XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Explainability Protocol and Methods (v1)

Deliverable number: D4.1

Version: 1.0

Acronym of the project: XAIface
Title of the project: Measuring and Improving Explainability for AI-based Face Recognition
Grant: CHIST-ERA-19-XAI-011
Web site of the project: https://xaiface.eurecom.fr/

Short abstract

Deliverable D4.1 “Explainability protocol and methods” reports the consortium work towards the development of explainable techniques for face recognition during the project. This deliverable will be published in two versions, one in month 12 and one in month 24. The second version, in addition to collecting the description of the methods developed, will also include their first implementation.

Table of content

Abbreviations

AI	Artificial Intelligence
CAM	Class Activation Maps
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Networks
DET	Detection Error Tradeoff
DL	Deep Learning
EBM	Explainable Boosting Machine
FAD	Feature Activation Diversity
FMR	False Match Rate
FNMR	False non-match rate
FR	Face Recognition
GDPR	General Data Protection Regulation
JPEG	Joint Photographic Experts Group
LIME	Local Interpretable Model-Agnostic Explanations
LMF	Large Magnitude Filtering
LRP	Layer-wise Relevance Propagation
RISE	Randomised Input Sampling for Explanation
RNN	Recurrent Neural Network
SAD	Spatial Activation Diversity
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

Definitions

ArcFace is a CNN based model for face recognition which learns discriminative features of faces and produces embeddings for input face images. To enhance the discriminative power of softmax loss, a novel supervisor signal called additive angular margin (ArcFace) is used as an additive term in softmax loss.

Automation bias: Over-reliance on automated aids and decision support systems.The same concept can be translated to the fundamental way that AI and automation work, which is mainly based on learning from large sets of data. This type of computation assumes that things won't be radically different in the future. Another aspect that should be considered is the risk of using flawed training data then the learning will be flawed.

Bagging is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.

Demographic information: Socio-economic information from a subject such as population, race, income, education and employment.

Explainable AI (XAI) is artificial intelligence that is programmed to describe and understand its purpose, rationale and decision-making process in a way that can be understood by the average (human) person.

Greedy algorithms are any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage.

Kernels are filters used to extract the features from the images. More specifically, a kernel is a matrix that moves over the input data, performs the dot product with the sub-region of input data, and gets the output as the matrix of dot products.

MagFace is a category of losses that learns a universal feature embedding whose magnitude can measure the quality of the given face. Under the new loss, the magnitude of the feature embedding monotonically increases if the subject is more likely to be recognized. In addition, MagFace introduces an adaptive mechanism to learn well structured within-class feature distributions by pulling easy samples to class centres while pushing hard samples away. This prevents models from overfitting on noisy low-quality samples and improves face recognition in the wild.

Pooling layers are used to reduce the dimensions of the feature maps. It reduces the number of parameters to learn and the amount of computation performed in the network. The pooling layer summarises the features present in a region of the feature map generated by a convolution layer.

Post-hoc: analysis of the results of experimental data.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Soft biometrics are attributes that are not necessarily unique to an individual, but can be used alone or in conjunction with primary biometric traits for improving performance or explainability in a variety of applications.

1. Introduction

The objective of deliverable D4.1 “Explainability protocol and methods” is to plan, report and track the consortium work towards the development of explainable techniques for face recognition during the project.
This deliverable will be published in two versions, one in month 12 and one in month 24. The second version, in addition to collecting the description of the methods developed, will also include their first implementation.

In its initial version (v1), the document contains the following sections:

- Section 1: A section about general motivation for the document and introduction.
- Section 2: A section providing an overview of the state of the art of AI explainability methods, focused on face recognition domain. The reported methods should give a necessary background for the development of new methods. Thus, the methods described are already used - or seem at least promising (potentially applicable) - for Explainable Artificial Intelligence (XAI) face recognition.
- Section 3: This section describes the actual plan for the development of new methods and reports on the progress and performance assessment during the project in later stages of the deliverable.
- Section 4: In this section a preliminary assessment of potential legal and ethical issues for the face-recognition systems is provided and questions regarding the right to obtain explanations are elaborated and discussed.

1.1. Motivation

Face recognition has become a key technology in our society, frequently used in multiple applications, while creating an impact in terms of privacy. As face recognition solutions based on artificial intelligence (AI) are becoming more and more popular, it is critical to fully understand and explain how these technologies work in order to make them more effective and accepted by society. Thus in XAIface, we focus on the analysis of the influencing factors relevant for the final decision of an AI-based face recognition system as an essential step to understand and improve the underlying processes involved.

Most state-of-the-art work regarding XAI is nowadays working either on the development of transparent machine learning models, or on the application of post-hoc explainability models. While transparency is most often an inherent property of simpler, classical machine-learning models, state-of-the-art (more complex) face recognition methods require “post-hoc” or “post-modelling” explainability techniques like simplification, feature relevance, or class activation maps. Of major interest are also model-agnostic techniques, which can be plugged into any model to extract information from the (black-boxed) prediction process.

Another related aspect is attention, as this will also facilitate the understanding which features of a face contributed to the decision made. Attention is a mechanism allowing the model to focus on certain parts of the input sequence when predicting a certain part of the output sequence, enabling easier learning and of higher quality. Attention can be introduced into different network architectures, such as multi-task networks or recurrent neural networks (RNNs). In face recognition, it has for example been proposed for weighting features in face verification and identification. Also, attention has been used to select face local patches without resorting to face landmarks. A related application of attention is modifying face attributes in generative approaches, which could be a component in adversarial learning. Some of these methods are also reviewed in Section 2.

Another main field this project will contribute to knowledge beyond state of the art in XAI will be the progress in optimising the tradeoff between interpretability and performance. This will be addressed by the development of adjustable explainability components of face recognition pipelines, which may affect some or all components of the pipeline. The plans for this work as well as results obtained so far (in later versions of the report) are outlined and reported in Section 3.

2. Related Literature Works for AI Explainability with Potential Application for Face Recognition

In this section, a brief overview of state-of-the-art methods related to the planned XAIface contributions to AI face recognition explainability is provided. This review will serve as background to understand the developed techniques in XAIface as well as highlighting the contributions of XAIface to the state of the art.

2.1. General Observations

Model explainability, one of the most important problems in machine learning today, refers to the concept of being able to understand a deep learning network. It aims for a better comprehension of why machine learning models make certain decisions and which factors are most relevant for it.

Being able to interpret a machine learning model leads to several benefits because understanding the decision process of a network increases the user's trust in the prediction, helps to debug a model during its development phase and can be relevant in determining whether or not a model is suitable for a real world application. Model interpretability, when successfully performed, can also help and guide in assessing the presence of bias.

In the existing literature, many classifications of explainability models have been used. In this document we distinguish according to the following (most common) XAI grouping:

- **Local vs. global methods:** Local methods investigate a specific set of samples and try to explain the model behaviour on those specific samples. In opposition, global methods give an overall explanation of model behaviour by studying how features collectively affect the final prediction.
- **Model specific vs. model agnostic:** Model specific techniques work under certain, clearly defined model structures, while model agnostic tools work in a more general manner by analysing directly the input-output pair without specific knowledge on the model.
- **White-box vs black-box models:** A white-box model is a technique designed to be explainable per se, thus it does not require any other additional explainability technique on top or besides of it to interpret the decision process. A black-box model is not explainable by itself. Instead of that it requires the adaptation of other techniques in its pipeline in order to make it explainable. The main advantage of black-box models is the fact that one can take an already trained model and train an explanation model on top of it. As a general rule, the explainer is much easier to understand.
- **Gradient-based backpropagation methods vs perturbation-based forward propagation methods:** Gradient-based methods investigate the signal flow through the classifier, investigating how changes in the weights and parameters in the model will influence the output. On the other hand, perturbation-based models change only the input of the model and investigate how the output changes.

2.2. Shapley-values (SHAP)

SHAP (Lundberg and Lee 2017), SHapley Additive exPlanations, is a framework for method interpretability that gives *data features* and *importance values* for specific predictions. SHAP values are based on Shapley-values, which refer in game theory to the average of all the marginal contributions to all possible coalitions. This average is used as a baseline and SHAP shows the impact of each feature by interpreting and measuring the relative impact of a certain value compared to that baseline. In this way, SHAP values allow us to determine any prediction as a sum of the effects or amount of influence of each feature value to the final decision.

Interpreting a model using SHAP has the advantage of providing both local and global model interpretability simultaneously. The contribution of each predictor to the final prediction, positively or negatively, can be measured by looking at the collective SHAP values thus ensuring global interpretability. Furthermore, each individual sample obtains its own set of SHAP values allowing us to assess and compare the relevance and impact on the final prediction of all individual factors. It is also important to mention that the calculation of the SHAP values is computationally expensive. In fact it is exponential in the number of features and this can be especially critical for highly dimensional (face-) features.

Summing up SHAP is a model agnostic technique thus it focuses on analysing the relation between the inputs and the predictions for black-box approaches in a perturbation-based forward propagation manner. Its adaptability to any type of deep learning structure makes it especially interesting for the project since SHAP can help to understand the most significant regions for face recognition and to check if those areas are consistent among different face recognition systems.

2.2.1. BreakDown

BreakDown (Staniak and Biecek 2018) is similar to SHAP. It is also based on the conditional response of a black-box model and it then attributes the response to the input features. The only difference is that breakDown deals with conditionings in a greedy way instead of averaging.
A GitHub implementation of the method can be found at:

https://github.com/MI2DataLab/pyBreakDown - python

https://github.com/pbiecek/breakDown - R

2.3. Local Interpretable Model-agnostic Explanations (LIME)

LIME (Ribeiro et al. 2016) (Local Interpretable Model-Agnostic Explanations) is one of the most important interpretability techniques developed by University Of Washington researchers to study the decision process inside an algorithm by capturing feature interactions. It is a model agnostic technique that analyses the relation between the input data and the prediction in a perturbation-based forward propagation manner, therefore it can be used on any model acting as a surrogate model. LIME modifies a test data instance by altering its input values (slightly) and observes the impact on the output. LIME tries to model the local neighbourhood of any prediction by focusing on a small decision surface around the input point. The basic assumption is that even very simple (linear) models are a good approximation of the blackbox model under investigation. Those more simple models are often linear/logistic regressions or decision trees. More specifically, LIME defines distance measures in order to compute the distance between the original sample and the altered one. Then, new predictions are computed by passing the altered samples through the black-box model. After that, LIME picks a number "*n*" of features by maximising the maximum likelihood, in order to select the "*n*" features that better describe the deep learning network. Those features are used to fit a simple model to the permuted data with the similarity scores as weights. Finally, by looking into this simple (surrogate) model, it is possible to get insights into the initial black-box model.

