ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Informatik
& Security

A Deep Dive into Deepfakes

What are the future opportunities and recommendations for organizations to prevent against deepfake-based attacks?

Diploma thesis

For attainment of the academic degree of

Diplom-Ingenieur

submitted by

Kathrin Schneller, BSc
51825242

in the
University course Information Security at St. Pölten University of Applied Sciences

Supervision
Advisors: FH-Prof. Dipl.-Ing. Mag. Marlies Temper, Bakk.
FH-Prof. Dr. Alexander Adrowitzer

St. Pölten,
12th of September 2023

(Signature author)

Declaration

I declare that to the best of my knowledge and belief

- This diploma thesis is written independently, without using any sources and aids other than those indicated and not made use of any other unauthorized assistance.

- I have not yet submitted this diploma thesis topic to an assessor in Austria or abroad for assessment or in any form as an examination paper.

- This work corresponds to the work assessed by the assessor.

The student / graduate grants the FH St. Pölten the right to use the diploma thesis for teaching and research activities and to advertise it (e.g., at project vernissage, in publications, on the homepage), whereby the graduate must be named as the author. Any commercial exploitation / use requires a further agreement between the student / graduate and the FH St. Pölten.

St. Pölten,
12th of September 2023

(Signature author)

Summary (German)

Deepfakes - eine Technologie, die nach unendlichen zukünftigen Möglichkeiten und Transformationen förmlich schreit. Das Aufkommen in den Medien, der Einsatz in vielen verschiedenen Anwendungsbereichen sowie die Entwicklung und Verbreitung verschiedener Tools für Deepfakes, haben ein neues Forschungsfeld im Bereich Deep Learning eröffnet. Obwohl neue Technologien den Vorteil haben, dass sie viele Möglichkeiten schaffen, bedeutet dies auch, dass die Risiken und Bedrohungen noch nicht in allen Bereichen erkannt oder sogar Gegenmaßnahmen für potenzielle Sicherheitsrisiken und Bedrohungen implementiert wurden.

Daher ist es wichtig, die Chancen und Risiken, insbesondere in den Bereichen Cyberkriminalität, Cyber Warfare, Bildung, Marketing, Medien und Gesellschaft in einem frühen Stadium der Risikobewertung einschätzen zu können. Somit können die negativen Auswirkungen gemindert und positive Risiken als Chancen für neue Möglichkeiten identifiziert werden.

Der Beitrag dieser wissenschaftlichen Arbeit ist es, die Probleme und den Einfluss von Deepfakes in Bezug auf die Bereiche Cyberkriminalität, Cyber Warfare, Bildung und Marketing aufzuzeigen. Dafür wird als Methodik eine Explorative Szenarioanalyse angewendet, um in Abständen von drei und fünf Jahren die verschiedenen Entwicklungen und zukünftigen Möglichkeiten von Deepfakes aufzuzeigen. In weiterer Folge werden daraus dann Maßnahmenempfehlungen im Umgang mit Deepfakes für Organisationen anhand eines Maßnahmenkatalogs erstellt.

Index Begriffe

Machine Learning, Deep Learning, Neuronale Netze, Deepfakes, Deepfake-Algorithmen, Fake News, Social Engineering, Phishing, Cyberkriminalität, Cyber Warfare, Bildung, Marketing, State-of-the-Art Analyse, Explorative Szenarioanalyse, Gegenmaßnahmen

Abstract

Deepfakes - a technology that calls out for endless future opportunities and transformations. The emergence in media, the use in many different application areas as well as the development and distribution of several tools for deepfakes opened a new field of research in deep learning. Although new technologies have the advantage of opening many possibilities, this also means that the risks and threats have not yet been recognized in all areas, or even that countermeasures for potential security risks and threats have been implemented. Therefore, it is important to be able to assess the potential risks and opportunities, especially in cybercrime, cyber warfare, education, marketing, the media and society at an early stage of risk assessment thus mitigating the negative impact and use positive risks as chances for new opportunities.

The contribution of this diploma thesis is to demonstrate the issues and impact of deepfakes in relation to cybercrime, cyber warfare, education and marketing. For this purpose, an explorative scenario analysis is used as scientific methodology to reveal the various developments and future possibilities of deepfakes at intervals of three and five years. Before the explorative scenario analysis, however, a state-of-the-art analysis is conducted to evaluate the current technical opportunities and technologies. This enables a precise elaboration of the individual future scenarios in the scenario analysis. Subsequently, recommendations for measures to deal with deepfakes for organizations will be drawn up based on a countermeasure catalogue.

Index Terms

Machine learning, deep learning, neural networks, deepfakes, deepfake algorithms, fake news, social engineering, phishing, cybercrime, cyber warfare, education, marketing, state-of-the-art analysis, explorative scenario analysis, countermeasures

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation for this work

In today's digital age, the prevalence of manipulated media has become a significant concern for society. Deepfakes, which are realistic but fake videos created using artificial intelligence and machine learning, have the potential to cause widespread harm by spreading false information or manipulating public opinion.

Therefore, education about deepfakes is crucial in understanding the implications of this arising technology. This research can shed light on the underlying technical aspects of deepfakes and their ability to deceive individuals. Furthermore, it can explore the ethical, social, and legal implications of deepfakes, and propose potential solutions to tackle these issues. It can also help in designing policies and regulations to mitigate the risks of deepfakes and educate the public about the threats and consequences of manipulated media.

The first encounter of the author with deepfake technology was a video on the author’s Instagram news feed. The video depicted a person playing golf, who appeared to be Tom Cruise. However, after reading the caption, it became clear that the video was not actually of Tom Cruise himself, but rather a product of deepfake technology. Immediately intrigued, the research on this emerging technology thus led to this diploma thesis – A Deep Dive into Deepfakes.

1.2 Problem Description

Due to the steadily increasing popularity and more and more emerging areas of application for deepfakes, not only are new fields of research being established, but also several threats emerge massively accompany and influence security as well as society. In the meantime, there are applications and software solutions with which deepfakes can be created in just a few steps without any need of technical know-how. Nevertheless, it takes some time and effort to generate a deepfake with the help of deep learning algorithms.

Because access to these technologies, videos or images is provided easily to public today, cybercriminals are not far away. They exploit deepfake technologies to an extent creating new attack methods and techniques to manipulate, influence or harm companies, public figures, or private individuals.

1.3 Scientific Question

The scientific challenges of this thesis are specified in the following research question:

„What are the future opportunities and recommendations for organizations to prevent against deepfake-based attacks?”

Thus, the contribution of this research work predominantly relates to the core areas of a detailed state-of-the-art analysis of deepfake technologies available today. During an explorative scenario analysis further focus is placed on the future of deepfakes in cybercrime, cyber warfare, education and marketing from an information security point of view. As result of this analyses, deepfake countermeasures and recommendations for organizations are developed.

1.4 Goals and Objectives

In the last few years deepfakes have emerged as a topic in IT security [1], still being relatively at the beginning of its real potential. Regarding information security, not only are current problems and challenges in connection with deepfakes be looked at more closely, but attention is paid to the impact in cybercrime including security attacks such as social engineering, phishing, blackmailing, remote identity fraud and cybercrime-as-a-service. When deepfake technology is combined and used in the context of this types of cyberattacks, the impact can be devastating for those affected.

With a view to deepfakes in cyber warfare, potential problems and future challenges combined with cybercrime activities are discussed further. What impact information security research fields like deepfake usage in quantum computing or military weaponry such as lethal autonomous weapons and information war tactics and techniques have on organizations and society, is also part of this thesis.

In the advertising industry, artificially created faces and people have been used in commercials for years, without the target groups realizing they were deceived. Threatening about this fact is that people seeing these commercials often cannot distinguish whether these people are real or a creation with the help of deep learning. Other areas in this segment are the strong personalisation and adaptation of advertising content to the needs and preferences of the target groups and its influence on customer satisfaction.

Moreover, the educational branch is also focused on. Here, the emphasis is on the impact of deepfakes on education, the exchange of information and the transfer of knowledge.

All mentioned areas focused on are part of the methodology approach, the explorative scenario analysis (further explanation can be found in the next chapter ***2. Methodology***) aiming to depict the potential future of deepfakes in the time frames of three and five years. For this purpose, scenario funnels are created, illustrating the best-case, worst-case and trend scenario for a specific future event. These future events are then interpreted in the last section of the scenario analysis. An additional goal of this research work is the development of a deepfake countermeasure and recommendation catalogue for organizations in handling deepfakes.

2. Methodology

This section describes all chapters of this thesis starting with the introduction of the scientific method, the explorative scenario analysis.

2.1 Explorative Scenario Analysis

As briefly mentioned in the previous section the scientific method used is an explorative scenario analysis. This methodology is extensively described in [2]. Since the previously stated research question aims at future perspectives of deepfakes a method centred around the predictable future seems suitable.

The future scenarios are illustrated as funnels and are defined as an area between three archetypes (the best-case, trend and worst-case) projected from the present into the future detailing scenarios with confounding effects along the scenario's path. As deepfakes especially in cybercrime, cyber warfare, education and marketing are trend driven a methodology centred around trends seems up to the task. Also, this novel perspective on deepfakes provides a new approach allowing the exploration of possible paths driven by extrapolating the present usage of the technology by applying domain knowledge around AI for realistic expectations.

After elaborating on the preliminary knowledge, the phases as outlined in [2] are executed in section ***5. Explorative Scenario Analysis of Deepfakes in the Future***. In the following subsections, the thesis is outlined.

2.2 Section: Introduction

The introduction is a brief overview of the motivation for this work, the problem description and scientific challenge as well as the pursued goals and objectives within this thesis.

2.3 Section: Literature Analysis

Secondly, this section describes basic knowledge and background information of deepfakes to better understand its functionality, application areas, problems, challenges, advantages and disadvantages. For example, topics like machine learning, deep learning and deepfake related neuronal networks are going to be further discussed.

2.4 Section: Preliminaries

Furthermore, the preliminaries are building the next big section dealing with specific information and application areas providing the necessary background information for the subsequent scenario and state-of-the-art analysis. Here, besides general methods and approaches, further considerations such as deepfake detection and verification, technical security implementations and approaches as well as deepfake-influenced areas in cybercrime and cyber warfare also take a huge part in this chapter.

2.5 Section: Explorative Scenario Analysis of Deepfakes

Based on the previous chapter the next part outlines all necessary steps of the explorative scenario analysis of deepfakes. The analysis is getting divided into six main phases – building the scenario environment, choosing the most important key factors and producing an analysis and classification out of it thus generating scenario funnels with potential future scenarios concerning deepfakes.

After the creation of several scenario funnels in intervals of three and five years, the next section describes the countermeasures derived from the future scenarios for dealing with deepfakes in organizations.

2.6 Section: Deepfake Countermeasures and Recommendations for Organizations

The impact of deepfakes concerning organizations sums up this section. The countermeasures and recommendations presented in this chapter are supported by the results of the state-of-the-art analysis and explorative scenario analysis to provide the expected scientific results based on the scientific question of this thesis.

2.7 Section: Conclusion and Future Prospects

Finally, the conclusion and future prospects chapter summarizes the most important findings, considers and draws conclusions on the impact of deepfakes in cybercrime, cyber warfare, education and marketing based on the deepfake potential in the future.

