data Privacy Law*, 2017. Available at SSRN: https://ssrn.com/abstract=2903469, pp. 82-83.
[90] Article 22.3 GDPR.
[91] S. WACHTER, B. MITTELSTADT and L. FLORIDI, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). *International Data Privacy Law*, 2017. Available at SSRN: https://ssrn.com/abstract=2903469, p. 92-93.

to decide to contest the decision or not. The Article 29 Working Party insists on the role of the controller in the transparency of the processing[92].

Thirdly, the data controller must ensure that someone who has the authority and ability to remove the decision will review the automated decision[93].

Fourthly, the Article 29 Working Party advises the data controller to carry out frequent assessments on the personal data collected and used. The curacy of the personal data and the quality of the tools used for the processing are core elements. The data controller, with the help of the data processor, should not only check the quality and prevent errors or inaccuracies at the time of the collection and the technical implementation of his system but in a continuous way as long as the data is not deleted or anonymized[94].

It is important to keep in mind that, in addition to the consent, the GDPR also provides two other exceptions to the prohibition of an automated individual decision-making. The interdiction does not apply if the decision is necessary for entering into, or performance of, a contract between the data subject and the data controller and when the decision is authorised by Union or Member State law to which the data controller is subject.

4.2.5.7.3 Guidelines[95]

They are several good practices in the context of automated individual decision-making, such as:

- To give clear, accessible and meaningful information about the personal data used, the processing, the existence of automated decision, the logic involved and how it is used for a decision concerning the data subject
- Visualization and interactive techniques to aid algorithmic transparency
- To inform the data subject about his or her rights
- To give all the necessary information for the data subject exercise their right of access, right to rectification, right to object
- To make effective the exercise of his or her right by the data subject by the use of technical tools
- Regular quality assurance checks of the system and algorithms
- Ensure the principle data minimization and the establishment of a clear period of retention of the personal data and the possible profile
- To always choose anonymization or pseudonymisation when possible
- To put in place a real and concrete procedure to permit the data subject to express his or her point view and to contest the decision and to have a human intervention.

[92] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017.
[93] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017, p. 10.
[94] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017, pp. 16-17.
[95] For more information, see the Guidelines of Article 29 Working Party.

4.2.5.8 To sum up

Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

GDPR inputs:

- Information to the data subject
- Confidentiality
- Minimisation
- Control
- Quality of the data
- Purpose
- Human intervention

4.2.6 *Conclusion*

Since 2017, artificial intelligence receives an important place in the interests and concerns of the European Union institutions. Recently, a study by the European Parliamentary Research Service came out along with the guidelines of the High-level Expert Group on Artificial Intelligence. In essence, from reading these documents, it seems that the notion of fairness is closely linked to the respect of the European values such as equality, equity, freedom of choice, justice, truth, trust, autonomy and privacy.

In order to integrate these fundamental European values into the development of artificial intelligence, all stakeholders must ensure the legality, ethics and robustness of their algorithms. An assessment list has also been set up by the High-level Expert Group on Artificial Intelligence.

Another tool available to developers and computer scientists is the General Data Protection Regulation (GDPR).

The GDPR adopts a certain approach to the notion of fairness. Indeed, when personal data are collected and processed, the GDPR states that the personal data have to be processed lawfully, fairly and in a transparent manner.

The regulation does not define the notion of fairness. The purpose of this section has therefore been to analyse and determine the essential elements for ensuring the fair processing of personal data in a context of artificial intelligence.

In essence, the notion of fairness seems closely linked to the requirement of purpose, transparency; quality of the personal data and security.

Fairness in the use of personal data is also reinforced by two mechanisms. First, the requirement of data protection by design and by default requires the controller and the processor to configure, from the outset, their artificial intelligence systems so that they comply with the requirements of the regulations on privacy and the protection of personal data. Secondly, the GDPR enshrines the principle of accountability, thus requiring the controller to comply with the regulation but also to be able to demonstrate in practice, in the event of an audit by a supervisory authority, that the requirements of the GDPR are met.

Moreover, it is interesting to note that BEUC's recent work identifies as AI rights for consumers requirements that the GDPR imposes on the controller and, to a certain extent, on the processor to be respected, even if the consumer's rights promoted by the European Consumer Organisation do not only apply when personal data are processed but to any artificial intelligence system.

For example, the right to transparency, explanation and objection is provided by article 22 of the GDPR (even if, according to BEUC, this right should not be limited to fully automated decisions but also in case of preparatory or supportive tool for a person).