The main advantage of LIME is its ability to maintain local fidelity even when the number of dimensions is high. LIME ensures this property by finding a model that approximates the black-box model only locally. It is also able to handle irregular inputs. On the other hand, there is the danger that a bad or insufficient approximation might lead to misleading explanations. In addition, explanations are sometimes unstable and depend on the underlying process of input perturbation. LIME has been considered of interest for XAIface since it can provide quantitatively visual information on the most relevant image regions for face recognition tasks for any type of deep learning structure.

2.4. Saliency Maps for CNNs

Saliency maps are a special visualisation technique providing insight into the decision making process of a neural network. Several methods have been proposed in literature and in principle they identify and highlight (e.g. via heat-maps) the regions a network focuses on when providing decisions. Several variants of these techniques exist, for example, Class Activations Maps, which will be described in more detail in section 2.5. Here we focus on one of the earliest approaches regarding saliency map visualisation, namely the "deconvolutional network approach" from (Zeiler and Fergus 2014) which we will describe in more detail in the following.

As the name might presume, this approach uses an inverted convolutional network (=deconvolution network) which maps all activations of the individual feature maps (in all layers of the neural network, including the intermediate ones) to the corresponding input patterns in the input pixel space.

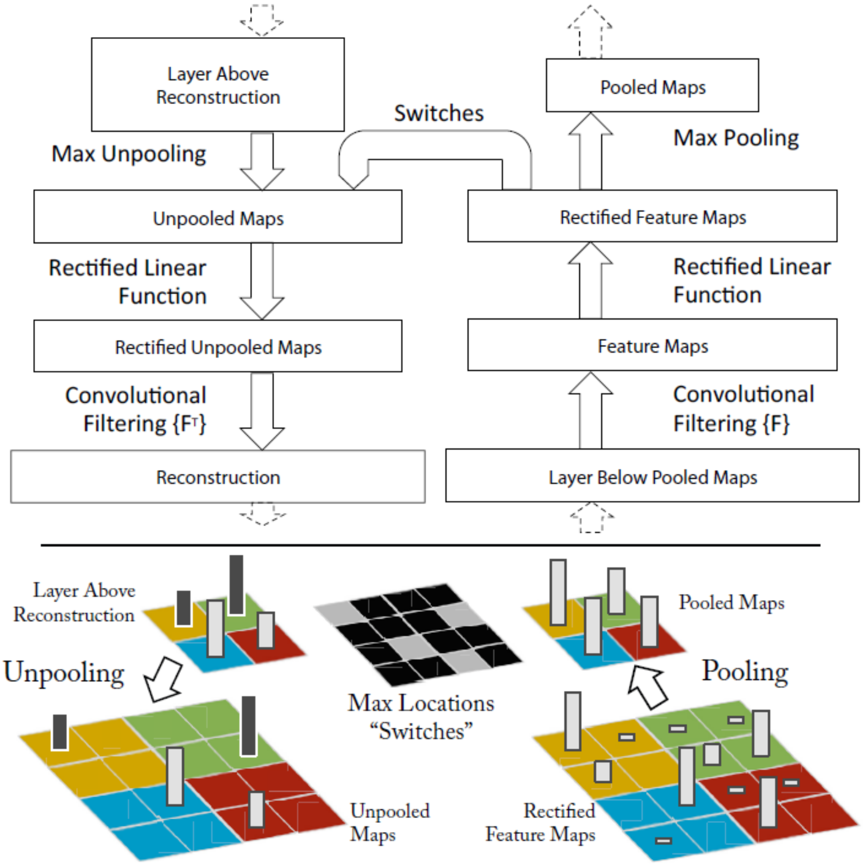

Figure 2.4a: Overview about the saliency map visualisation principle. The network on the left side will reconstruct an approximate version of the traditional convolutional neural network illustrated on the right side. The unpooling operation in the bottom of the figure uses so-called "switches" to record the locations of the local max in each pooling region (colored zones) during traditional operation.

The illustration in Figure 2.4a – taken from (Zeiler and Fergus 2014) – outlines the principle. The traditional image classification pipeline (ImageNet in this example) is shown on the right side and processed from the bottom to the top direction. The deconvolution network (deconv) is depicted on the left side and consists in principle by the same components, but in reverse order. As in the deconv-network the features are mapped to pixels, inverse operations of all convolutional network steps are required. While this is no problem for e.g. "filter" operations by using their transposed kernel versions, the situation is more difficult for non-invertible operations as e.g. max-pooling. In those cases only approximations of an inverse are feasible. In the case of the max-pooling operation for example, the intermediate "recording" (storage) of locations of the maxima within each pooling region in a set of so-called "switch variables" can be used to overcome that issue.

Regarding the categorization in the introduction, the "deconvolutional network approach" requires specific interventions (additional layers) within the structure of the neural network to provide a proper explanation. Thus the approach is definitely a model specific one. Anyway it allows the explanation of each sample per se thus providing a global explanation of the entire model behaviour for a sufficient number of testing samples analysed.

2.5. Class Activation Maps (CAM), Grad-CAM, and Grad-CAM++

Class Activation Maps (Zhou et al. 2016), or CAMs, is a deep learning interpretability method used for CNNs. It is used to indicate the discriminative regions of an image used by a CNN to identify the category of the image. In detail, the CAMs method performs the following steps. First of all, it modifies the network architecture by replacing fully-connected layers in the end by a Global Average Pooling layer and concatenates the averages of the activations of convolutional feature maps that are located before the final output layer and create a feature vector. The weighted sum of the vector is fed to the final softmax loss layer. Finally, the important image regions are identified by projecting back the weight of the output layer to the convolutional feature maps. Publicly available GitHub-implementations can be found at:

https://github.com/zhoubolei/CAM or https://github.com/frgfm/torch-cam

The main drawback of CAM is that it requires neural networks to have a specific architecture in the final layer. This is not the case for the Grad-CAM (Selvaraju et al. 2019) method, which is a generalisation of CAM that can produce visual explanations for any CNN architecture. As a gradient-based method, Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN, producing a coarse localization map of the important regions in the image. Grad-CAM++ (Chattopadhay et al. 2018) is an extension of the Grad-CAM method that provides better visual explanations of CNN model predictions.

Proper GitHub-implementations for Grad-CAM and Grad-CAM++:

https://github.com/jacobgil/pytorch-grad-cam

2.6. Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP) is a framework that decomposes the prediction of a deep neural network over a sample image down to relevance scores for image regions. It helps to interpret highly complex deep neural networks by back-propagating the predictions.
The LRP method supports various data types, e.g. images, text, etc, and various neural network architectures. Moreover, it has been successfully applied to explainable facial expression recognition (Arbabzadah et al. 2016), and document categorization (Arras et al. 2017).

The main idea of Layer-wise relevance propagation is to trace back the contributions of input nodes to the final network prediction. It performs backward propagation using a set of purposely-designed propagation rules from the output, identifying the most relevant neurons within the neural network until returning to the input. In detail, firstly the relevance score of the specified node in the final layer is set as output. The relevance score is then back propagated to the input layer. As a result, the prediction is decomposed into pixel-wise relevance indicating the contribution of a neuron to the final decision.

Figure 2.6a illustrates the general process and results of applying pixel-wise decomposition and LRP method in image classification tasks. A deep learning-based image classifier attempts to extract feature vector representations with deep convolutional neural networks and perform prediction with some classification layers. The LRP method decomposes the classification output into sums of feature and pixel relevance scores and consequently generates a heatmap visualising the contributions of every single pixel to the prediction.

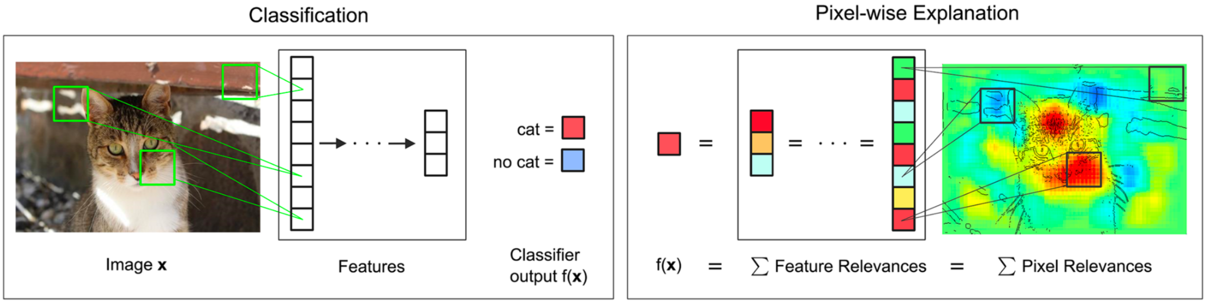

Figure 2.6a. Visualisation of the pixel-wise decomposition process in image classification tasks.

Note that a (non-official) GitHub implementation is available at:
https://github.com/sebastian-lapuschkin/lrp_toolbox .