3. Literature Analysis

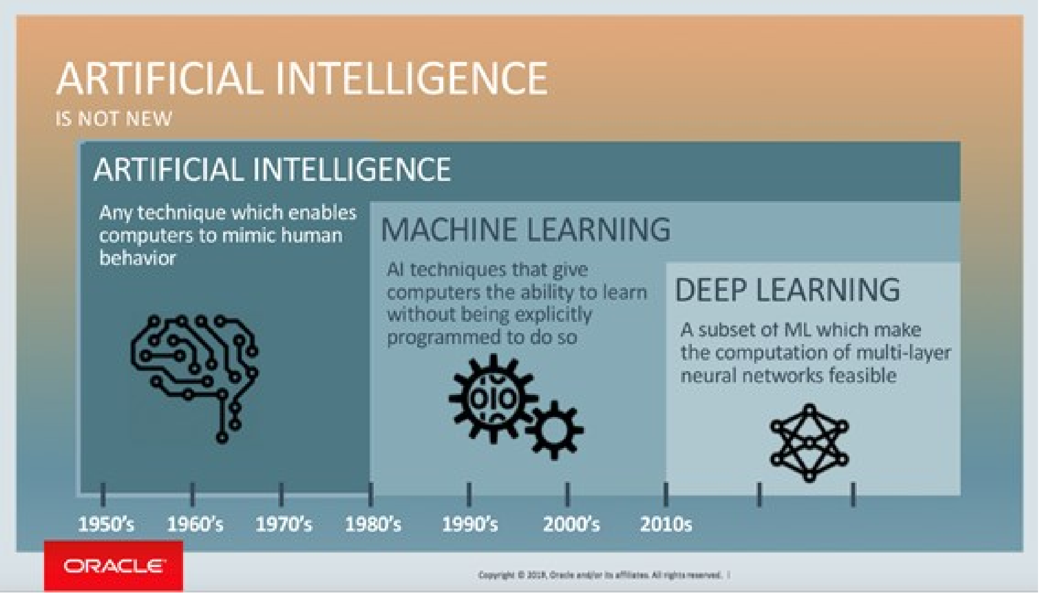
Deepfakes, are a very advanced form of Machine Learning, hence this literature analysis traces the evolution from Machine Learning including definitions and used algorithms to Deep Learning, followed by Autoencoders and Convolutional Neural Networks, Generative Adversarial Networks, and finally deepfakes.

3.1 Machine Learning

The first chapter of the literature review addresses the overriding topic of machine learning. The next chapters will cover a general definition, individual advantages and disadvantages, areas of application, modes of operation, algorithms and further subcategories, e.g., Deep Learning.

3.1.1 Definition

Machine Learning (ML) is a term that has become increasingly popular in recent decades. While learning itself is a natural human behaviour, computers cannot learn as humans do. Since humans learn from experience, we do not need different kinds of algorithms or a huge amount of data to make decisions on our own. But it is different with computers. Therefore, Machine Learning describes techniques of Artificial Intelligence (AI) whereby a computer has the ability to learn, make its own experiences and think on its own without being programmed to do so explicitly. It is particularly important that the computer can adapt and improve actions and calculations it performs - in other words - that it can learn from the environment and its own mistakes. If this is not the case, the accuracy of the actions performed could be affected. [4] [5] [6]



ARTIFICIAL INTELLIGENCE
IS NOT NEW

ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior

MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so

DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Figure 1: Deep Learning History [6]

As pictured in ***Figure 1*** above back in the 1950s, the first techniques were already being developed on how a computer could simulate human behaviour - basically the birth of Artificial Intelligence. Later, from the 1980s onwards, computers were able to learn more from their behaviour. The era of Machine Learning was born. It was not before the 2010s that Machine Learning gave rise to Multi-layer Neural Networks, which can be subsumed under Deep Learning. [5] [6]

In the following sections, the objectives and various areas of application of Machine Learning will be outlined, individual different learning approaches is explained and Deep Learning will be discussed more closely.

3.1.2 Objectives, Types and Applications

Nowadays, the Machine Learning approach is used in many different areas of application and forms the basis for various technologies. Due to [5], the following main objectives are pursued when using Machine Learning:

- Machine Learning includes the **recognition and detection of patterns** and structures in data based on examples.
- Machine Learning can be used to **forecast future results** (forecast or prediction) thus to outline and describe the detected patterns and structures.
- Machine Learning can help solve challenges or **support solutions** to highly complex and data-overloaded challenges, as well as for problems where no solution could be found yet.
- Due to the **self-tuning property** of ML-algorithms, Machine Learning supports the challenges with a fluctuating and permanently changing environment.

Machine Learning can be divided into the following types and application areas as shown in the example figure below:

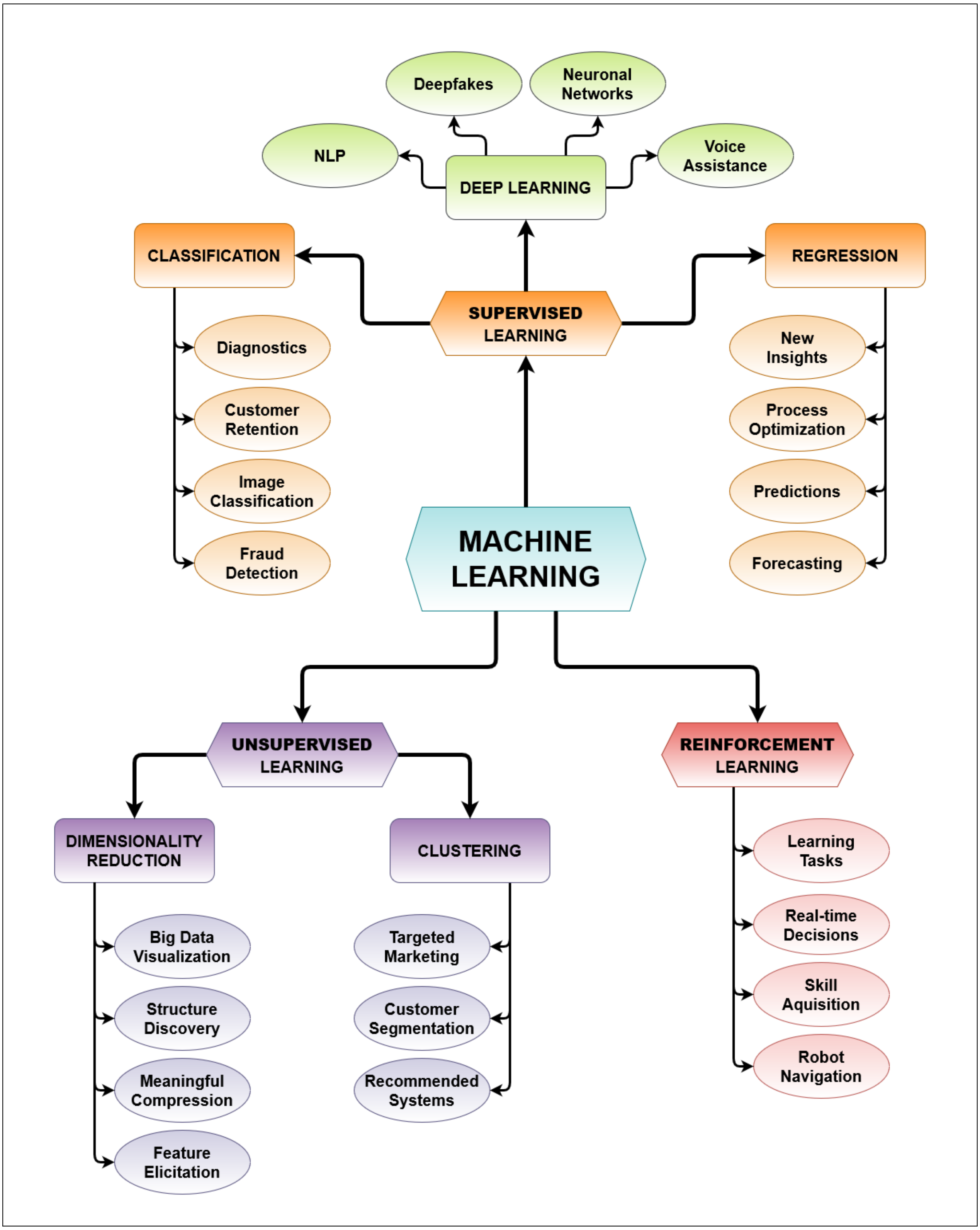


Figure 2: Machine Learning categories [5] [7]

Turning to ***Figure 2***, the individual Machine Learning categories are now described in more detail:

Machine Learning differentiates between **three basic learning categories:** Supervised Learning, Unsupervised Learning and Reinforcement Learning. In addition to the supervised and unsupervised approach, there is also a combination of the two, the semi-supervised method.

In chapter ***3.2 Supervised Learning*** the first mentioned category **Supervised Learning** is discussed in detail and described more precisely.

The second main category of Machine Learning is **Unsupervised Learning**. The two techniques used here are Dimensionality Reduction and Clustering. While **Dimensionality Reduction** includes some techniques such as Big Data Visualization, Structure Discovery, Meaningful Compression or Feature Elicitation, it aims to filter huge amounts of high-dimensional data with specific algorithms and extract only the most relevant information for understanding the dependencies between the data points. [5] [8] **Clustering** methods are used when an indeterminate number of groups or categories are formed from existing data sets, which differ or are similar, for example, in certain feature properties. These methods are often used in Targeted Marketing or Customer Segmentation. [5] [9]

Another category that arises from Machine Learning is **Reinforcement Learning**. Here an agent interacts with the environment to learn how to assign situations to the proper actions. To achieve that, the learner must choose which actions bring the most reward. This is done with reward and penalty functions. So, not only it is to decide which next situation will bring the highest reward, but also the one after that and thus the reward that follows. The main features of this learning method are the search for the best action with trial and error and delayed reward. Moreover, Reinforcement Learning is not only used for learning specific tasks, but also for real-time decisions, Skill Acquisition or Robot Navigation. [10]

Besides the several different types of machine learning and the massive variety of learning algorithms, the focus of this thesis will be on Supervised Learning and **Deep Learning**, because deepfakes are using Deep Learning algorithms as base and part of their functionality. An explicit explanation can be found under the chapter ***3.3 Deep Learning***.