The objective of this section is therefore to complement the approaches of the previous chapters. After the description and analysis of the defensive and reactive mechanisms to reinforce the security of artificial intelligence and the problem of explanation, the development of artificial intelligence is enriched by the adoption of a fairness approach. Indeed, the notion of fairness also encompasses the first chapter (reactive and defensive mechanisms refers to the criteria of technical robustness and safety) and the second (explainability which refers to the transparency criteria). The objective of this section is to complement this approach with a more global vision of the notion of fairness and to deepen the approach provided by the GDPR (privacy and data governance is a key standard for the European Commission in the fairness approach).

Chapter 5 Information about the AI Program contests

Modern machine learning algorithms are extremely susceptible to small and almost imperceptible perturbations of their inputs (so-called adversarial examples). An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it. In many cases, these modifications can be so subtle that a human observer does not even notice the modification at all, yet the classifier still makes a mistake.

Adversarial examples pose security concerns because they could be used to perform an attack on machine learning systems, even if the adversary has no access to the underlying model.

To accelerate research on adversarial examples, we need to organize a competition on Adversarial Attacks and Defence. This competition is going to design to facilitate measurable progress towards robust machine learning models and more generally applicable adversarial attacks.

The competition on Adversarial Attacks and Defences could consist of three sub-competitions:

- **Non-targeted Adversarial Attack** (Generate untargeted adversarial examples)
  - The goal of the non-targeted attack is to slightly modify the source dataset in a way that the dataset will be classified incorrectly.
  - In this task, participants are given a sample image and access to a model. The goal is to create an adversarial image that is as similar as possible to the sample image but is wrongly classified by the given model.
- **Targeted Adversarial Attack** (Generate targeted adversarial examples)
  - The goal of the targeted attack is to slightly modify the source dataset in a way that the dataset will be classified as specified target.
  - In this task, participants are given a sample image, a target class and access to a model. The goal is to create an adversarial image that is as similar as possible to the sample image but classified as the target class by the given model.
- **Defence Against Adversarial Attack**.
  - The goal of the defence is to build a machine learning classifier, which is robust to an adversarial example, i.e. can classify adversarial dataset correctly.
  - In this task, participants design robust models which are evaluated against various attacks.

5.1 Judging and Scoring

Each submission will earn a certain number of points.

- For the adversarial example classification task, all submissions will be shown the same set of input images. Each correctly classified image is worth one point. If the classifier fails to return a prediction for an image, it receives no points for that image.
- For the untargeted adversarial example generation task, each adversarial example generation algorithm will be asked to create adversarial examples. Each example will be presented to classifiers. The submission is awarded one point for each classifier that misclassifies the example. If the adversarial example generation algorithm fails to create an image, it receives no points for the requested image. If any classifier fails to return a prediction for any image, no points will be awarded to any attack algorithm for that classifier.

- For the targeted adversarial example generation task, each adversarial example will be asked to create adversarial examples and for each target label. Each example will be presented to the same set of classifiers. The submission is awarded one point for each classifier that classifies the image with the chosen target class. If the adversarial example generation algorithm fails to create an image, it receives no points for the requested image. If any classifier fails to return a prediction for any image, no points will be awarded to any attack algorithm for that classifier.

5.2 Ranking Participants

- Highest score first.
  - Participants will be ranked in order of the highest score first and the lowest score last. A separate ranking will be prepared for each of the three tasks.
- Ties ranked by submission time.
  - The one with the earliest submission time will be ranked first while the one with the latest submission time will be ranked last.

Chapter 6 Summary and Conclusion

This document gathered the preliminary descriptions of the formalisms, methods and tools designed and developed in the SPARTA WP7 SAFAIR program. These pertain to two contemporary challenges of Artificial Intelligence – the security of current AI solutions and its trustworthiness.

The document opens with a variety of defensive mechanisms in multiple domains, like medical image processing and cybersecurity. The presented mechanisms make AI more robust against the novel threat of adversarial attacks, seeking at the same time to preserve the performance of AI, making it possible to fully utilise its advantages, safely.

This is followed by mechanisms ensuring the trustworthiness of AI by providing explainability of the decisions of the AI solutions. These shed a light the black-box nature of AI, documenting its functionality and supplying better understanding of its decision-making process to the human operators.

In the fourth chapter, the technical and legislative problems fairness are addressed, supplying the mechanisms to reduce bias and finding ways for AI to correctly comply with applicable laws.