2.7. Explainable Boosting Machines

Explainable Boosting Machine (EBM) is a glassbox model, designed to have comparable accuracy to state-of-the-art machine learning methods like Random Forest or XGBoost (Chen and Guestrin 2016) while being highly intelligible and explainable (Nori et al. 2019). They are very similar to generalised additive models (Hastie and Tibshirani 1987), but exhibit a number of important improvements besides their ability to uncover the feature contributions to the prediction of a model.
The basic idea behind of the approach is the selection of a simple additive model in the form $y = a_0 + a_1 \times f_1(x_1) + ... + a_k \times f_k(x_k)$ given y the output of the model and x_1, …, x the input features. a_n denotes the coefficients and f_n are the (even non-linear) functionals covering more complex correlations between input and output. EBMs learn the feature functions f by using modern machine learning techniques such as bagging and gradient boosting. The boosting procedure is carefully restricted to train on one feature at a time in round-robin fashion using a very low learning rate so that feature order does not matter. Moreover it round-robin cycles through the features to learn the best feature function and enables for easy estimation of the feature's contribution to the final prediction.
So in particular, the method uses very small trees (decision stumps) as sample and feature specific functionals and each tree is trained using only one single feature at a time! So the model created is a simple additive combination of r trees (=number of training samples) in the form of

$$y = a_0 + a_1 \times (T_1^{(1)}(x_1) + ... + T_r^{(1)}(x_1)) + ... + a_k \times (T_1^{(k)}(x_k) + ... + T_r^{(k)}(x_k))$$

and the sum of all trees for a particular feature represent the (highly non linear) functional f.
A potential disadvantage of the method is the fact that compared to other approaches the training time is larger, but during inference the approach requires only simple additions and

lookups. This enables EBMs to be one of the fastest models to execute at prediction time by coevally using low memory, as the trees can be represented as simple graphs and thus deleted after the training.
EBMs are especially important for the XAIface project, as they are an inherent part of the planned contribution described in section 3.1. They are a prototype for a whitebox-model as each step in the decision process can be explicitly tracked through the algorithm and explained. As they allow for the explanation of each single sample they can provide global explanations of the entire model once a sufficient number of testing samples are analysed.

2.8. Randomised Input Sampling for Explanation (RISE)

Unlike white-box approaches that estimate pixel importance using gradients, RISE (Petsiuk et al. 2018) works on black-box models. The RISE algorithm generates a saliency map for any black-box model, indicating how important each pixel of the image is with respect to the network's decision. This is similar to the saliency maps method mentioned before which require access to the internal structure of the model, such as the gradients of the output with respect to the input, intermediate feature maps, or the network’s weights. Therefore they are limited to certain types of network architectures or layers.

In contrast to the plain saliency maps for specific networks presented before, RISE provides a more general approach to produce saliency maps for an arbitrary network architecture. Figure 2.8a illustrates the overall workflow of the RISE algorithm. In detail, it firstly generates random binary masks following a uniform distribution. Then each input image is element-wise multiplied with the random masks and the resulting image is subsequently fed to the model for classification. The model produces probability-like scores for the masked images. In the end, a saliency map for the original image is created as a linear combination of the masks using the probability-like scores as coefficients.

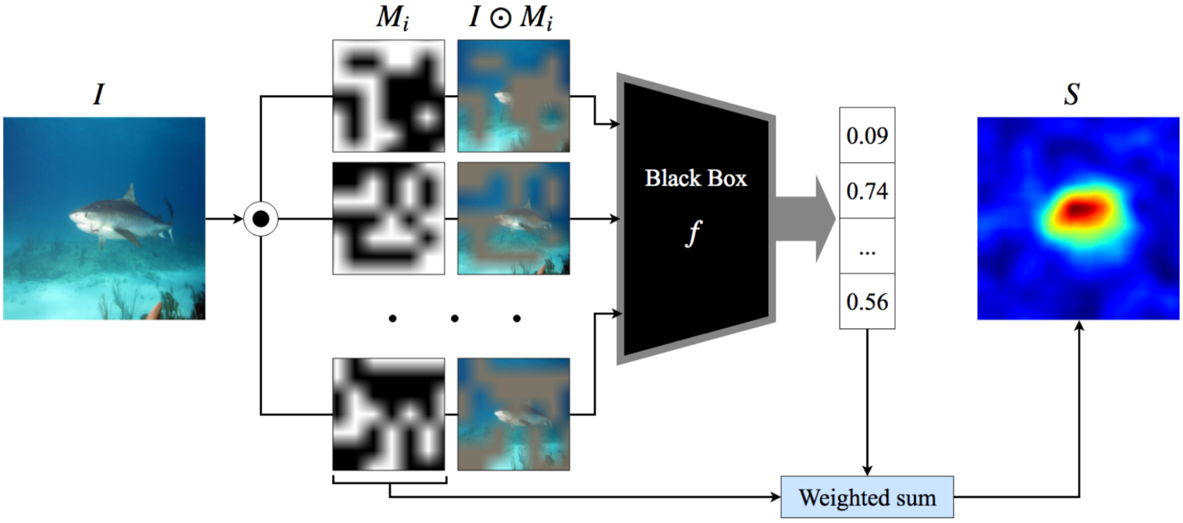

Figure 2.8a: Overview of the RISE algorithm. The input image I is multiplied with some randomly generated masks Mi in an element-wise manner. The black-box model takes masked images as input and produces scores of classes. The score vector serves as weights and linearly combines with the masks to create the saliency map.

RISE is an approach that explains black-box models via estimating the salient regions of input images for the model's predictions. It matches the goal of the XAIface project as it provides a way to interpret a face recognition model by highlighting the important regions of two matching or non-matching faces.

A proper GitHub-implementation can be found at:

https://github.com/eclique/RISE

2.9. Spatial and Feature Activation Diversity Losses for Structured Face Representations

The work presented in (Yin et al. 2019) proposes the usage of two loss functions - spatial activation diversity loss and feature activation diversity loss - to learn more structured face representations. Filters are learned end-to-end from data and constrained to be locally activated with the proposed spatial activation diversity loss. The feature activation diversity loss is introduced to better align filter responses across faces and encourage filters to capture more discriminative visual cues for face recognition, notably when dealing with occluded faces.

By leveraging the face structure, considering part-based representations, the proposed spatial and feature activation diversity losses strive for interpretable representations, which are discriminative and robust to occlusions. The authors claim that the final face representation does not compromise recognition accuracy.

The system architecture proposed by (Yin et al. 2019) for learning meaningful part-based face representations with a deep CNN, using carefully designed losses, is presented in Figure 2.9a. It consists of a Siamese network with two branches sharing weights to learn face representations from two faces: one with synthetic occlusion and one without.

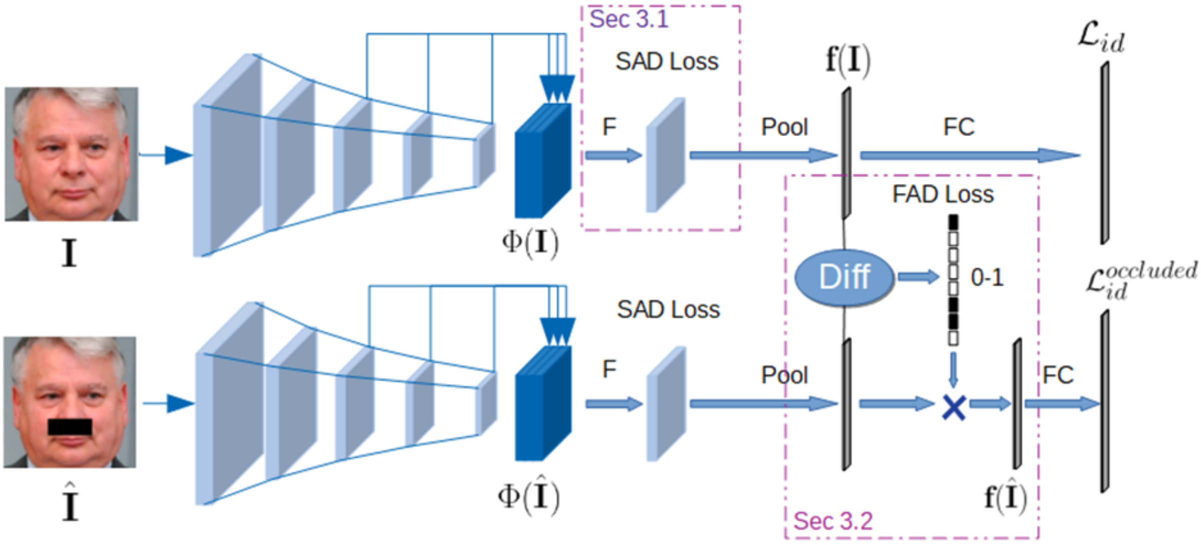

Figure 2.9a: Overall network architecture of the proposed framework. Spatial Activation Diversity (SAD) loss – promotes structured feature responses Feature Activation Diversity (FAD) loss – enforces features to be insensitive to local changes (occlusions).

The Spatial Activation Diversity (SAD) loss encourages the face representation to be structured, with consistent semantic meaning. Softmax loss helps to encode the identity information. Spatial Activation Diversity loss learns one set of filters for discriminating the K considered classes and employs the large magnitude filtering (LMF), also proposed in (Yin et al. 2019), which removes small magnitude values, favouring discriminative feature learning. SAD loss improves the spatial spread of peak locations in the feature response maps, and more spreadness indicates higher interpretability.

The Feature Activation Diversity (FAD) loss requires filters to be insensitive to the occluded part, hence more robust to occlusion. The input to the lower network branch is a synthetic occluded version of the top branch input. This way, parts of the face representation sensitive to the occlusion are masked out, and training is performed to identify the input face solely based on the remaining elements. As a result, the filters that respond to the non-occluded parts are trained to capture more discriminative cues for identification. Feature Activation Diversity loss encourages that any local face area only affects a small subset of the filter responses, thus learning part-based face representations, when learning the network model. Therefore, the learned filters are also robust to occlusions. This is achieved by leveraging pairs of face images (one of them with a synthetically occluded region), enforcing the two feature representations to be similar.

The implementation of the interpretable face model using the SAD and FAD loss functions is available at:

https://github.com/yubangji123/Interpret_FR

3. Planned XAIface Contributions to AI-based Face Recognition Explainability

In this section, we describe preliminary ideas and methods for explanations as well as methods under development by the XAIface consortium. Later versions of the document will cover the detailed description of successfully implemented approaches as well as (basic) performance evaluations.

3.1. Explainable Face Recognition by Interpretable, Local Features

3.1.1. Introduction

Although explaining the contribution of several parts of a face image using standard explanation methods (e.g. Class Activation Maps (CAM), Grad-CAM, Grad-CAM++ or Saliency Maps) is very useful, the principle disadvantages of such methods already mentioned (e.g. extra layer required, runtime issues, equivalency not guaranteed) are a major drawback. In addition it is most important that it cannot be always guaranteed that the identified regions of influence are really meaningful and reasonable for the end-users.