3.1.3 Challenges

Besides all sorts of application fields, Machine Learning brings with it some challenges, which can be summarized into following topics:

Data Quantity

Even for simple algorithms or decisions, such as whether the animal is a cat or not, often a huge amount of data is needed to train the respective model. [5]

Data Quality

It is not only the quantity of data available to the machine model for training that is important, but also the quality of the data. This usually manifests itself in noise, outliers, errors or "impurities" during the data cleansing process. [5] In the Generic Machine Learning Model, the sub-process of Data Collection and Data Preparation also plays an important role. [4]

Data Representation

Non-representative data can be a problem because sometimes sampling noise or sampling bias plays a role in training a model. [5] Assumed deviations from the true data set are called data noise and always refers to deviations in the data, for example in images. Another noise category is label noise, which points to deviations in the labels itself. [11] In terms of sampling noise, the misinterpretation of certain tasks or the over-trending of the data in one direction by the learning algorithm can influence the desired model result. [12] A Machine Learning bias in terms of definition describes an algorithm not producing the desired outcome based on incorrect assumptions and decisions made during the learning process. [13]

Complexity

For a well-trained model it is essential that the right amount of complexity is applied to the training for the learning algorithm. Otherwise, the model can be too simple for a complex problem or vice-versa. [5]

Overfitting

If the model has too much complexity, a high variance and a very low bias when training the learning model, this is called overfitting. [14] Put simply, Overfitting means that the model is overfitted to the data, that the error rate is 0%. Consequently, noise and biases are also included in the model, which in turn leads to a decrease in the quality of the model. [5]

Underfitting

If the model has too little complexity, low variance and a high bias when training the learning model, then this is referred to as underfitting. [14] In contrast to Overfitting, the problem with Underfitting is that the model contains too little data or is too simple and thus cannot properly represent the actual problem or task. [5]

Feature Selection and Feature Extraction

If the input data is already poorly pre-processed or contains features that are not relevant to the actual result, this will be reflected in the output data. [5]

3.2 Supervised Learning

In general, depending on the supervision during training the ML-systems, they can be divided into several different learning systems, such as Supervised Learning. It is often interpreted as learning from given examples, whereby a distinction is made between the training set and the test set for training the learning model. Based on the training set, the algorithm is trained and learns how to behave and follow certain solution approaches. During training, the algorithm compares the output of the labels with those provided as input. [4]

In general, Supervised Learning is divided into two subtypes: classification and regression. While **Classification** determines the best suited class of a feature characteristic out of a known set of category classes, **Regression** predicts the value of a characteristic showing the functional relationship between all these characteristics. Typical application areas for Classification are Diagnostics, Customer Retention, Image Classification and Fraud Detection. Forecasting or making predictions, process optimization and building new insights can be achieved with Regression. [5]

3.2.1 Supervised Learning Approach

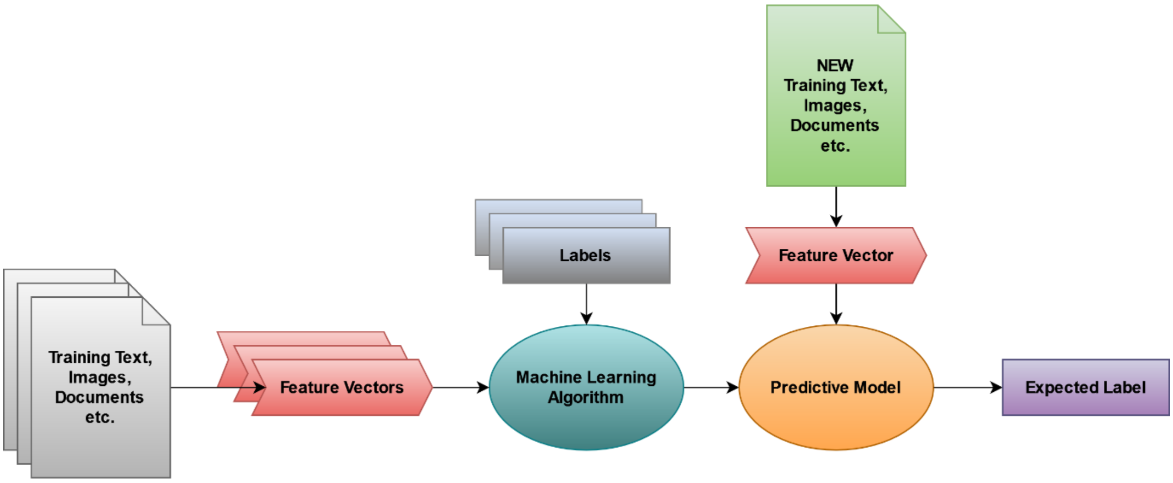


Figure 3: Supervised Learning Approach [4]

As described in ***Figure 3***, in supervised learning, the input for the machine learning algorithm is firstly the training set. This can be documents, texts, words, pictures or videos, however, everything relevant for the later ML-model. For the training of the algorithm **feature vectors** and **labels** are necessary.

Data Labeling is used to identify raw data handled during the training process in order to produce accurate output information. [15] Moreover, to provide more context to the information meaningful and informative labels play a significant role so that the algorithm can learn from it. [16] For instance, if a model is to be created for the recognition of whether a cat is present in a picture or not, an important piece of information as a label would be whether a cat is present in the picture or not. As a result, labels include the desired solutions for the ML-model.

Features are the representation of properties or objects numerically, which are then further processed into vectors, termed **feature vectors**, demonstrating a group of features that describe an object or a property. Some reasons why further processing into Feature Vectors is so important is that working with a huge amount of data usually contains numerous redundant information.

In order to save performance or the manual checking of data by a data scientist, feature vectors are used as standard data representation when training Machine Learning systems. [17]
When training the algorithm using the selected labels and feature vectors is finished, a **predictive model** is generated with the test set of data as input. The result at this point is the expected label, which, if we recall our previous example of recognizing a cat, distinguishes cats from other people or animals. Finally, it is important to ensure that no overfitting or underfitting of the model takes place. Otherwise, this may reduce the quality of the model.

3.2.2 Supervised Learning Algorithms

The most widely used and essential supervised learning algorithms according to [5] [18] are:

K-Nearest Neighbors (KNN)

This algorithm is commonly used for Classification or Regression approaches. When given new instances the K-Nearest Neighbors algorithm recognizes the nearest neighbours (data points) while using similarity or distance functions, e.g., through the Euclidean distances, and decides then to which class the new stance belongs. The basic idea of this algorithm consists in the fact that with a high likelihood all unknown data points or new instances are in the same class as their neighbors. [4] [5]

Linear Regression

The Regression analysis focuses predominantly on the relationship of the data points to each other. Thereby, a dependent and independent variable is always assumed, meaning that the dependent variable can be influenced, while this is not the case with the independent variable. A good example could be a music festival, where the independent factor would be the sunshine, while the dependent variable would be the festival visitors. As the name implies, Linear Regression establishes a linear relationship using a trend line between the variables. Using the trend line, the most accurate predictions of the true values of all data points are made to reduce the distance differences between the data points to a minimum. [4] [5]

Logistic Regression

Another type of regression is the Logistic Regression. This type of regression works similarly to the linear variant, but instead of a linear function, an S-function is used. A threshold value defines the membership of a data point to a certain category or class. Furthermore, the probability of the affiliation is mapped by the target variable with the values between 0 and 1. [5]

Support Vector Machines (SVMs)

This type of Supervised Learning is suitable for Regression, linear and nonlinear Classification, as well as particularly for Outlier Detection. [4] SVMs provide an advantage over other Supervised Learning methods because they offer a subset of the original dataset as support vectors during the learning phase of the algorithm. Every set of support vectors signifies a specific Classification task. The motivation behind SVM is to separate multiple classes in the training set with an area that maximizes the distance between them. Thus, maximizing the generalization ability of a model becomes possible. Furthermore, the goal of structural risk minimization is always pursued, which allows the minimization of a threshold value for the generalization error of a model, instead of minimizing the mean square error in the training data set as in most other methods of error and risk minimization. [19]

Decision Trees (DTs)

The basic principle of Decision Trees is to query attributes until a decision is made to classify these attributes. The starting point is always the root node, which represents the main issue or task. Starting from the root node, the attributes are compared with the possible class attributes using a "divide and conquer" approach. In most cases these are constant values known as decision nodes. Alternatively, two attribute values can be compared or even a function from one or more classes may be used.

Besides root and decision nodes, there are also "leafs", which reference either to a class, set of multiple classes, or even to a probability distribution. In conclusion, at the end of the classification, starting from the root node, several branches with the respective decision nodes or leafs should have been formed. [5] One huge advantage of Decision Trees over other methods is that they require little data preparation to show their desired performance. [18]

Random Forests (RFs)

A random forest consists of several decision trees and can be used for classifications according to the majority decision principle as well as for regression in search of the average decision. In addition, when building, a bootstrap aggregating algorithm, or bagging for short, is applied. [5] Bagging describes sampling with substitutions or replacements during the training process, so, when performing sampling without substitutions it is described as pasting. Both methods allow the training entities to be sampled multiple times with different predictors, but only bagging also permits sampling only with the identical predictor. [18] With the bagging approach the final output of a Random Forest is build out of several Decision Trees with a random subset of data. The combination of all trees forms the final Decision Tree. [4]

Neural Networks (NNs)

Neural Networks pursue the concept of working with the help of three different layers: the input layer, the hidden layer and the output layer. To perform parallel processing, the network weights the interconnections and learns from their adjustments. Furthermore, there is, once again, the distinction between Supervised, Unsupervised and Reinforcement Neural Networks. [4]

3.3 Deep Learning

Deep Learning is an important subcategory of Machine Learning mechanisms. In most cases, the term "Deep Learning" is predominantly used for Deep Neural Networks, as several stacked hidden layers are in use. The objective of Deep Neural Networks is thus to make the most accurate prediction of a desired output during training by correctly figuring out the individual weights per layer. [20] The next chapters will mainly focus on the original theory and the functionality of Neural Networks (Linear Threshold Unit and Perceptron, Deep Neural Networks) together with Deep Neural Networks and Deep Learning Algorithms.

3.3.1 Linear Threshold Unit (LTU)

The theory behind the Multilayer Perceptron, or Backpropagation, has been around since 1957, when Frank Rosenblatt introduced the concept of Linear Threshold Units (LTU). Nowadays, LTUs form the basis of a Perceptron, which consists of only one layer of an LTU. [5] [21] The following figure (***Figure 4***) explains the functionality of it in more detail:

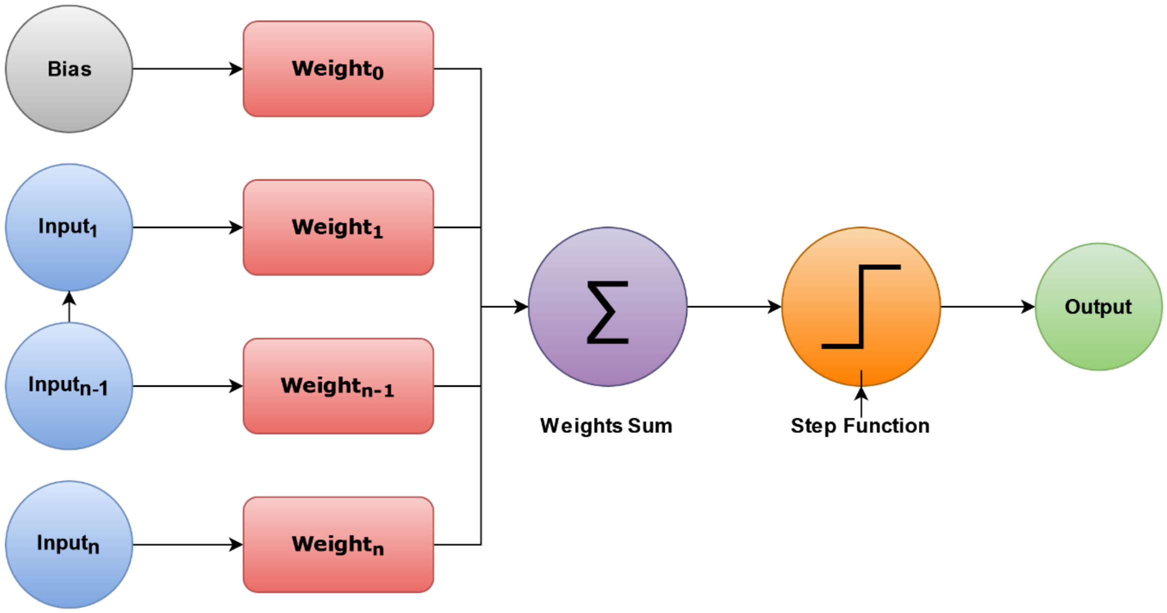


Figure 4: Linear Threshold Unit [5] [22]

In general, an LTU maps a vector of input data to a binary output. As illustrated in the figure above, these input data are assigned randomized weights at the beginning. In addition to the inputs, there is a certain constant value of 1 – the bias - which can be thought of as separate neuron helping to ensure that the later trained Neural Network is not generating a potentially higher error rate. [21] [23]

In the next step, the sum of inputs times the respective weights are calculated and transferred to a step or activation function. As soon as the threshold value is exceeded, the input is forwarded to the next layer - the output layer. This method of operation is particularly useful for Linear Binary Classification, but as soon as more complex problems are involved, several layers of perceptrons (Multilayer Perceptrons) are recommended. [21] When talking about training a Neural Network, it is always referred to as learning all the weights associated with all the edges of the nodes. [24]

3.3.2 The Perceptron

According to Rosenblatt [22], the basic concept of a perceptron consists of three main components: the Sensory System (S-system), the Association System (A-system) and the Response System (R-system).