The document closes with a preliminary description of the SAFAIR AI contest to be conducted in the coming months.

List of SAFAIR Publications

- Pawlicki Marek., Choraś Michal, Kozik Rafał, **Defending network intrusion detection systems against adversarial evasion attacks** , Future Generation Computer Systems, Volume 110, September 2020, Pages 148-154, 2020
- Michał Choraś, Marek Pawlicki, Rafał Kozik: **The Feasibility of Deep Learning Use for Adversarial Model Extraction in the Cybersecurity Domain**. IDEAL (2) 2019: 353-360
- Mateusz Szczepański, Michał Choraś, Marek Pawlicki, Rafał Kozik: **Achieving Explainability of Intrusion Detection System by Hybrid Oracle- Explainer Approach.** IJCNN Sessions – 2020 IEEE WCCI
- Michal Choraś, Marek Pawlicki, Damian Puchalski, Rafal Kozik: **Machine Learning - The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness?** ICCS (4) 2020: 615-628
- Marek Pawlicki, Michal Choras, Rafal Kozik, Witold Holubowicz: **On the Impact of Network Data Balancing in Cybersecurity Applications**. ICCS (4) 2020: 196-210
- Michał Choraś, Marek Pawlicki, Rafał Kozik: **SAFAIR: Secure and Fair AI Systems for Citizens**(PP-RAI Wrocław 2019)
- Xabier Echeberria-Barrio, Amaia Gil-Lerchundi, Ines Goicoechea-Telleria, and Raul Orduna-Urrutia: **Deep Learning Defenses Against Adversarial Examples for Dynamic Risk Assessment**, Accepted to CISIS2020
- Simon Grah, Vincent Thouvenot: **A Projected SGD algorithm for estimating Shapley Value applied in attribute importance**, Accepted to CD-MAKE 2020

References

[1] Donald MacKenzie. Mechanizing Proof. MIT Press, 2001
[2] Richard A. De Millo, Richard J. Lipton, and Alan J. Perlis. Social processes and proofs of theorems and programs. Communications of the Association of Computing Machinery, May 1979
[3] K. Dvijotham, R. Stanforth, S. Gowal, T. Mann, and P. Kohli, ‘‘A Dual Approach to Scalable Verification of Deep Networks,’’ in Conference on Uncertainty in Artificial Intelligence (UAI), 2018
[4] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, ‘‘Safety Verification of Deep Neural Networks,’’ in International Conference on Computer Aided Verification, 2017
[5] W. Xiang, H. Tran, and T. T. Johnson, ‘‘Output Reachable Set Estimation and Verification for Multilayer Neural Networks,’’ IEEE Transactions on Neural Networks and Learning Systems, 2018),
[6] Algorithms for Verifying Deep Neural Networks, Changliu Liu, Tomer Arnon, Christopher Lazarus, Clark Barrett, Mykel J. Kochenderfer, VNN2019
[7] PaRoT: A Practical Framework for Robust Deep Neural Network Training, Edward Ayers and Francisco Eiras and Majd Hawasly and Iain Whiteside, NASA Formal Methods Symposium, NFM 2020
[8] DL2: Training and Querying Neural Networks with Logic, Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, Martin Vechev in International Conference on Machine Learning 2019
[9] G. Katz et al. “The marabou framework for verification and analysis of deep neural networks”. In: International Conference on Computer Aided Verification. Springer. 2019
[10] Gopinath, D., Converse, H., Pasareanu, C., & Taly, A. (2019, November). Property inference for deep neural networks. In 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE) (pp. 797-809). IEEE
[11] Girard-Satabin, J.; Charpiat, G.; Chihani, Z. & Schoenauer, M.CAMUS: A Framework to Build Formal Specifications for Deep Perception Systems Using Simulators ECAI, 2020
[12] Singh, G., Ganvir, R., Püschel, M., & Vechev, M. (2019). Beyond the single neuron convex barrier for neural network certification. In Advances in Neural Information Processing Systems (pp. 15098-15109).
[13] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017, July). Reluplex: An efficient SMT solver for verifying deep neural networks. In International Conference on Computer Aided Verification (pp. 97-117). Springer, Cham.
[14] Sharafaldin, A. H. Lashkari, A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization., in: ICISSP, 2018, pp. 108–116.
[15] M. Pawlicki, R. Kozik, M. Choras, Artificial neural network hyperparameter optimisation for network intrusion detection, in: Intelligent Computing Theories and Application - 15th International Conference, ICIC 2019, Nanchang, China, August 3-6, 2019, Proceedings, Part I, 2019, pp. 749–760. doi:10.1007/978-3-030-26763-6_72. URL https://doi.org/10.1007/978-3-030-26763-6_72
[16] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 39–57
[17] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2014). arXiv:1412.6572.
[18] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, arXiv preprint arXiv:1607.02533
[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks (2017). arXiv:1706.06083.