Thus it might be difficult for end-users to understand the reasons a face recognition task fails, and - more important - it would be rather impossible to take any countermeasures to improve the recognition process by altering the base (face-acquisition) situation with respect to any changes possible through end-users interactions (e.g. by changing the light conditions, acquisition geometry or presentation of different viewpoints).

The proposed method will target all the issues mentioned above and as described in more detail below, we plan to develop a novel, general method with the ability to map novel, human understandable and reasonable (local) features to influence and importance parameters. The information gained will enable end-users of the recognition system to understand and uncover potential problems of the face-recognition process and allow for performing proper countermeasures.

3.1.2. Description

A straightforward approach to explain the end-to-end face-recognition process is the investigation of locally similar perturbations on input images or features following the paradigm of local interpretable model-agnostic explanations (LIME), but it is not guaranteed that the results are really meaningful and understandable for the end users. For instance it may be not intuitive, how single pixel perturbations can be altered or even worse, how single face-feature dimensions can be influenced.

The same non-interpretability with respect to the original face-image presented is true for the application of per se explainable and understandable algorithms (as eg. **decision trees**) on face-features directly, because the face-features in the latent space used are usually non-locality preserving, abstract representations (e.g. 512 dimensional ArcFace-Features), and the gained knowledge about the influence of the individual feature dimensions has no clear relation to the locality of origin (root cause) in the input face-image.

To overcome these issues mentioned, we propose to combine the two basic ideas of (a) locally changing the input data of the face-recognition pipeline and (b) inherent explainability of a specific classification method. In particular we plan to develop (learn) new human-understandable face-region descriptors and in addition (and in contrast to other approaches) we will specifically tune the changes in the input with respect to local and interpretable (meaningful) image properties around distinctive face-regions.

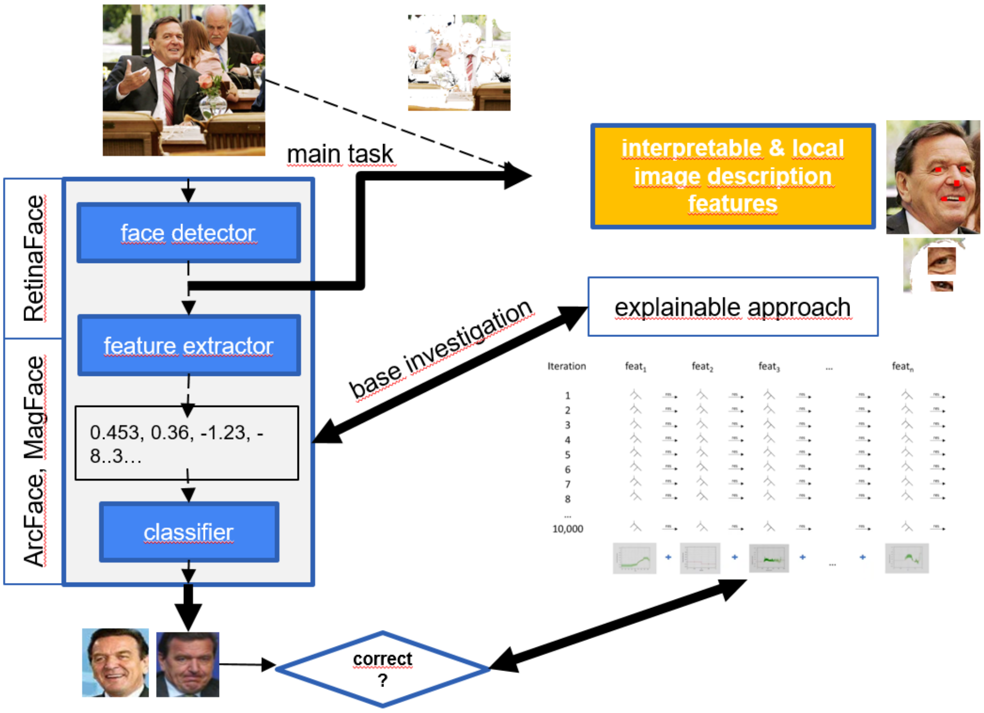

Figure 3.1a: Principle of the proposed locally, interpretable boosted features (LIBF) approach. Local, interpretable features around distinctive points (e.g. anchor points) in the face - thus meaningful and understandable for the end user - are used to train an inherently explainable approach and estimate probability for correct decisions of the native pipeline.

So the idea is to use the reference-face-recognition-pipeline defined in the XAIface project (RetinaFace face-detector and ArcFace or MagFace feature extraction and classification approach), and try to train an inherently explainable approach (e.g. decision tree or similar approaches like the recently proposed **explainable boosting machine**) in parallel to the reference pipeline providing the same (almost similar) recognition results or estimate the probability for correct decisions of the native pipeline (see Figure 3.1a for illustration). This is similar to already known “black-box” end-to-end approaches, but as a main difference to them we will not work on original input data (face-image) but on other (semi-) local face and image features (easily interpretable by humans). Potential candidates for such image features could be local (e.g. blurriness-levels around anchor points/facial-landmarks, colour histogram at nose tips etc.) or global (e.g. the amount of noise on background or foreground only). Another possibility may be to learn them using self-supervised face-image processing tasks.

Summing up, we will develop a new black-box explanation method for face recognition tasks which does not alter the recognition queue. It should provide a different way for explaining most important regions in face-images for the recognition task, coevally taking into account human understandability and thus providing the ability to interact for countermeasures in the case of failures.

Besides the proof of the feasibility of the approach and suitability of the explainable boosting machine for uncovering local feature importance, the main research question will be the definition (design), selection and evaluation of the proper, novel face-image features.

Finally we want to mention, that the basic idea can also be seen as a generic approach for XAI in even other domains and it will help to avoid (classical) pitfalls e.g. uncovering the presence of not task related image features (e.g. specific background for a certain class) feigning perfect recognition/classification performance.

3.1.3. Performance Evaluation

The entire evaluations are conducted during working on the project. Thus this section serves as a log for the main results to be shared within the consortium. The final results will be reported in the second version of the deliverable.

3.2. Improving Face Verification Explainability using Spatially-Biassed Similarity Metrics and Training Loss Functions

3.2.1. Introduction

Convolutional neural networks (CNNs) have achieved top performance for many computer vision tasks, notably face recognition. However, there is still a lack of effective processes to explain the complex decision-making process of deep learning (DL)-based solutions, especially due to their large, non-linear components. This situation is largely known as the "black box" approach, basically recognizing that the performance may be excellent but there is no clear idea how the final result is obtained. This limits the application of this technology, especially when the decisions have serious implications in real-worlds applications, e.g. access to critical infra-structures, gender or race discrimination, etc.; in fact, unexplainable false-positive results may lead to serious security and privacy issues. In this context, it is crucial to improve the transparency of the decision-making process for DL-based face recognition solutions which are the target of the XAIFace project. Thus, with the help of explainable AI, it is possible to understand and trust more the DL-based models results.

3.2.2. Description

It is well-known that DL-based face recognition results are largely impacted both by the loss function at training time as well as the features/description similarity metric at decision time. It is also intuitive that the similarity between two faces is not equally determined by all parts of the corresponding faces, notably face landmarks may be more relevant.

In the selected XAIFace reference face recognition pipelines, notably based on ArcFace and MagFace, two images are said to be from the same person simply if the adopted similarity metric is larger than a specific threshold value. However, this standard procedure does not help in interpreting and explaining to humans the relevance and impact of the high-dimensional features involved in the recognition process.

In this context, the plan for this novel explainable DL-based face recognition solution is to include new similarity metrics and training loss functions that consider the different impact of the various spatial regions in a face image, thus offering a local, spatially-biased approach. Naturally, the face landmarks may have a special impact in face recognition, which is not even always the same among them, e.g. eyes, nose and mouth. The target is that this novel solution provides quantitative and qualitative reasons to explain why two face images are from the same person or not. For example, if two face images are associated with the same person, the proposed solution may identify which parts of the face were more impactful and representative, e.g. by providing local similarity values. In this spatially-biassed process, it is also possible to include a spatial attention model, which should replicate the attention that humans dedicate to different regions when they recognize faces.

However, explainability capabilities, notably involving new similarity metrics and training loss functions, should not come at an unreasonable price in terms of recognition performance, notably for the verification protocol. Therefore, a novel explainable DL-based face recognition solution must offer an appropriate balance between verification accuracy and human interpretability, e.g. through some meaningful spatial maps which offer spatial explanations for the final decisions. By using the local similarities with different weights, the proposed, improved solution has the potential to become more robust than the original one for partially occluded faces.

Since it is common to perform face recognition on images which have suffered image compression, the recognition performance of the proposed explainable DL-based face recognition solution will also be studied for decoded images, not only using the so-called conventional image codecs, e.g. JPEG, JPEG 2000, JPEG XL, but also the recently emerged DL-based image codecs, notably those presented in the context of the JPEG AI project. In this context, it will also be possible to explain and assess the impact of coding the various face landmarks with different qualities after face and landmarks detection, if image encoders including that facility are available.

3.2.3. Performance Evaluation

The novel similarity metrics and training loss functions will be developed and assessed using the two XAIFace reference face recognition pipelines as baseline solutions and anchors (notably ArcFace and MagFace) under the recommended verification protocol.

3.3. Improved RISE Algorithm for Explainable Face Recognition System

3.3.1. Introduction

Face recognition has been a crucial task in recent years and the recognition systems have been widely deployed to various applications, such as smartphones or surveillance. With the development of deep convolutional neural networks (DCNN) and an increasingly large amount of available dataset, the methods developed have shown remarkable results on face recognition tasks with DCNN-based approaches. However, beyond the exceptional performance, deep face recognition models trained with large scale datasets were usually treated as "black-boxes", because the model designers can only have an understanding of the dataset or loss functions for training, but very limited understanding of the learned model itself. The deep learning-based recognition systems need more explainability and transparency so that people can truly trust the results they predict or understand the possible failures from them.