Furthermore, Rosenblatt also described the **S-system** as a set of nodes connected to several entities from the A-system, which transmits positive or negative impulses to the R-system as soon as the S-point is reached. Within this node point system, the individual nodes are either positively or negatively connected to each other A-system units. [22]

The **A-system** is responsible of the switching functions between input (S-system) and output (R-system). This means that each A-unit receives impulses from the S-points and transmits them to the respective response units (R-units). A-units are characterised by a fixed parameter, the threshold value, the sum of all input pulses and the output result.
In this way, the output value changes depending on which input signals have been transmitted to the A-units and serves as the system's memory counter or register. [22]

This brings up the last of the three main components of a perceptron: the **R-system**. The R-units within the Response System are triggered as soon as the received input signals have reached or exceeded a certain threshold value. In besides displaying a result, the R-system also has the function of sending feedback signals back to the respective A-units with the help of the R-units. This feedback consequently ensures that the connections to mutually exclusive R- and A-units are cut or blocked. This concept also prevents several mutually exclusive sets from being triggered at the same time, i.e., if a signal arrives at two R-units at the same time, the unit with the greater mean of all inputs wins and the other signal is blocked. [22]

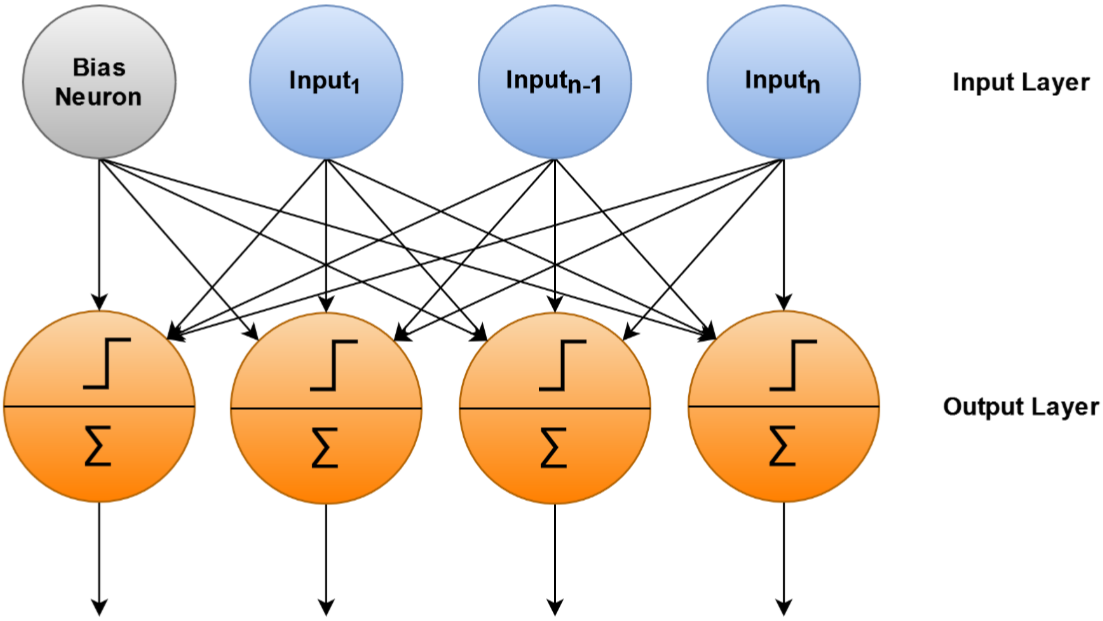


Figure 5: Perceptron [21] [22]

3.3.3 From Multilayer Perceptrons to Artificial Neural Networks

It was only a few years after Rosenblatt published the theory behind the perceptron that Marvin Minksy and Seymour Papert realized that a single perceptron was incapable of executing an XOR operation. The solution to this problem resides in the number of layers: Multilayer Perceptrons are one of the most used Neural Net Architectures these days. Another term for Multilayer Perceptron is Feed Forward Neural Network or Artificial Neural Net (ANN). Once several hidden layers are in usage, the term Deep Neural Net (DNN) is used. [21]

The main difference between an Artificial Neural Network and a Multilayer Perceptron is the function that triggers the next actions. ANNs use non-linear activation functions, such as the sigmoid function, ReLU (Rectified Linear Unit), Leaky ReLU or the inverted tangents (tanh) function. Multilayer Perceptrons, on the other hand, rely on the step function. [21] [24] [25]

3.3.4 Deep Neural Networks (DNN)

Continuing from the previous chapter, the following aspects can explain the functioning of Deep Neural Networks in more detail: The initial base is a stack of layers, where each layer contains a certain number of nodes, which are connected to the nodes of the previous layer by different weights. In general, the three different layer categories are distinguished as the input layer, the hidden layer and the output layer. [20]

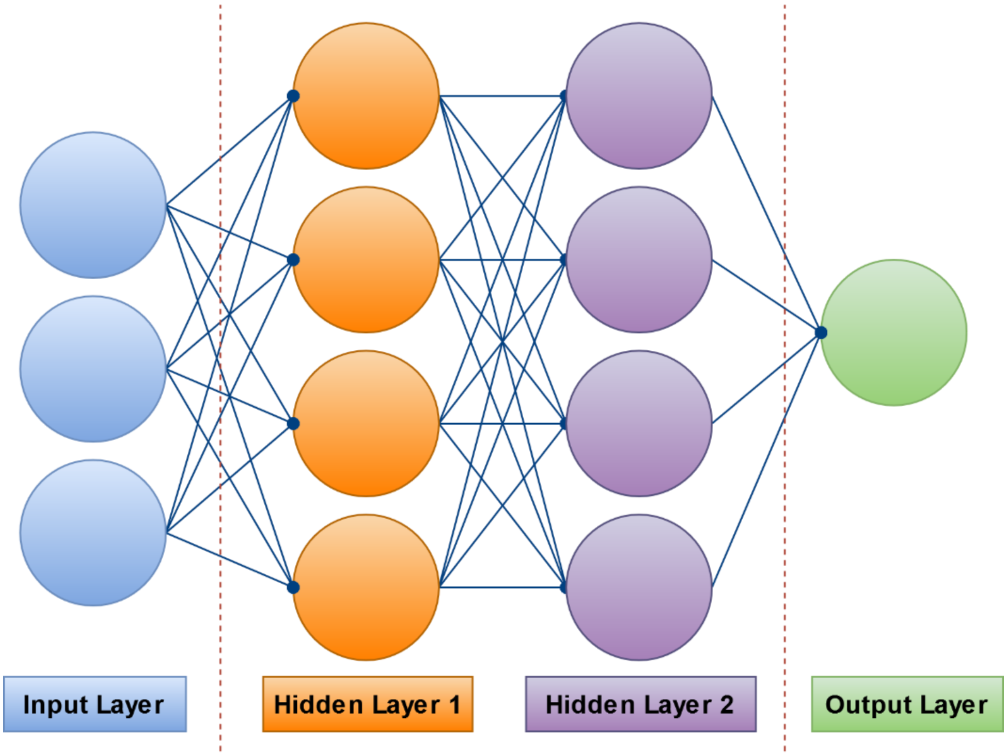


Figure 6: Feed Forward Artificial Neural Network Architecture [24]

Figure 6 above shows the structure of an Artificial Neural Network (ANN) or a Deep Learning Network. According to the definition, it is considered "deep" as soon as an increased amount of hidden layers are involved in the creation of the network. [24]

Furthermore, the following distinction can be made between Neural Network learning methods:

In **Supervised Neural Networks**, the two input and output layers are interpreted as training data, so that the hidden layer can be used to implement adjustments to the weighting to obtain the most accurate results possible. Conversely, when no output data is provided to the Neural Network, it is referred to as **Unsupervised Neural Networks**. In this case, the network tries to identify a correlation or structure in the input data and thus to group or classify them. If new instances are added to the input data, the neural network recognizes the features and can classify them according to the similarities per group. Another type of NNs is **Reinforcement Neural Networks**. Here, the network learns from the decisions previously made by using reward and penalty functions. If the output is produced correctly, the weighted connection is strengthened each time, while it is weakened if the results are incorrect. [4]

3.3.5 Neural Network Training Process

With the intention of generating a reliable Neural Network, the model must be trained beforehand as following [24]:

- **Step 1:**
  The first step involves initialising the nodes of the input layer with different randomized weights. For this, some own initialisation methods are used.

- **Step 2:**
  The next step is to apply a Feed Forward or Multilayer Perceptron method to each training instance using the current weight values, thereby estimating the output from left to right.

- **Step 3:**
  During the third stage of training, a loss function compares the final output with the target value for determining the current error rate.

- **Step 4:**
  The last training step involves Backpropagation. This technique works with the "Back Pass" variant, which means distributing the error to every node from right to left. Furthermore, the individual weight distributions are determined based on the error rate and adjusted with the help of the gradient descent. The error gradients are retraced starting with the last of the hidden layers.

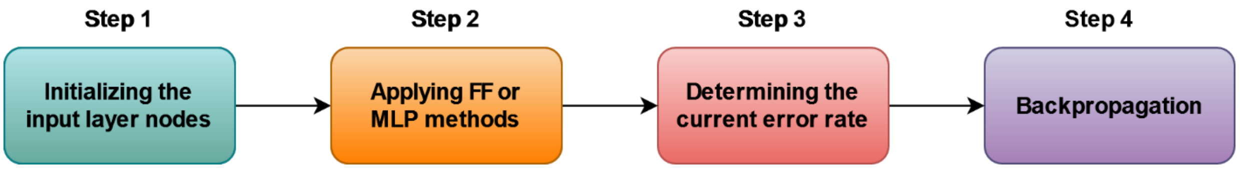

Step 1
Initializing the input layer nodes
Step 2
Applying FF or MLP methods
Step 3
Determining the current error rate
Step 4
Backpropagation

Figure 7: Neural Network training process [24] [26]

3.3.6 Gradient Descent

One of the most important optimisation algorithms in Machine Learning is the Gradient Descent. The objective pursued in this case is to adjust the weights when training a model in a way that keeps the cost function as minimal as possible. In general, cost functions are applied in Machine Learning to measure how well or poorly the respective model performs. The cost function estimates the relation between the X and Y value and then measures the deviation of the value from the Machine Learning model and the actual predicted value. [18] [27]

With Gradient Descent, randomised values are initialised at the beginning, the local gradient is measured and then the gradient is reduced in small steps. Once it reaches zero, the minimum and the aim of decreasing the cost function is accomplished. [18] The following figure (***Figure 8***) illustrates this algorithm graphically:

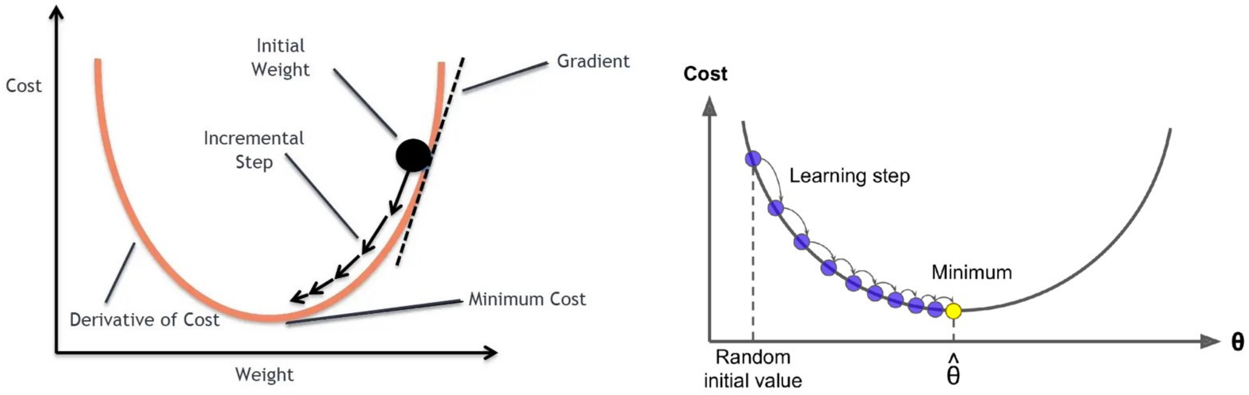


Figure 8: Gradient Descent Algorithm [18] [28]

Additionally, it is important to consider the size of the adjusting steps, i.e., the learning rate, and to choose them wisely. On the one hand, if they are too small, the algorithm takes longer to converge and therefore needs more time. On the other hand, if the learning rate is chosen too large, the problem can arise that the values jump back and forth on the curve and the algorithm potentially diverges ending up with a worse result than before. [18]

Another obstacle to consider when using Gradient Descent is that not all cost functions look like a bowl - some are not continuous, have ridges or plateaus - which make it difficult to converge to a minimum. More efficient than the local minimum is the global minimum. This is guaranteed to be reached if the training duration takes long enough and the learning rate is not too high. ***Figure 9*** below demonstrates the connection between the local and the global minimum: [18]

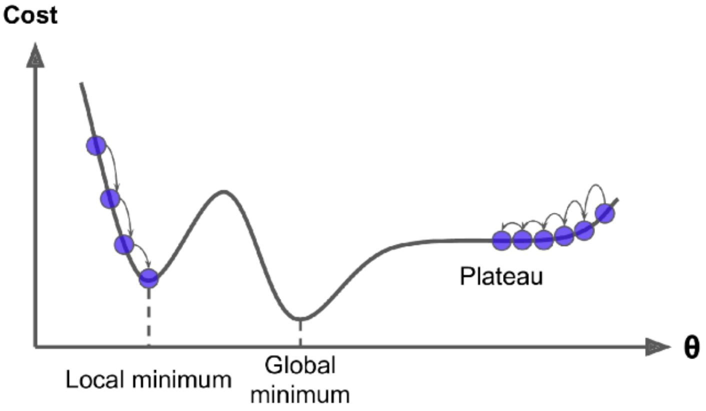


Figure 9: Gradient Descent - Global and local minimum [18]

Moreover, there exist further approaches of how the Gradient Descent can be implemented:

Stochastic Gradient Descent (SGD)

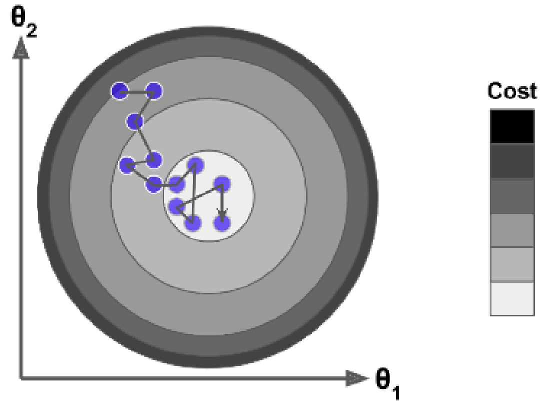


Figure 10: Stochastic Gradient Descent [16]

In this variant, one training instance after the other is passed on to the Neural Network and the weights of each layer are adjusted according to the calculated gradient. Consequently, if the complete training data set consists of 100 instances, the gradient of the complete Neural Network is adjusted 100 times. [28]

Figure 10 shows the convergence curve for the Stochastic Gradient Descent method.

Batch Gradient Descent (BGD)

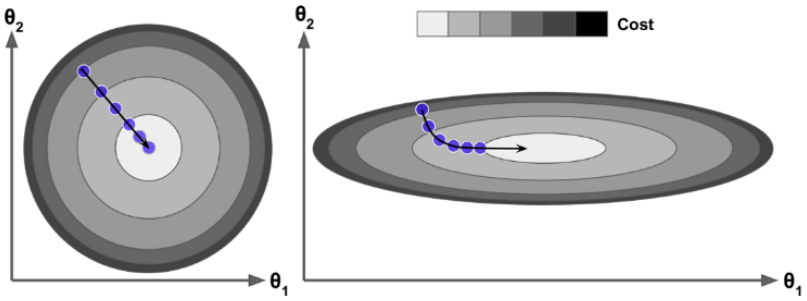


Figure 11: Batch Gradient Descent [16]

This method works similarly to the stochastic variant, but the weights of the Neural Network are adjusted after computing the loss of all training instances. For example, if the training data set consists of 100 instances, the gradient of the entire Neural Network is adjusted only once. [28] ***Figure 11*** shows the convergence curve for the Batch Gradient Descent technique.

Mini-batch Gradient Descent

The Mini-batch Gradient Descent forms the golden middle between the other two methods. The training set is split into several groups, known as batches. Each batch contains several training instances and for each iteration only one training instance is passed on to the Neural Network, but it is updated using the average of all calculated losses per batch. Thus, with 100 training data instances divided into four batches of 25 instances each, the weights of the Neural Network are adjusted four times. [28]

3.3.7 Deep Learning Algorithms

The most well-known Deep Learning algorithms include Deep Boltzmann Machines (DBM), Deep Belief Networks (DBN), Convolutional Neural Networks (CNN) and Stacked Auto-Encoders. Furthermore, the category of Ensemble algorithms includes Random Forests (RF), Gradient Boosting Machines (GBM), Boosting, Bootstrapped Aggregation (Bagging), Adaptive Boosting (AdaBoost), Stacked Generalization (Blending) and Gradient Boosted Regression Trees (GBRT). Algorithms that primarily belong to the field of Neural Networks include Radial Basis Function Network (RBFN), Perceptron, Backpropagation and the Hopfield Network. [5] [18]

3.4 Deepfake relevant Neural Networks

The subsequent content mainly concentrates on important background knowledge about the functioning of Deepfakes. In particular, Convolutional Neural Networks (ConvNets or CNN), Autoencoders and Generative Adversarial Networks (GANs) are examined in more detail.

3.4.1 Autoencoders

The first relevant Feed Forward Neural Network (FNN) architecture to understand is Autoencoders. They use unlabeled data and are therefore an unsupervised type of ANN. Their mode of operation is to compress the input to a lower dimensional representation vector and then rebuild it to the respective output vector. This process is called Latent-space Representation. They are used primarily for Dimensionally Reduction algorithms and are trained in the same way as ANNs through Backpropagation. [18] [20] [29]

Autoencoders contain three different parts: the Encoder, the Decoder and the Code. To create one, you need an encoding method, a decoding method and a loss function to compare the results with the desired output. To prevent the Autoencoder from directly rendering the input as output, the size of the Latent Representation can be restricted, for example, or noise may be added to the inputs. This guarantees the Autoencoder to try learning efficient ways of representing the data appropriately. [18]

The graphic below (***Figure 12***) illustrates the structure of an Autoencoder:

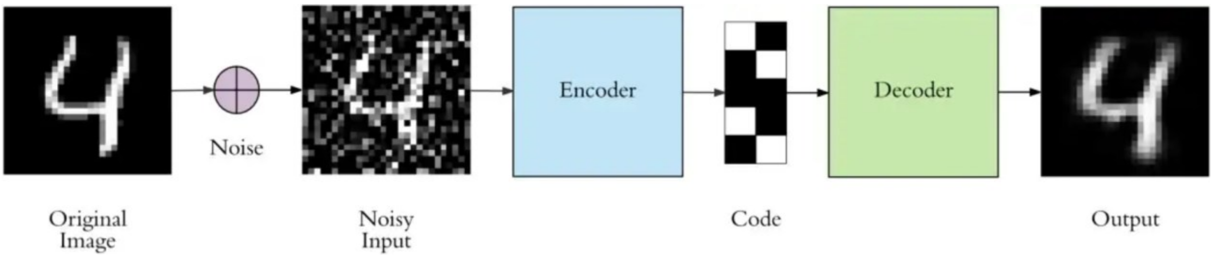


Figure 12: Autoencoder mode of operation [29]

One of the advantages of Autoencoders are that the data used for their learning can be raw and unprocessed. However, they are not suitable for lossless compression as they only approximate the output as closely as possible. Furthermore, they are data-specific, meaning for example, that an Autoencoder for compressing wildlife images cannot be employed for compressing handwriting. [29]

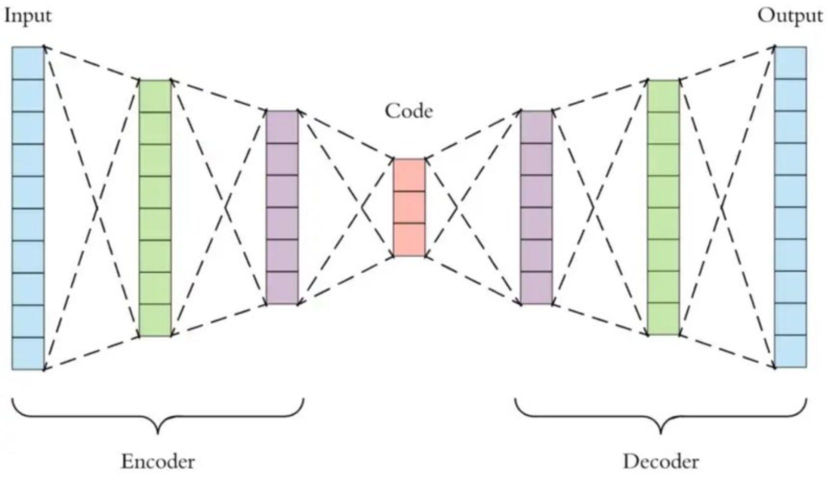


Figure 13: Autoencoders Architecture [29]

The graphic ***Figure 13*** above provides a detailed view of the Autoencoder architecture. The Encoder and Decoder are both complete Feed Forward Neural Networks, while the Code is represented by a layer of an ANN with the dimension of our choice. The number of nodes within the Code layer is a hyperparameter, which is set before training the Autoencoder. There are three other relevant hyperparameters: the code size, the number of layers and the loss function. [29]

Stacked Autoencoders

If Autoencoders with several hidden layers are employed, these are referred to as Stacked or Deep Autoencoders. This offers the advantage of enabling to learn more complex data and codes. Autoencoders can be trained with high performance, which means that overfitting, as with Machine Learning Models, should be taken into consideration. [18]