[20] C. Szegedy *et al.*, “Intriguing properties of neural networks,” *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.*, pp. 1–10, 2014.
[21] F. Tramèr and D. Boneh, “Adversarial Training and Robustness for Multiple Perturbations,” 2019.
[22] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, “Enhancing robustness of machine learning systems via data transformations,” *2018 52nd Annu. Conf. Inf. Sci. Syst. CISS 2018*, no. 1, pp. 1–5, 2018.
[23] R. Sahay, R. Mahfuz, and A. El Gamal, “Combatting Adversarial Attacks through Denoising and Dimensionality Reduction: A Cascaded Autoencoder Approach,” *2019 53rd Annu. Conf. Inf. Sci. Syst. CISS 2019*, 2019.
[24] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, “Workshop track -ICLR 2017 TACTICS OF ADVERSARIAL ATTACK ON DEEP REIN- FORCEMENT LEARNING AGENTS,” pp. 3756–3762, 2014.
[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE Trans. IMAGE Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[26] L. Zhang, Lei Zhang, X. Mou, and David Zhang, “FSIM: A Feature Similarity Index for Image Quality Assessment Lin,” *IEEE Trans. IMAGE Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.
[27] L. Akoglu, H. Tong, and D. Koutra. “Graph based anomaly detection and description: a survey.” In: Data Mining and Knowledge Discovery 29.3 (July 2014), pp. 626–688. issn: 1573-756X. doi:10.1007/s10618-014-0365-y.
[28] D. Chakrabarti. “AutoPart: Parameter-Free Graph Partitioning and Outlier Detection.” In: Lecture Notes in Computer Science Knowledge Discovery in Databases: PKDD 2004 (2004), pp. 112–124. doi: 10.1007/978-3-540-30116-5_13.
[29] R. Hu, C. C. Aggarwal, S. Ma, and J. Huai. “An embedding approach to anomaly detection.” In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE) (2016). doi: 10.1109/icde.2016.7498256.
[30] R. J. Bolton and D. J. Hand. “Unsupervised Profiling Methods for Fraud Detection.” In: Proc. Credit Scoring and Credit Control VII. 2001, pp. 5–7.
[31] L. Invernizzi and P. M. Comparetti. “EvilSeed: A Guided Approach to Finding Malicious Web Pages.” In: 2012 IEEE Symposium on Security and Privacy (2012). doi: 10.1109/sp.2012.33.
[32] M. Ott, C. Cardie, and J. Hancock. “Estimating the prevalence of deception in online review communities.” In: Proceedings of the 21st international conference on World Wide Web - WWW 12 (2012). doi: 10.1145/2187836.2187864.
[33] J. Yang and J. Leskovec. “Defining and Evaluating Network Communities Based on Ground-truth.” In: Knowl. Inf. Syst. 42.1 (Jan. 2015), pp. 181–213. issn: 0219-1377. doi: 10.1007/s10115-013-0693-z.52
[34] N. Papernot, P. D. McDaniel, and I. J. Goodfellow. “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples.” In: CoRR abs/1605.07277 (2016).
[35] R. Kozik, C. Micha l, J. Keller, Balanced efficient lifelong learning (B-ELLA) for cyber-attack detection, J. UCS 25 (1) (2019) 2-5.URL http://www.jucs.org/jucs_25_1/balanced_efficient_lifelong_learning
[36] M. Chora´s, M. Pawlicki, R. Kozik, The feasibility of deep learning usefor adversarial model extraction in the cybersecurity domain, in: H. Yin,D. Camacho, P. Tino, A. J. Tall´on-Ballesteros, R. Menezes, R. All-mendinger (Eds.), Intelligent Data Engineering and Automated Learning –IDEAL 2019, Springer International Publishing, Cham, 2019, pp. 353–360.
[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence, pages 1527–1535, 2018.
[38] https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