3.3.2. Description

Explainable face recognition is exactly a problem of explaining the matches returned by a recognition system in order to provide insight into why a probe face image can be matched with one identity over another. An explainable face recognition method should be able to generate attention or saliency maps which best explains which regions in a probe image would match with a mated image.

Under this context, we plan to work on an explainable face recognition algorithm based on ArcFace/MagFace (the deep face recognition pipelines to work on agreed in the project), and the RISE explanation algorithm. The RISE algorithm measures the importance of an image region by perturbing the image and observing how much this affects the black box prediction. In detail, an input image is firstly element-wise multiplied with some randomly generated masks. The masked images are then fed to the black box model. In the end, the final saliency map is generated by performing a linear combination of the masks and the weights, which comes from the predicted scores of target classes. This simple yet powerful method interprets how important each pixel of the image is with respect to the network prediction. This can be applied to the explanation face recognition task by indicating and comparing the importance region of both probe and gallery faces.

However, RISE algorithm is not designed for an explainable face recognition task by nature. First of all, in a face recognition system, there are no prediction class scores serving as weights to formulate the final saliency map. Secondly, RISE generates random binary masks with uniform prior, which is not efficient, because the important regions in a facial image are more likely to appear in certain areas. We plan to improve it by introducing a prior distribution to guide the mask generation process.

Ideally, given a triplet of probe, mate, non-mate images, a well-developed explainable face recognition system should be able to generate a saliency map that captures the region in the probe image that is most similar to the mate while the least similar to the non-mate. However, since the probe and non-mate images are from two identities, they are different by nature not only from their faces but also backgrounds. There is no dataset with suitable groundtruth to validate the effectiveness of a proposed explainable face recognition approach. Therefore, another challenge beyond this problem is how to properly validate or benchmark the output saliency map of the explainable face recognition method.

We derive our motivation from the work of (Williford et al. 2020), who proposed an 'inpainting game' and created paired data by erasing some key components on face images with inpainting tools. We propose to create the pairwise ground truth with popular deep-fake creation techniques to manipulate the mate face images in a more realistic way. In detail, we plan to create non-mate images by swapping the certain regions of a mate face image, such as mouth, eyes, and nose, with another identity with learning-based deepfake creation tools. Based on the new triplet of data, i.e probe, mate, and deep-fake non-mate images, we hope to design a protocol for explainable face recognition methods.

3.3.3. Performance Evaluation

The evaluation will start when the proposed explainable face recognition method is developed. Suitable protocols and evaluation experiments will be designed.

3.4. Demographic Information Disentangling

3.4.1. Introduction

Face recognition (FR) is currently predominantly based on convolutional neural networks (CNNs). CNNs are used to extract face representations that can be used for several tasks, including identity recognition. Although CNNs are intended to generate representations encoding only the identity information, recent studies have shown that information about soft-biometric traits, including gender, age, and other demographics, is also encoded in these representations (Dantcheva et al. 2015) (Dhar et al. 2020) (Nagpal et al. 2019) (Parde et al. 2017). Soft biometrics are attributes that are not necessarily unique to an individual but can be used alone or in conjunction with primary biometric traits for a variety of applications. However, since these attributes can be extracted from the computed face representations and may potentially be misused, they represent a considerable privacy risk.

3.4.2. Description

For the so-called black-box FR systems, it is not clear how face features such as gender, age, and ethnicity are encoded in the overall face description, namely the face template. Several works have addressed the problem of masking or altering the face representation in order to conceal the soft-biometric traits (gender, ethnicity, age). There are two approaches: image-level techniques operate by suppressing the soft-biometric information in face images so that machine-learning based models, such as CNN, fail to infer it (Mirjalili et al. 2019) (Rozsa et al. 2019). However, they rely on pre-trained classifiers to learn the perturbation to be applied to the image and thus do not generalise well for other classifiers. Template-level approaches try to suppress the soft-biometric information from the more compact face representation encoded in face templates. Existing techniques from this group were shown to generalise well to arbitrary attribute classifiers (Terhörst et al. 2019).

In XAIface, instead of masking or altering the gender, age, or ethnicity information, we plan to develop techniques for disentangling demographic information from the overall face representation in order to understand the impact of such traits on face recognition but also to develop demographic-free face recognition. The latter will also indirectly address fairness and non-discrimination issues by following the idea of de-biasing during the training, as the only information used by the FR pipeline will be related to identity and not to other soft biometric traits.

An autoencoder will be used to reduce the dimensionality of a facial biometric template (feature vector) and at the same time to separate identity information from demographic information. To do this, specific loss functions will be used to drive the separation of the information into two smaller vectors.

3.4.3. Performance Evaluation

Performance evaluation will be carried out by analysing and measuring the impact of demographics on face recognition in terms of loss/gain in performance, both on the whole dataset and on subsets of specific classes. Face recognition performances will be assessed using the metrics recommended in the ISO/IEC 19795-1, that is false match rate (FMR), false non-match rate (FNMR) for verification (one-to-one), and false-negative identification-error rate and false-positive identification-error rate for identification (one-to-many), as well as with graphical illustrations through the detection-error tradeoff (DET) plot.

3.5. Explainable Soft Biometrics Estimation

3.5.1. Introduction

Human face images encode different types of biometric information. Soft biometrics such as gender, height, and weight do not have the capacity to differentiate between two different identities, however they can be useful to improve the quality of different systems. Among those, weight is also an indicator of both physical appearance and health conditions and unlike gender and height, body weight changes during a person's adult life and needs to be periodically measured. Conventional weight measurement techniques require the cooperation of the subject to be measured, which might not be possible during medical emergencies, video surveillance for criminal pursuit or due to different patient disabilities. When non-cooperative scenarios occur, visual estimation of the weight of the patient by a health professional is preferred but such estimations might not always be accurate.

3.5.2. Description

Nowadays, deep learning technologies provide new solutions to obtain end-to-end learning models that gain knowledge and insights from complex, high-dimensional biomedical data. However, the general user might be still sceptical when facing black-box approaches in applications where model interpretability is a concern. Understanding the decision making of a predictor is crucial when actions are taken in the medical domain. Assessing trust in a model cannot be achieved only through accuracy metrics, a trustworthy model should be evaluated using interpretability techniques.

In this context, we aim to increase the trust of the user in soft biometrics estimation models, especially for one of the most discriminative as it is the weight. We intend to do it by complementing the prediction delivered by the deep learning network with a visualisation of the most contributive facial regions that lead to it. To this end, we explore two model-agnostic explainability techniques, SHAP and LIME. The interpretability techniques do not assess the validity of the result, instead, they give complementary information on which images areas were most significant for the prediction.

As a result of the regions highlighted by the interpretability approaches, we will focus on assessing the impact of different occlusion factors on the weight estimation model. This study on the important face regions for the weight estimation will be further explored by occluding different areas and visualising with post-hoc interpretability tools how the model predicts without being provided with the occluded region. We will also evaluate different face detection and cropping techniques to assess whether different detection methods will consider more of the most meaningful areas highlighted by the interpretability approaches. As suggested by the explainability techniques, the occlusion or omission of those areas will lead to less accurate predictions, causing a misestimation of the soft biometric (in this case the weight) thus misleading the face recognition system.

3.5.3. Performance Evaluation

Different model-agnostic explainability techniques will be applied to the deep learning networks in order to explain the predictions by understanding the most meaningful face areas for the model that lead to the final estimation. To evaluate the validity of the regions delivered by the interpretability approaches, consistency between the returned face areas is expected.
Suitable protocols and evaluation experiments will be designed during the implementation of the above mentioned techniques.

4. Addressing Ethical and Legal issues

4.1. Explainable approach design and implementation

According to the project proposal, the aim of WP 4 is to research “an approach to explaining decisions of face recognition systems, approach for explaining decisions of face recognition systems, building on information contributed from the different influencing factors in the form of metrics developed in WP2 and data/metadata obtained in WP3. The WP aims to design a protocol that is applicable to different state-of-the-art face detection and recognition pipelines that include AI components.”

The work package is separated into 3 tasks:

T4.1	**Explainability protocol and methods (M06 – M24: 1; 2, 3, 4, 5)** This task researches and develops components of a face recognition pipeline producing explanatory information, addressing both the cases of creating transparent models and post-hoc explanation of black-box models. It also develops a protocol for implementing them in face recognition systems relying on AI components.
T4.2	**Implementation of explainability approach (M06 – M24: 2; 1, 4, 5)** The protocols and methods developed in T4.1 will be prototypically implemented for at least two different state of the art face recognition pipelines. As these pipelines will differ in which components are AI-based (e.g., face detection, feature extraction) and for which transparent models are used (e.g., for matching/classification), as well as in the network architecture (e.g., relying on common backbones vs. multitask networks for the specific problem), the implementation of the protocol will be specific for each of the methods. It is also expected that the trade-off between interpretability and performance will be different for each of the pipelines.
T4.3	**Communicating explanations to users (M12 – M36: 1; 2, 3, 4, 5)** This task analyses how information produced in T4.2 can be visualised or verbalised to be understandable to users. It will specify approaches to be integrated into user interfaces to summarise explanations for non-expert users and enable them to give feedback that is valuable for further training.

4.2. Scope

According to task 4.1, components shall be developed, which would produce additional **explanatory information**. Furthermore, the models should be **transparent**. The analysis in task 4.3 explores how such information can be visualised or verbalised. The aim is to deliver understandable information to users and summarise such information for non-expert users.

The major topics which must therefore be addressed within this work package are explainability and transparency obligations. These issues must be analysed from a legal and an ethical standpoint. While the legal analysis deals with the status quo of regulation as well as foreseeable changes in the law, the ethical perspective handles moral issues that may go beyond current regulation.

It should be noted that explainability and transparency requirements differ depending on the specific context in which a system is used. Recent EU-legislation has chosen a risk-based approach to regulation. As a result, the same system may have to adhere to different standards depending on the level of risk involved. Additionally, national legislation may add to those requirements. The starting point of the entire analysis are base requirements, which all or at least most systems must fulfil.