3.4.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) arose in the research of the visual cortex in our brain and are already used for recognizing images since the 1980s. Nowadays, they are one of the most widely implemented Deep Learning technologies. For example, CNNs are used in image search, self-driving cars, Natural Language Processing, as well as Voice Recognition. [18] In 2012, the interest in this method of learning began to increase further, as the CNN architecture AlexNet was created from eight layers to ResNet, which already has 152 layers. [30]

A major advantage of Convolutional Neural Networks compared to other NN variants is being able to recognizing all essential features without any human supervision. Another benefit is their computational efficiency through the use of certain pooling and convolution operations along with parameter sharing. [30]

As with ANNs, the fundamental structure includes fully connected layers and sigmoid activation functions. In addition, two new components are introduced here: the convolutional layer and pooling layer. [18] The pooling and convolution operations perform the feature extraction, while the ANN layers are responsible for the classification. [30]

The general structure of a CNN is illustrated in the following figure (***Figure 14***):

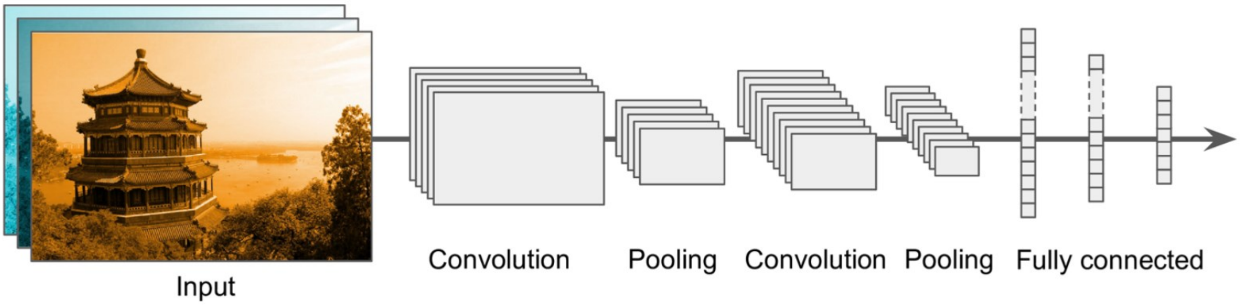


Figure 14: CNN Architecture [18]

First of all, an arbitrary image is selected as input with which several iterations of convolutional and pooling operations are executed. These are followed by a series of fully connected ANN layers (FC). The following subtopics explain these individual elements in more detail:

Convolutional Layer

Convolution is a mathematical operation that combines two sets of data. The input data is represented as a matrix grid and the convolution is performed by applying a convolution filter, also called kernel, over the input grid. The result is then a Feature Map. Thus, one kernel is generated per convolutional layer through the convolution operation. [30] By connecting each Convolutional Layer to only a subset of the data of each layer, it is simpler to target only a small subset of low-level features and then assemble high-level features in the next hidden convolutional layers. [18]

The hierarchical structure of CNNs (***Figure 15***) and the representation of the input matrix and kernels (***Figure 16***) can be imagined as in the following graphics:

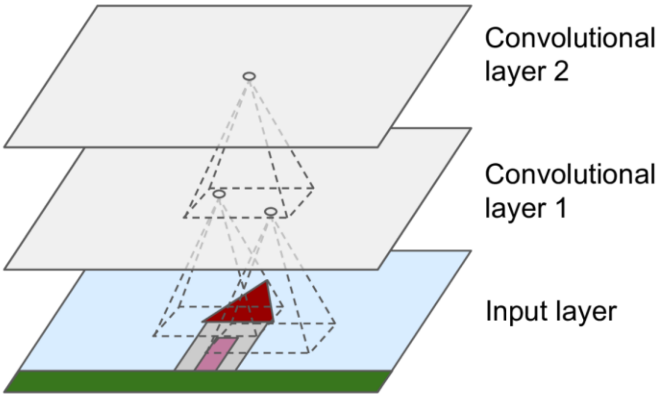


Figure 15: CNN Hierarchical Structure [18]

Figure 15 demonstrates how the convolutional layers on top of each other only have access to the receptive field of the input data (the image) and are stacked over each other. [18]

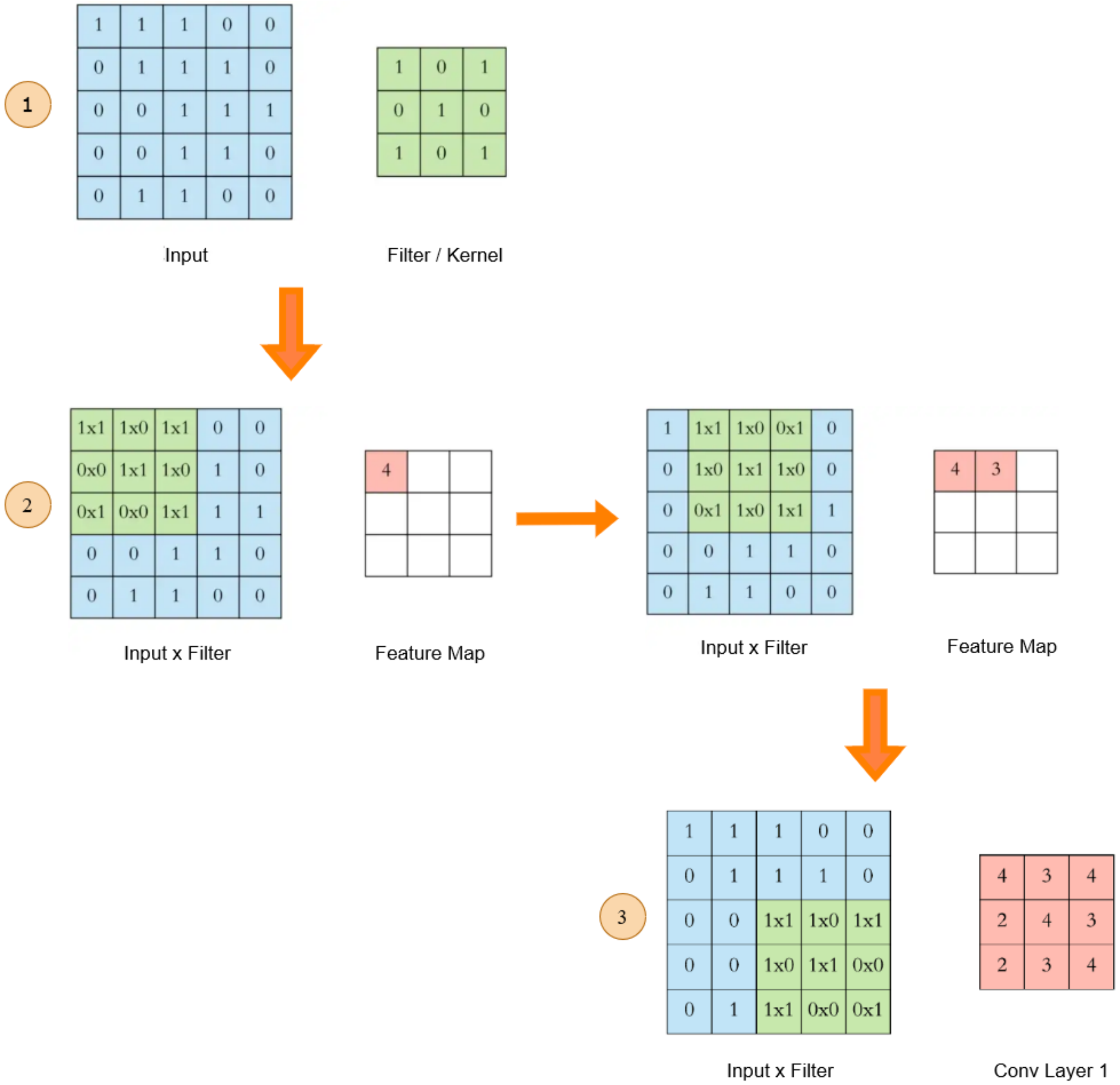


Figure 16: CNN Convolution Operation in 2D [30]

In the first step, according to ***Figure 16***, the input data and the kernel are represented in a matrix grid. A 3x3 convolution operation is performed here, whereby the exact number is always dependent on the form of the kernel matrix itself. In the second step, the kernel matrix is slided across the input matrix. Matrix multiplications are executed for each input field and summed up. This generates a Feature Map in which the convolution results are stored. It is called a receptive field once the Kernel matrix is placed over the input matrix and the convolution operations begin. Once the first procedure is completed, the Kernel matrix shifts to the next fields and starts calculating and recording the results in the Feature Map again. This process is repeated until the Feature Map is completed. If this is successful, the first Convolutional Layer has been obtained, as shown in step three. [30]

The above example shows a convolutional operation in two dimensions using a 3x3 Kernel. Nevertheless, most images require three-dimensional representation (height, depth and width), which explains the fact that the convolutional operations are usually performed in 3D. The depth here is mainly dedicated to the RGB colour space of images. However, for forwarding to the fully connected ANN layer, a 1D vector is required. [30]

Subsequently, all these Feature Maps are connected to each other by stacking them on top of each other, resulting in the final output of the convolution layer. [30]

Figure 17 below illustrates this in more detail:

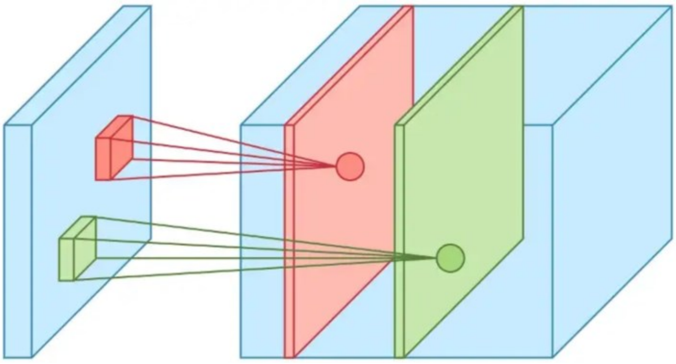


Figure 17: CNN Stacked Convolutional Layer [30]

Activation Functions

For a Neural Network to maintain its performance, it needs to be non-linear. For this purpose, activation functions are used as in ANNs and Autoencoders. In CNNs the results of the convolutional operations are passed on through the ReLU activation function. [30]

Stride and Padding

In convolution, stride variable determines how many steps a kernel must pass to be capable of entering the next values into the Feature Map. By default, this value is set to 1, but it is also possible to set larger values. Consequently, the Feature Map will become smaller. To preserve the dimension between Kernel and the input, a solution is provided using Padding. Here, the input is wrapped with zeros or other values along the corners. [18] [30]

Figure 18 below shows an example:

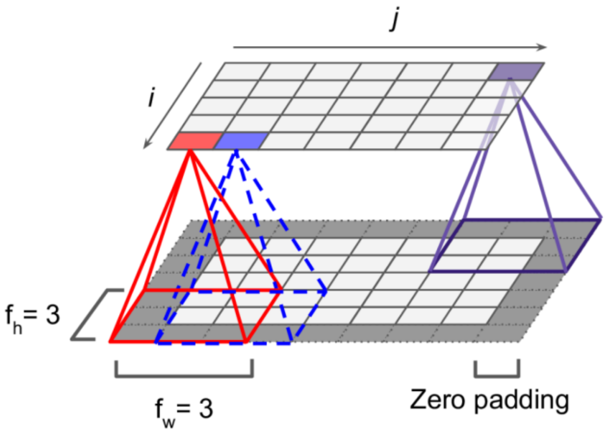


Figure 18: CNN Convolutional Layer Connections and Zero Padding [18]