[39] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey." 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018.
[40] https://towardsdatascience.com/we-are-ready-to-ml-explainability-2e7960cb950d
[41] Bhatt, Umang, et al. "Explainable Machine Learning in Deployment." arXiv preprint arXiv:1909.06342 (2019).
[42] https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html
[43] Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics." Electronics 8.8 (2019): 832.
[44] Etmann, Christian, et al. "On the Connection Between Adversarial Robustness and Saliency Map Interpretability." arXiv preprint arXiv:1905.04172 (2019).
[45] Molnar, Christoph. "A guide for making black box models explainable." URL: https://christophm. github. io/interpretable-ml-book
[46] Robnik-Šikonja, Marko, and Marko Bohanec. "Perturbation-Based Explanations of Prediction Models." Human and Machine Learning. Springer, Cham, 2018. 159-175.
[47] Silva, Wilson, et al. "Towards Complementary Explanations Using Deep Neural Networks." Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer, Cham, 2018. 133-140.
[48] Honegger, M. Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions. arXiv 2018, arXiv:1808.05054.
[49] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoff-man, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovi´c, S. Mourad, P. Pedemonte,R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varsh-ney, D. Wei, and Y. Zhang, “One explanation does not fit all: A toolkit and taxon-omy of ai explainability techniques,” 2019.
[50] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, and Y. Li, “An explainableai decision-support-system to automate loan underwriting,”Expert Systems withApplications, vol. 144, p. 113100, 2020.
[51] “Flowcast official web page and associated resources.” https://flowcast.ai, 2020.Accessed: 2020-05-08.
[52] M. Wang, K. Zheng, Y. Yang, and X. Wang, “An explainable machine learningframework for intrusion detection systems,”IEEE Access, vol. 8, pp. 73127–73141,2020
[53] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 51(5):93:1– 93:42, August 2018.
[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence, pages 1527–1535, 2018.
[55] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, volume 30, 2017.
[56] Gabriel G. Erion Scott M. Lundberg and Su-In Lee. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1606.03490, 2019.
[57] Sandra Wachter, Brent DM Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology, 31(2):841–887, 2018.
[58] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability inmachine learning. arXiv preprint, 2017.
[59] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.
[60] Dhaliwal and Shintre, Gradient Similarity: An Explainable Approach to Detect Adversarial Attacks against Deep Learning, 2018

[61] L. Vigano, D. Magazzeni, ”Explainable security,” 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence, Stockholm, Sweden, pp. 158-164, July 13-19 2018.
[62] A. Weller, ”Transparency: motivations and challenges,” in Samek W., Montavon G., Vedaldi A., Hansen L., Muller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham, pp. 23-40, September 2019.
[63] L. K. Hansen, L. Rieger, ”Interpretability in intelligent systems – a new concept?,” in Samek W., Montavon G., Vedaldi A., Hansen L., Muller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham, pp. 41-50, September 2019.
[64] A. Richardson, A. Rosenfeld, ”A survey of interpretability and explainability in human-agent systems,” 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence, Stockholm, Sweden, pp. 137-144, July 13-19 2018.
[65] M. Ribeiro, S. Singh, C. Guestrin, ”Why should I trust you?: explaining the predictions of any classifier,” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, California, pp. 97-101, June 2016.
[66] A. Blanco-Justicia, J. Domingo-Ferrer, ”Machine learning explainability through comprehensible decision trees,” in Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2019. Lecture Notes in Computer Science, vol 11713. Springer, Cham, August 2019.
[67] J. Domingo-Ferrer, V. Torra, ”Ordinal, continuous and heterogeneous k-anonymity through microaggregation,” Data Mining and Knowledge Discovery, 11(2), pp.195-212, August 2005.
[68] T. Parr, P. Grover, ”Explained.ai”, https://explained.ai/decision-treeviz/index.html. Last accessed 8 Jan 2020.
[69] Ziyuan Zhong, A Tutorial on Fairness in Machine Learning, https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
[70] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems. 2016.
[71] Zafar, Muhammad Bilal, et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.
[72] Del Barrio, E., Gamboa, F., Gordaliza, P., & Loubes, J. M. (2018). Obtaining fairness using optimal transport theory. arXiv preprint arXiv:1806.03195.
[73] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).
[74] Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. arXiv preprint arXiv:1511.00830.
[75] Lum, K., & Johndrow, J. (2016). A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077.
[76] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 3992-4001).
[77] Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. arXiv preprint arXiv:1702.06081.
[78] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259.
[79] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.
[80] Louppe, Kagan, and Cranmer, Learning to Pivot with Adversarial Networks, 2017
[81] Raff, Sylvester and Mills, Fair Forests: Regularized Tree Induction to Minimize Model Bias, 2017