The two main questions , that shall be answered are the following:

A. **Is there a legal obligation to be transparent and to provide explanations for decisions of facial recognition systems?**
B. **Is there a legal right to obtain explanations of facial recognition systems?**

In a preliminary meeting of UNIVIE, the following additional subpoints were identified:

Legal Perspective:

a) What is "explainability” and how is it linked to transparency?
b) What is “transparency”?
c) What transparency obligations does a controller have?
d) To whom does the controller have an obligation to transparency?
e) What rights relating to transparency does a data subject have?
f) What information needs to be communicated to data subjects?
g) How must the information be conveyed to data subjects?
h) Must an AI-based face recognition system be explainable?
i) If so, what information must be conveyed to data subjects?
j) If so, how must the information be conveyed to data subjects?
k) What additional requirements will (probably) be added by future EU-regulation?

Ethical Perspective:

a) What are the ethical requirements concerning transparency and explainability according to the High-Level Expert Group on AI and similar authorities?

4.3. Legal Requirements

4.3.1. Explainability & Transparency

For the legal analysis, the terms “explainability” and “transparency” must first be explored. Although classification varies depending on the model and taxonomy, transparency, alongside interpretability, can be seen as a subcategory of explainability.[1] Contrary to this view, the High-Level Expert Group on Artificial Intelligence categorised explainability as a subset of transparency.[2] In a study prepared for the members and staff of the European Parliament, "algorithmic transparency" was defined as follows:

> *"Depending on the type and use of an algorithmic decision system, the desire for algorithmic transparency may refer to one, or more of the following aspects: code, logic, model, goals (e.g. optimisation targets), decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs. Algorithmic system transparency can be global, seeking insight into the system behaviour for any kind of input, or local, seeking to explain a specific input - output relationship."*[3]

The study describes transparency as a prerequisite for accountability.[4] In order to ensure accountability, a certain amount of information must be provided. The authors of the document identified seven potential areas of transparency for machine learning systems:[4]

1. **Data:** refers to raw data, sources, pre-processing and collection methods;
2. **Algorithms:** refers to testing outputs against inputs;
3. **Goals:** refers to relative priorities of respective goals;
4. **Outcomes:** refers to the outcome of the actual deployment of a system;
5. **Compliance:** refers to the reports on overall compliance of an operator or manufacturer;
6. **Influence:** refers to revealing own interests or third party interests;
7. **Usage:** refers to what personal data is used.

Furthermore, transparency can be differentiated based on the addressees. A system can be transparent vis-à-vis everyone, authorities, third-party analysts, researchers[5] or data subjects.

The term "explainability" still eludes legal scripture. Attempts to define the term have been undertaken[6], but no definition has been agreed upon. However, the key points are captured by a similar debate on the "right to explanation". *Wachter et alia* concisely defined the potential points an explanation to an automated decision would have to include[7]. According to the authors, the information can be separated into two categories: system functionality and specific decisions. While system functionality includes information about the logic of the system, significance, envisaged consequences, decision trees, pre-defined models, criteria

[1] Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 118.
[2] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 18.
[3] EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 4.
[4] EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 5.
[5] EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 6.
[6] Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 117.
[7] Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) Vol. 7/2, 76.

and classification structure, the second category would include more information about the specific decision such as the rationale, the reasons, individual circumstances (weighting of features, profile groups etc.).

Wachter et alia further elaborated that explanations can be distinguished by their timing. An explanation can be ex ante and would therefore be limited to system functionality. Alternatively, an explanation can be provided ex post and could incorporate both categories. This structure has since guided the legal debate.[8] The definition in the field of ethics will be addressed in the chapter on ethical requirements.

4.3.2. Current obligations according to the GDPR

General remarks

A key regulation that instituted obligations to transparency was the GDPR[9]. With the GDPR, a directive on the processing of data in the context of criminal law was created, which also contains provisions on transparency.[10] Since this context is not a basic use case, the directive will only be mentioned for the sake of completeness.

The term "transparency" is not defined in the GDPR itself.[11] Recital 39 of the GDPR establishes that transparency is one of the principles of the GDPR and asserts that "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".[12] **From the excerpt one can already derive that transparency according to the GDPR is mostly an obligation vis-à-vis the data subjects.**

Hence, the main objective of the provisions is to ensure accountability of data controllers and to empower them to exercise control over their data[13].

According to the material scope, the obligations in the GDPR apply to the wholly or partly automated processing of personal data[14]. The creation and use of a face recognition system usually falls within the scope. The problem of which criteria must be fulfilled for data to

[8] For further references: Kim/Routledge, Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach (2020), available at: http://dx.doi.org/10.2139/ssrn.3716519.

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (referred to as GDPR).

[10] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

[11] Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 6.

[12] Rec 39 GDPR.

[13] Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 5.

[14] Art 2 GDPR.

constitute personal data will be addressed in another deliverable. It must however be noted that some relevant exemptions exist for the processing of certain competent authorities (criminal law) or organisations of the European Union[15]. This is due to specific regulations in these sectors.

The main addressee of the provisions of the GDPR is the "controller". According to Art 4 cif (7) GDPR a "controller" means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law. Within the lifecycle of a face recognition system a typical controller would be the developer with regard to the data (images) used to train the models. Afterwards, the system may be used by an organisation, who may be considered the controller concerning the data used in the productive stage. Hence, these controllers need to fulfil their obligations towards their respective data subjects.

Art 5 GDPR established the key principles relating to processing of personal data. According to the article, personal data shall be "processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')".[16] Art 5 par 2 GDPR highlights that it is not only the obligation of a controller to ensure compliance with the transparency obligation, but also to be able to demonstrate compliance.

Modalities of transparency

Section 1 of Chapter III of the GDPR specifically deals with transparency and modalities. Article 12 is titled "transparent information, communication and modalities for the exercise of the rights of the data subject." The actual information that must be provided to data subjects is enlisted in Art 13 & 14 GDPR. According to paragraph 1 of Art 12 GDPR, this information must be provided "in a **concise, transparent, intelligible and easily accessible form**, using clear and plain language, in particular for any information addressed specifically to a child."

The Art 29 Working Party states that "intelligible" means that an average member of the intended audience should be able to understand the information.[17] If there is uncertainty of the intelligibility, it should be tested (p.ex.: readability testing). The standard applied to controllers in relation to linguistic presentation is high.[18] "Easily accessible" means that it should be obvious where the information is provided. It should not be the task of the data subject to seek out the relevant information.[19] Common examples would be pop-ups and prominently displayed privacy notices.

[15] Art 2 par 2 lit d, par 3 GDPR.
[16] Art 5 par 1 lit a GDPR.
[17] Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 7.
[18] Schrey in Rücker/Kugler, New European General Data Protection Regulation (2018) 128.
[19] Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 7.

Furthermore, the information must (generally) be either provided **in writing or other (electronic) means.** The complete information should be provided within one document. However, a layered approach is recommended.[20] Even though the obligations may result in some workload for the controller, the information must always be provided free of charge. The data subject must be informed at the time, when the data is obtained.[21]

Visualisation

The information may also be provided **in combination with so-called standardised icons**. The goal would be to aid visibility, legibility and to provide a meaningful overview over the processing. Such icons must be machine-readable, if they are presented electronically.[22] According to the Art 29 Working Party, a variety of visualisation tools may be used: icons, certification mechanisms, data protection seals and marks.[23] However, visualisation tools may only be used in combination with and not as total replacement for language.[24] The GDPR requires the controller to make use of standardised icons to increase the utility. Since no decision could be reached on the matter, no annex with standardised icons was delivered with the GDPR. Rather, it is now the task of the European Data Protection Board and the European Commission to create a draft.[25] Recital 58 of the GDPR highlights that visualisation may be used in addition to language. However, for the specific information according to Art 13 & 14 GDPR only standardised icons may be used.[26]

Information

Art 13 & 14 GDPR specify the information that must be provided to the data subjects. Both articles contain similar lists but differ in the scope. Art 13 GDPR is applicable if the data were obtained directly from the data subject. Otherwise, Art 14 GDPR applies. A typical scenario where Art 14 GDPR applies instead of Art 13 GDPR, would be the training of models with data from an available online database.

The list of information is extensive, so only a portion of particularly relevant information will be listed below. For a practical overview in table form the authors refer to the annex of the Guidelines on transparency under Regulation 2016/679 of the Article 29 Working party.

The controller must provide information about themselves and potentially their representative and data protection officer.[27] The data subject must be informed about the purposes of and legal basis for processing[28], storage time[29] and the individual rights of the data subject.[30]

[20] For more information: Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 11.
[21] Arg.: Art 13 par 1 GDPR.
[22] Art 12 par 7 GDPR.
[23] Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 25.
[24] Franck in Gola, DS-GVO2 (2018) Art 12 cif 47.
[25] Franck in Gola, DS-GVO2 (2018) Art 12 cif 49.
[26] Veil in Gierschmann/Schlender/Stentzel/Veil (Eds.), Datenschutzgrundverordnung (2018) Art 12 cif 28.
[27] Art 13 par 1 lit a, b GDPR.
[28] Art 13 par 1 lit c GDPR.
[29] Art 13 par 2 lit a GDPR.
[30] Art 13 par 2 lit b, c, d GDPR.

Furthermore, the controller must inform about the existence of automated decision-making, including modalities.[31] Due to the importance of this specific topic and its inextricable link to explainability, it will be discussed in a separate paragraph below. Art 14 GDPR contains a slightly more extensive list. A controller must inform about the categories of personal data involved.[32] Additionally, the source of the data must be revealed.[33]

In addition to these obligations, required information may also include inter party communication on the exercise of the subjects' rights (Articles 15-22 GDPR) and communications in relation to data breaches (Article 34 GDPR).

Right to access and specific provisions on automated decision-making

As a counterpart to the obligation of the controller, a data subject has the right to obtain information about the processing activities. Such information includes, for example, the purposes of the processing[34] and the categories of personal data concerned.[35] Art 15 par 1 lit h also grants the right to data subject to be informed about the existence of automated decision-making, including profiling. The specific provisions on automated individual decision-making in the GDPR can be seen as the connecting piece between transparency, explainability. Art 22 GDPR primarily contains the right of any data subject to not be subject to a decision solely based on automated processing, including profiling. The data subject does not need to actively invoke the right for it to apply.[36] The right is restricted to cases, where the automated processing produces legal or similar effects for the data subject. The right does not apply if explicit consent is provided[38], the decision is necessary for a contract[37] or the decision is authorised by the law of the controller and suitable safeguards for the data subject are provided.[38] In case one of the exceptions applies, the controller must inform the data subject about: "[…] the existence of automated decision-making, including profiling, referred to in Article 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."[39] Hence, the question arises, what the exact telos of the norm is.