Pooling Layers

The primary purpose of pooling operations is to reduce the size of the input image. Therefore, the risk of overfitting is reduced and the computational load, the memory usage and the number of parameters is minimized. The difference from pooling to convolutional neurons are that they do not use weights but aggregate functions, for example the mean or maximum value of each receptive field. [18]

The following graphic (***Figure 19***) illustrates the pooling operation:

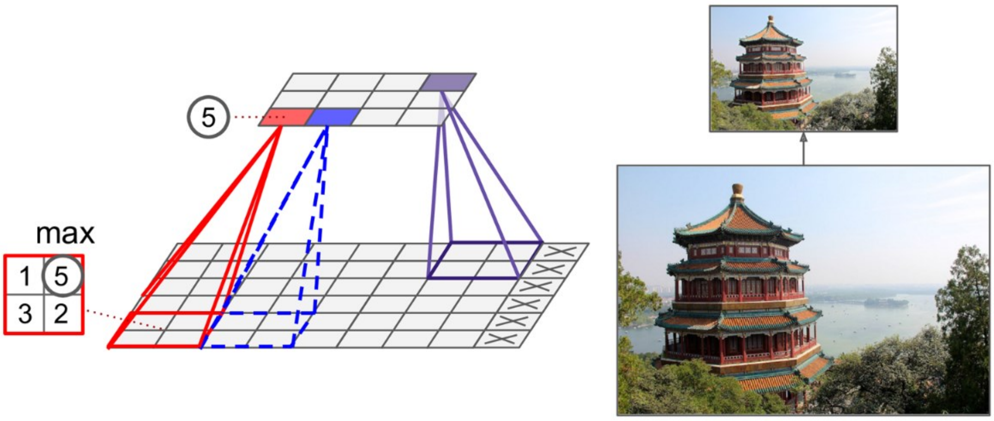


Figure 19: CNN Pooling Layer [18]

However, pooling offers more than just advantages - by reducing the input to usually at least half or even more, a lot of information can be lost or even certain processes as with images which are classified on a pixel basis. [18]

Fully Connected Layers

After all the required convolution and pooling layers, the results are forwarded to fully connected ANNs to produce the final output result. Training continues as with previous Neural Networks, using Backpropagation and Gradient Descent. [30]

3.4.3 Generative Adversarial Networks (GANs)

In the previous chapters of the Literature Review, the topics around Machine Learning and Deep Learning as well as the most important different algorithms and methods were discussed. However, Generative Adversarial Networks (GANs) are even more important for understanding the principles of deepfakes. They not only form the basis but are also considered pioneers in training Neural Networks for images and videos. [18]

GANs were born in the year 2014 when Ian Goodfellow published a paper on this training algorithm. As the name of this method implies, two Neural Networks train against each other to achieve the best possible outcome. The mode of operation is similar to a Variational Autoencoder. This type of encoding creates new instances based on randomness and probabilities, which apparently appear as if they were created from the training set reminiscent of the Gaussian normal distribution. This method is efficient, fast and easier to train. [18]

The following two types of convolutional or fully connected Neural Networks are involved [31]:

Generator

The Generator is responsible for generating an output (mostly an image) from a randomized distribution as input, such as the Gaussian Normal Distribution. The functionality is the same as that of a Variational Autoencoder - if the generator is fed Gaussian noise, a new image is created quickly and effortlessly. [18] [20]

Discriminator

The purpose of the Discriminator, on the other hand, is to select an image created by the Generator or one from the original training set and to guess whether it is a real image, or one created by the Generator. In practice, convolutional layers are often used for the Discriminator in order to train more efficiently. [18] [20]

Due to the fact, that two Neuronal Networks are pursuing different objectives, a new approach to training must be followed, which can be divided into the following two phases:

GAN Training Phase 1 – Train the Discriminator

The first training phase of a GAN focuses first on the Discriminator, which is supposed to distinguish the fake images of the Generator from the real images of the input training data. A random sample with a certain number of real input images is taken from the training set and the fake images of the Generator are added by the equal amount. Binary classification is used to label a real image with 1 and a fake image with 0. In addition, the binary cross-entropy loss is used and the weights of the Discriminator are adjusted with the help of the Backpropagation method. [18]

GAN Training Phase 2 – Train the Generator

The next training approach takes care of the Generator of fake images. Firstly, a new set of fake pictures is generated with the help of the Discriminator and presented to the Generator for distinction. Secondly, this time real images are not added, and all images are given the value 1, hence they are considered real. [18]

Furthermore, the Generator is never actually presented with real images from the training set as input. Only the adjusted gradients of the Discriminator are available, which is why no differences are generated by the Backpropagation during this training phase. [18] [31]
Figure 20 underneath explains the general sequence of training a Generative Adversarial Network, while ***Figure 22*** describes the complete training process in more details:

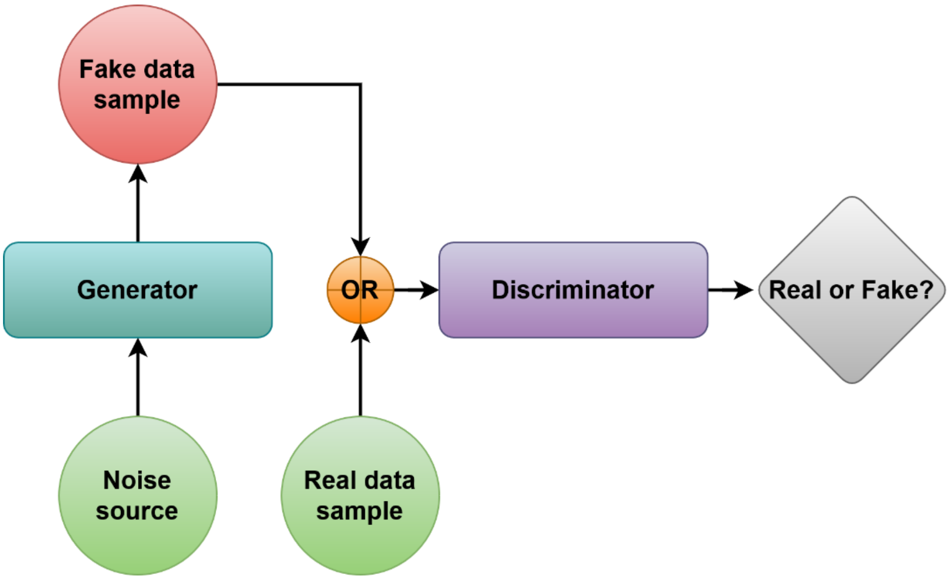


Figure 20: GAN Training Process [26] [31]

The input is usually a noise source (Gaussian noise) that is passed to the Generator to create the fake set. The synthetic data set is then combined with a real data set by an OR operation and presented to the Discriminator for distinction. [31]

Furthermore, ***Figure 21*** shows the exact GAN architecture behind each individual training sequence:

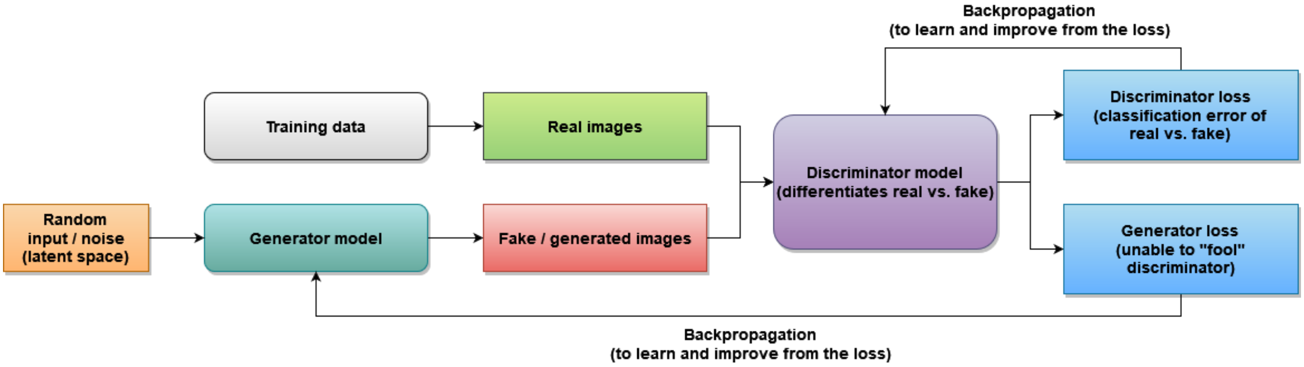


Figure 21: GAN Architecture of a Training Sequence [26] [32]

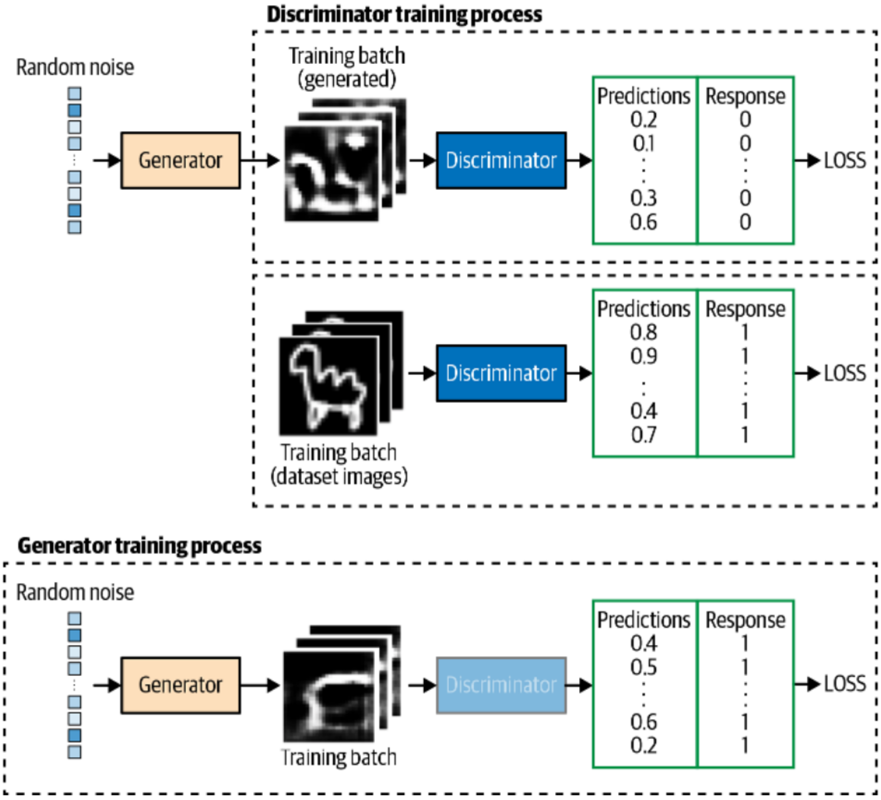
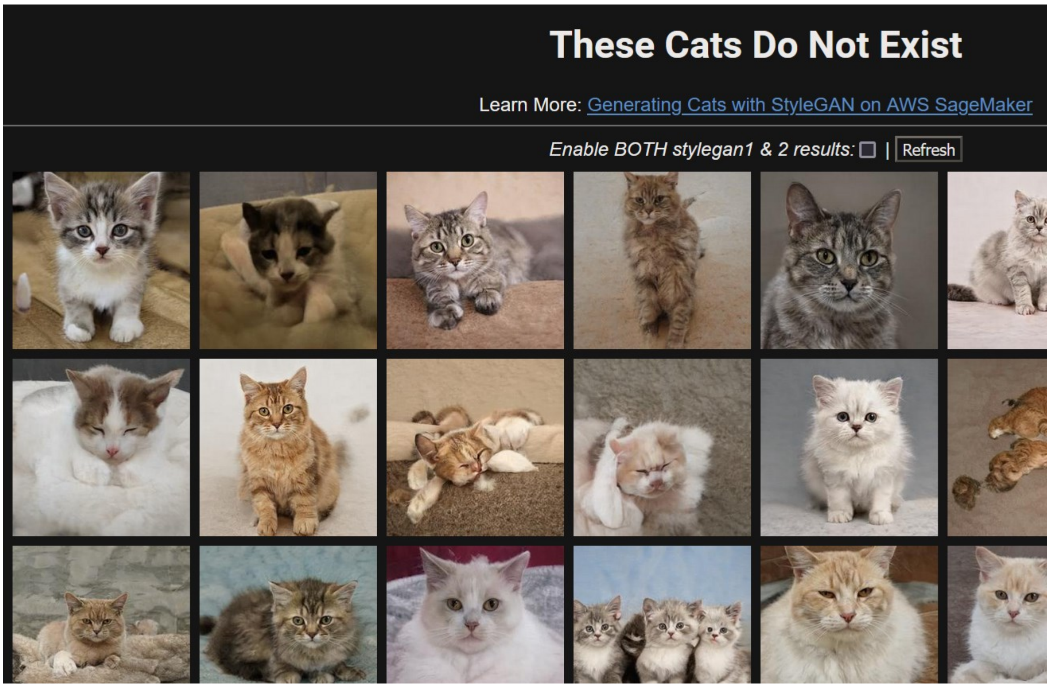