Scope

The scope of the provision is limited to specific scenarios. Despite the fact that profiling is mentioned in the title of Art 22 GDPR, it is not a necessary element for the application. The concept of profiling can be neglected for the purposes of this document, since facial recognition systems will rarely fall under the definition due to the lack of evaluation of personal aspects of the data subject.[40] The concept of "automated decision-making",

[31] Art 13 par 2 lit f GDPR.
[32] Art 14 par 1 lit d GDPR.
[33] Art 14 par 2 lit f GDPR.
[34] Art 15 par 1 lit a GDPR.
[35] Art 15 par 1 lit b GDPR.
[36] Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 19
[37] Art 22 par 2 lit c GDPR.
[38] Art 22 par 2 lit b GDPR.
[39] Art 13 par 2 lit f, Art 14 par 2 lit g GDPR.
[40] Compare to Art 4 cif 4 GDPR.

however, cannot be dismissed as quickly. Nevertheless, both concepts have a significant overlap.[41] Art 22 GDPR only applies in cases of solely automated decision-making, which "[…] is the ability to make decisions by technological means without human involvement".[42] If a human is involved in the decision-making process, Art 22 GDPR usually does not apply, insofar as the contribution to the decision by the human is significant enough and not just formal.[43] Additionally the automated decision-making must have legal or similar effects for the data subject. Examples of legal effects mentioned by the Art 29 Working Group include refused admission to a country or denial of a particular social benefit granted by law. Also a person may be similarly affected if the decision leads to exclusion or discrimination.

In the case of a facial recognition system, the application of Art 22 GDPR depends on the context it is used in/for. The argument could be made that at least in case of a classification system as referred to in in-depth analysis of the European Parliamentary Research Service[44], a decision with potential legal or similar effects for the person is reached fully automatically. Other examples found in the study would be the use for crime prevention at train stations or crime investigations at the 2017 G 20 summit in Germany[45]. Even though it is questionable, if one would have to include potential false positives, such as wrongfully being identified as a suspect, the authors would argue for such an interpretation. After all, the objective of the provision is to address the risk of automated processing. Even if the processing does not strictly fall within the scope of Art 22 GDPR, the Art 29 Working Group recommends fulfilling the additional transparency obligations.[46]

Right to Explanation

As mentioned above, a controller employing automated decision-making must provide additional information to their data subjects:

a) The fact that the controller is employing the technology
b) Meaningful information about the logic involved
c) Explanation of the significance and envisages consequences of the processing.

According to the guidelines on automated individual decision-making, it is the obligation of the controller to find simple ways to explain the "rationale behind, or the criteria relied on in reaching the decision"[47]. Understanding the process can be particularly difficult if machine-learning is involved.[47]

The language of Art 22 GDPR and Art 13 & 14 GDPR appears somewhat ambiguous even after the clarification of the Art 29 Working Group. This fact was noticed even before the GDPR came into force by various authors, who published essays questioning the right under

[41] Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 8.
[42] Ibid.
[43] Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 21.
[44] Madiega/Mildebrath, Regulation facial recognition in the EU (2021) 2.
[45] Madiega/Mildebrath, Regulation facial recognition in the EU (2021) 37.
[46] Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 25.
[47] Ibid.

the lens of a “right to explanation”.[48] It is not unreasonable to prima facie detect a right to explanation in those articles, especially if one considers the respective Recital 71: “In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **obtain an explanation of the decision reached after such assessment** and to challenge the decision.” However, Recital 71 is not legally binding and the right to explanation is not explicitly mentioned in the safeguards of Art 22 par 3 GDPR.[49]

According to the elaborations of *Wachter*, *Mittelstadt* and *Floridi*, a right to explanation would have to include not only information on system functionality such as the logic, significance (ex-ante) and consequences, but also information on the specific decision (ex-post) such as the reasons, weighting of features and individual circumstances.[50] Within their paper, the trio comes to the conclusion that, since no ex-post explanation of the specific decision must be provided, no complete right to explanation exists.[51]

For the purposes of this analysis, it must therefore be concluded that no right to explanation currently exists. However, controllers have certain obligations to be transparent about automated decision-making towards their data subjects.

Exceptions and opening clauses

Art 11 GDPR

Art 11 par 1 GDPR states: “If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.” This exemption may apply, for example, if a developer uses a database for the creation of data derived from images without knowledge of and interest in the identity of the data subjects. In such a case, no further data must be collected to identify the data subjects and to fulfil the obligations to transparency. However, if a data subject provides additional information about their identity in accordance with Art 11 par 2 GDPR, a controller must still comply with the relevant provisions. It should be noted that Art 11 GDPR does not apply if national law applies due to the use of an opening clause.[52]

[48] Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) Vol. 7/2, 76; Mendoza/Bygrave, The Right not to be Subject to Automated Decisions based on Profiling, University of Oslo faculty of Law Legal Studies Research Paper Series No. 2017-20.

[49] See also Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) Vol. 7/2, 76 (80).

[50] Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) Vol. 7/2, 76 (78).

[51] Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) Vol. 7/2, 76 (82).

[52] Kampert in Sydow, Europäische Datenschutzgrundverordnung2 (2018) Art 11 cif 13.

Art 89 GDPRGDPR was the full harmonisation of data protection law, national legislators retained the right to regulate specific areas of data protection law. Art 89 GDPR contains an opening clause for research purposes and is therefore of particular interest. National legislation may contain specific provisions on transparency.[53]

4.3.3. Future obligations according to the AI-Act

General Remarks

In 2021 the European Commission proposed the Artificial Intelligence Act (AI-Act)[54] in an effort to introduce harmonised rules on artificial intelligence systems. The act marks a significant new milestone in the process of regulation of new technologies. The main objective is to ensure that AI systems placed on the EU-market are safe and that fundamental rights and EU-values are respected. The legal certainty provided by the act is supposed to foster innovation and investment. AI made in Europe shall be safe, lawful and trustworthy. In essence, the proposal is a product regulation. It includes design requirements for AI systems as well as obligations for import and usage of such systems. Specifically, Art 13 & 14 AI-Act contain provisions on transparency and human oversight, which may increase the required level of interpretability of AI-systems. Interpretability, as mentioned above, can be seen as one aspect of explainability.[55] However, these provisions only apply to so-called “high risk systems”. It should be noted that the legislative process is extensive and changes to the text of the regulation are expected. In the current version[56], facial recognition systems, as systems that would potentially use biometric data[57], even occupy a special position within the text. Naturally, the application of the specific provisions is dependent on factors such as intent and data used and cannot be decided categorically.

High-Risk Systems

The classification rules for high-risk AI systems are laid out in Art 6 of the proposed AI-Act. In essence, an AI system must be regarded as “high-risk” if it is either a safety component or a product which is covered by Union harmonisation legislation or is required to undergo a third-party conformity assessment. This, for example, includes sector specific regulation on machinery or medical devices.[58] Additionally, every system that is referred to in Annex III is a high-risk system. Annex III enlists specific areas, which would be deemed “high-risk” per se. The first area is of particular relevance for the project XAIface: “(a) AI systems intended to be used for the ‘real-time’ and ‘post’ remote biometric identification of natural persons;”[59]

[53] Details are provided in Deliverable 3.1.

[54] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (referred to as AI-Act).

[55] Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 118.

[56] Status: 21/04/2021.

[57] Art 3 cif 33.

[58] See Annex II to AI-Act.

[59] § 1 lit a Annex III AI-Act.

Nevertheless, not all facial recognition systems may fall within the provision. If they are used for other purposes than identification and Art 6 par 1 does not apply, the system may still be considered high-risk if it is used for specific intents in areas such as:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, workers management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes;[60]

These categories can be amended by the European Commission even after the regulation is in force.[61] For the current version of the proposal including the annex, it can be assumed that a significant portion of facial recognition systems would be classified as “high-risk” and would therefore have to comply with the obligations set out in Chapter 2.

Requirements according to Chapter 2

Art 13 AI-Act establishes new requirements for transparency and provision of information to users. Users as referred to in this article are not necessarily the data subjects or end users, but rather professional users of an AI-system.[62] Art 13 par 1 AI-Act requires high-risk AI systems to “[…] **be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output** and use it appropriately”[63] Therefore, transparency and interpretability must be “by design” rather than being established post-hoc.

Instructions

Additionally, a document with instructions must be provided to the users of the system.[64] As modalities Art 13 par 2 AI-Act requires that the information provided be in a digital or other format, concise, complete, correct, clear, relevant, accessible, and comprehensible to users. Similar to the respective provisions on transparency in the GDPR, it can be expected that the demanded standard for readability will be high.

The instructions mainly include information about the provider[65] and the system. The provider must list the characteristics, capabilities and limitations of performances of the system.[66] The system information includes its intended purpose, accuracy levels, robustness, cybersecurity, foreseeable misuse and risks, performance as regards the persons on which the system will be used and specifications for input data as well as

[60] § 2 – 8 Annex III AI-Act.
[61] Art 7 AI-Act.
[62] Art 3 cif 4 AI-Act.
[63] Art 13 par 1 AI-Act.
[64] Art 13 par 2 AI-Act.
[65] Art 13 par 3 lit a AI-Act.
[66] Art 13 par 3 lit b AI-Act.

relevant information on training, validation and testing. Importantly, the instructions must include so-called **"human oversight measures"** as referred to in Art 14 AI-Act. These measures include "[…] technical measures put in place **to facilitate the interpretation of the outputs of AI systems** by the users".[67]

Human Oversight

Art 14 par 1 of the AI-Act requires high-risk AI-systems to "[…] be designed and developed in such a way, including with appropriate **human-machine interface tools**, that they can be effectively overseen by natural persons during the period in which the AI system is in use." The aim of the measures is to minimise risks and foreseeable misuse.[68]

Whenever feasible, the measures shall be already built into the system by the provider before the product is placed on the market or put into service.[69] Otherwise the provider must identify the measures while they will be implemented by the user.

Art 14 par 4 then sets out the specific objectives such measures must enable a human overseer to achieve:

"(a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;

(b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system **('automation bias'),** in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons; (c) be able to **correctly interpret the high-risk AI system's output**, taking into account in particular the characteristics of the system **and the interpretation tools and methods** available;

(d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;

(e) be able to intervene in the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure."

Specifically for AI systems intended to be used for remote biometric identification of natural persons, the measures must ensure that no action or decision is taken as a result of the systems identification without **verification and confirmation by at least two natural persons.[70]**

For the purposes of this analysis, it must therefore be concluded that the new proposal of the AI-Act will significantly increase the requirements of interpretability and transparency of high-risk AI-systems by introducing new design and human

[67] Art 13 par 3 lit d AI-Act.
[68] Art 14 par 2 AI-Act.
[69] Art 14 par 3 lit a AI-Act.
[70] Art 14 par 5 AI-Act.

oversight obligations. These obligations will require providers to technically enable the correct interpretation of outputs of AI systems. Such measures must include human-machine interfaces that will allow for better interpretability.

4.4. Ethical Requirements

4.4.1. HLEG-Guidelines

The Ethics Guidelines for Trustworthy AI[71] are a central reference document for issues on AI. They promote trustworthy AI, which must be lawful, ethical, and robust.[72] The document, however, by its own definition does not provide guidance on lawful AI. Hence, these requirements were established in the chapter on legal requirements. The following section analyses the key requirements on transparency and explainability of the guidelines.

Transparency & Explainability

According to the guidelines, transparency of AI can be structured into three components: **traceability, explainability and communication.[73]** "Traceability" dictates that data sets and processes that yield a decision must be documented to increase transparency. Likewise, decisions of the AI systems should be traced to identify errors. "Explainability" is the ability to not only explain the technical process, but also the related human decisions. A human must be able to understand and trace the decisions. The guidelines assert that the person or persons concerned should have **a right to explanation**, whenever the AI system has a significant impact on their lives.[73] The explanation should be provided in a timely manner and suitable for the receiver. Furthermore, the guidelines demand transparency of the influences of AI-systems on organisation decision-making processes. "Communication" means that AI systems must not represent themselves as humans, but be transparent of their identity. Capabilities and limitations should be communicated appropriately to the parties involved.

The guidelines go into more detail on explanation methods in the section on technical methods. To be considered trustworthy, it must be understandable why a system behaved in a given manner or why it gave a certain interpretation.[74] The document states that "[…] to address this issue to better understand the system's underlying mechanisms and find solutions" is still an open challenge for some AI systems and that XAI research is vital.[75] For the concrete assessment of explainability, the authors refer to the relevant section in the assessment list in the guidelines.[76]

For the purposes of this analysis, it must therefore be concluded that a right to explanation is ethically required, whenever an AI system has a significant impact on a person or group of persons.

[71] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019).
[72] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 2.
[73] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 18.
[74] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 21.
[75] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 21, 22.
[76] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 29.

4.5. Preliminary Conclusion

The aim of this Section is to determine whether obligations to be transparent and to provide explanations for decisions of facial recognition systems exist. In turn, it was questioned whether there is a legal right to obtain explanations of facial recognition systems.

Having analysed the current state of EU-legislation, the authors reached the following conclusion:

- The GDPR contains specific provisions that oblige data controllers to be transparent about the data processing towards data subjects.
- Art 22 GDPR in conjunction with Art 13-15 GDPR only requires the controller to inform about the existence of automated decision-making, including profiling, and meaningful information about the logic involved, the significance and the envisaged consequences of the processing. Exceptions may apply.
- The articles do not provide for an obligation to ex-post explain specific decisions and therefore also do not contain a right to explanation in line with the current legal definition of such a right.

With regard to future EU-legislation, the authors reached the following conclusion:

- The proposal for the AI-Act in its current form will increase the level of transparency of high-risk AI systems, which will include a significant portion of facial recognition systems.
- The proposal for the AI-Act requires high-risk AI-systems to be designed in such a way, that their outputs are interpretable.
- Various technical tools for interpretability and human-machine interfaces will have to be provided.
- The document is still subject to change and results of the analysis are therefore only preliminary.

With regard to the Ethics Guidelines on AI, the authors reached the following conclusion:

- A right to explanation is ethically required, whenever an AI system has a significant impact on a person or group of persons.
- Providing solutions to make certain AI systems explainable is still an open challenge.

5. Summary

In this deliverable, the overview of the state of the art on AI explainability methods, focused on potential application for face recognition, is provided. This overview serves to give a necessary background for the development of new face explainability methods in the frame of the project.

In the overview (see Section 2), different categories of AI-explainability methods have been identified, including local and global methods, model specific and agnostic methods, white-box and black-box approaches, and gradient-based backpropagation as well as perturbation-based forward propagation algorithms. For each method, along with a brief description, we report which category it belongs to and what its possible use in XAIface might be. In addition, where possible, links to the implementation of the method are given. Many of the AI explainability methods in Section 2 are model-agnostic and applicable to black-box models. This allows a lot of flexibility in choosing the architecture of the face recognition model. However, a drawback of these methods is that their output can be very generic. Thus, one of the objectives of XAIface is to map novel, human understandable and reasonable (local) features.

Section 3 outlines the work plan and reports the first ideas for the development of new methods for AI face explainability in XAIface. Finally we want to mention again, that the main goals for the methods to be developed in the project are to: (i) develop AI explainability methods specific for face recognition; (ii) to provide meaningful and reasonable feedback for the end-users, (iii) to disentangle demographic information from identity to protect people's privacy; (iv) to understand what other information can be extracted from the face (soft biometrics) apart from identity.

Finally, a preliminary assessment of potential legal and ethical issues is provided. The aim of Section 4 is to determine whether obligations to be transparent and to provide explanations for decisions of facial recognition systems exist. In turn, it was questioned whether there is a legal right to obtain explanations of facial recognition systems.To summarise, while there is an ethical requirement to make AI systems explainable, the current EU-legislation does not provide for a general right to explanation of or an obligation to explain decisions of facial recognition systems ex post. Depending on the context, data subjects may have a right to obtain ex-ante information such as the logic involved. Future legislation may change the status quo insofar as outputs of high-risk AI-systems must be interpretable.

Concerning the project XAIface, it must therefore be concluded that the efforts fall in line with the direction of future legislative projects. Even though not all methods may currently be legally required, there is an ethical obligation to ensure explainability. The concrete methods to allow for interpretability are not specified in the proposal of the AI-Act. Rather, they must be implemented or at least provided for by the provider of an AI system. The researched methods may be used to achieve compliance with future legislation. As insinuated in Art 13 and 14 of the proposal of the AI-Act, however, the choice of methods and interpretation tools will depend on the use case.

References

Arbabzadah, Farhad, et al. “Identifying individual facial expressions by deconstructing a neural network.” *German Conference on Pattern Recognition, Springer*, 2016, pp. 344-354.

Arras, Leila, et al. “" What is relevant in a text document?": An interpretable machine learning approach.” *PloS one*, vol. 12, no. 8, 2017.

Auret, Lidia, and Chris Aldrich. “Interpretation of nonlinear relationships between process variables by use of random forests.” *Minerals Engineering*, vol. 35, 2012, pp. 27-42.

Bach, Sebastian, et al. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” *PloS One*, vol. 10, no. 7, 2015.

Chattopadhay, Aditya, et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks.” *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 839-847.

Chen, Tianqi, and Carlos Guestrin. “gboost: A scalable tree boosting system.” *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

Dantcheva, Antitza, et al. “What else does your biometric data reveal? A survey on soft biometrics.” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, 2015, pp. 441-467.

Deng, Houtao. “Interpreting tree ensembles with inTrees.” *International Journal of Data Science and Analytics*, vol. 7, 2019, pp. 277–287.

Dhar, Prithviraj, et al. “How are attributes expressed in face DCNNs?” *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 85-92.

Hastie, Trevor, and Robert Tibshirani. “Generalized additive models: some applications.” *Journal of the American Statistical Association*, vol. 82, no. 398, 1987, pp. 371-386.

Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768-4777.

Mirjalili, Vahid, et al. “Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers.” *IEEE Access 7*, 2019, pp. 99735-99745.

Nagpal, Shruti, et al. “Deep learning for face recognition: Pride or prejudiced?” *arXiv preprint arXiv:1904.01219*, 2019.

Nori, Harsha, et al. “Interpretml: A unified framework for machine learning interpretability.” *arXiv preprint arXiv:1909.09223*, 2019.

Parde, Connor J., et al. “Face and image representation in deep cnn features.” *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017, pp. 673-680.

Petsiuk, Vitali, et al. “Rise: Randomized input sampling for explanation of black-box models.” *British Machine Vision Conference*, 2018.

Ribeiro, Marco Tulio, et al. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.

Rozsa, Andras, et al. “Facial attributes: Accuracy and adversarial robustness.” *Pattern Recognition Letters*, vol. 124, 2019, pp. 100-108.

Selvaraju, Ramprasaath R., et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626.

Selvaraju, Ramprasaath R., et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization.” *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 618-626.

Staniak, Mateusz, and Przemysław Biecek. “Explanations of Model Predictions with live and breakDown Packages.” *The R Journal*, vol. 10, no. 2, 2018, pp. 395-409.

Terhörst, Philipp, et al. “Suppressing gender and age in face templates using incremental variable elimination.” *2019 IEEE International Conference on Biometrics (ICB)*, 2019, pp. 1-8.

Williford, Jonathan R., et al. “Explainable face recognition, Springer.” 2020, pp. 248-263.

Yin, Bangjie, et al. “Towards interpretable face recognition.” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9348-9357.

Zeiler, Matthew D., and Rob Fergus. “Visualizing and understanding convolutional networks.” *European Conference on Computer Vision, Springer*, 2014.

Zhou, Bolei, et al. “Learning deep features for discriminative localization.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929.