Figure 22: GAN Detailed Training Process [20]

GAN common areas of application

Some popular applications are, for example, websites as https://this-person-does-not-exist.com/ [33] or https://thesecatsdonotexist.com/ [34]. With every call to this website, a new non-existing face (or cat) is generated with the help of StyleGAN. Moreover, it is possible to select gender, age or an ethnicity to generate a new non-existing person. StyleGANs are created at NVIDIA Labs and consist of the combination of ProGAN and neural style transfer. This algorithm uses the Adaptive Instance Normalization method, which uses the mean and variance of each feature map to adjust output within the synthetic network. [20]

The following two figures (***Figure 23***, ***Figure 24***) illustrate examples of the previously mentioned websites for generating non-existent people and cats.



These Cats Do Not Exist

Learn More: Generating Cats with StyleGAN on AWS SageMaker

Enable BOTH stylegan1 & 2 results: □ | Refresh

Figure 23: Examples of "These Cats Do Not Exist" [34]

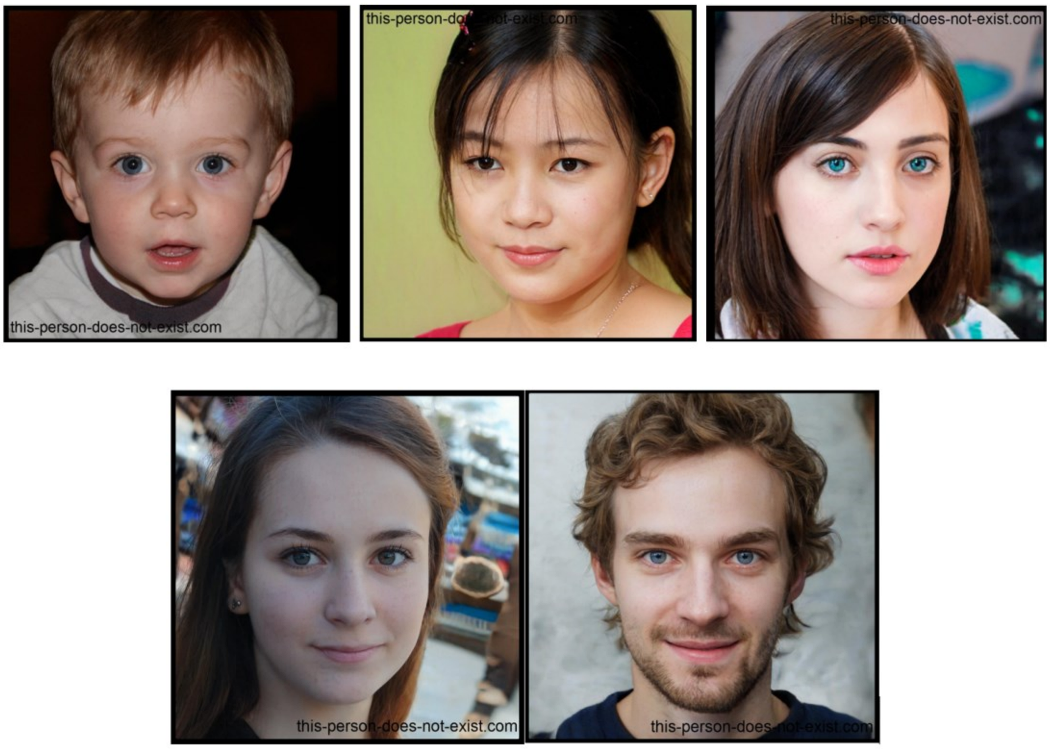

this-person-does-not-exist.com
this-person-does-not-exist.com
this-person-does-not-exist.com
this-person-does-not-exist.com
this-person-does-not-exist.com

Figure 24: Examples of "This Person Does Not Exist" [33]

4. Preliminaries

Now coming to the fourth chapter, the Preliminaries. This chapter focuses on the topic of Deepfakes, where they are applied, how they operate and how they can be created in just a few steps as well as the state-of-the-art today. In addition, the impact of Deepfakes is discussed in the context of cybercrime, with a focus on Phishing and Social Engineering, the society and privacy as well as the influence on the media today.

4.1 Deepfakes

Deepfake, a word that has attracted a considerable amount of media attention, especially recently. Headlines about city leaders of Vienna, Berlin and Madrid talking via video call to a fake version of the Kiev head of state Vitali Klitschko [35] [36] (even if it was not a real deepfake, which was proofed afterwards), for example, have caused a lot of commotion around this technology in recent months but have also brought more awareness back into focus that deepfakes may cause a lot of harm in the wrong hands. Let us start at the very beginning and take a critical look at this technology.

4.1.1 Definition

The word “deepfake” is a compound word from the English "Deep Learning" in combination with "Fake". The term not only embraces a Deep Learning method but is almost considered an established general media change. A deepfake generally describes an image, audio or video in which a person has been exchanged with another or synthetic – non-existing – person doing or saying things the person in the original media file did. [37] [38]

The deepfaked persons not only look different, but also say, do or move differently within the deepfake video compared to the original video. In deepfake images, faces are usually replaced so that a new face with the features of the original image is created. Definitionally, deepfakes also fall into the categories of puppet-master and lip sync. In a puppet-master deepfake, the target is considered a "puppet", while the person whom the target imitates the movements of the eyes, head and mimics the facial expressions in front of a camera, is classified as a "master". The other category is lip sync, where the mouth movements of the person in the deepfake video are adjusted to mimic a voice recording making it impossible to distinguish whether the deepfake person really has said that or not. [38]

4.1.2 Deepfake History

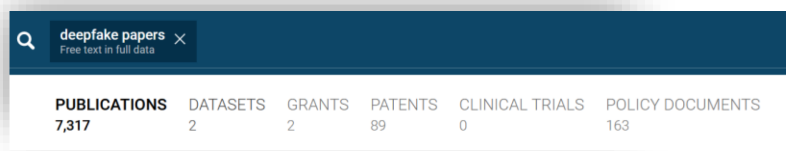
As the name “deepfake” already implies, it is a fake or modified file (image, video, audio) created with the help of deep learning algorithms and methods. The concept of faking or manipulate photos has existed for quite some time, especially in the field of image processing. The keyword here are applications such as Photoshop. [37] Furthermore, it is worth mentioning that in some cases, however, there have been already so-called **shallow fakes** before the technical possibilities for deepfakes were available on the market, which have not explicitly received an official definition, but which specifically describe a form of manipulation of videos with a deceitful intention without the use of artificial intelligence. [39]

But the real roots were found in research in the 1990s, as three researchers Christoph Bregler, Malcolm Slaney and Michele Covell were developing a then novel application which could transform videos of a person speaking in such a way that the speaker would then mimic the spoken words of another audio sample. In the early 2000s, these technologies were further developed in the course of research into facial recognition, so that as of 2016 the hardware requirements for today's deepfake implementations also became applicable on the general market for consumers. [37]

In 2017, when a Reddit user named "Deepfake" posted a video on his subreddit on the platform in which he swapped the face of the Israeli actress Gal Gadot (main actress of the movie Wonder Woman in 2017) with that of a porn star thus starting the era of deepfakes, which have since conquered the media world. The first deepfake was born and caused a tremendous sensation in the history of Artificial Intelligence, but not only in a good way. [37] [38] [39] What is particularly exciting to note is that the Reddit user who created the deepfake was simply a programmer with a strong interest in Machine Learning algorithms using open source libraries such as Keras and Tensorflow. The training data used was only collected with the help of published data from Google, YouTube and stock photos. [39] Shortly afterwards, the first application was published by another Reddit user: FakeApp was now freely available to everyone. This was quite a breakthrough, because now just about anyone could create a deepfake, even without any special technical knowledge or previous experience. [40]

As mentioned before, deepfakes have gained increasing attention in the media, in society and research, as well as in the general exchange of information in the last few years. As a result, awareness around this technology has also further increased.

Figure 25 and ***Figure 26*** are showing the increased occurrence in relation to publications between the years 2014 and 2023:



deepfake papers
Free text in full data

PUBLICATIONS	DATASETS	GRANTS	PATENTS	CLINICAL TRIALS	POLICY DOCUMENTS
7,317	2	2	89	0	163

Figure 25: Published papers about deepfakes overview [41]

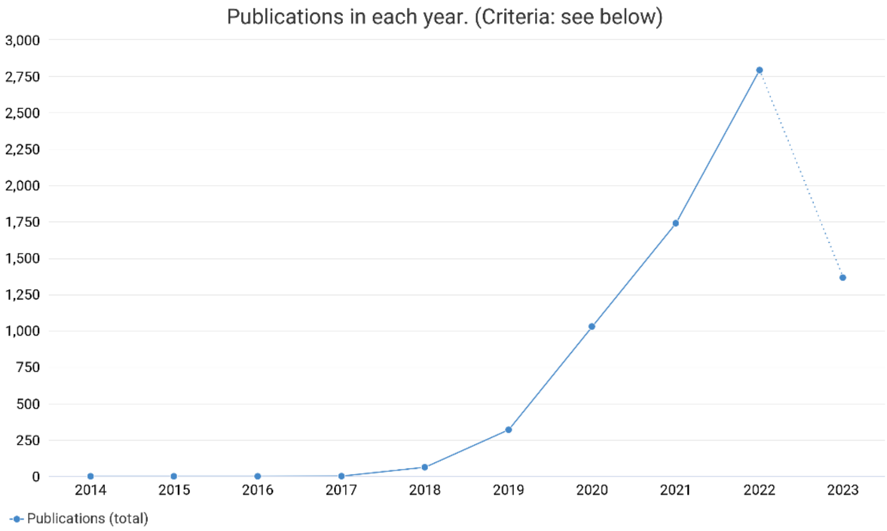


Figure 26: Number of deepfake publications in each year [41]

In the last nine years a total of **7,317 publications** have been published on topics related to deepfakes including papers, datasets, grants, patents and policy documents. A remarkable increase can be seen especially in the years between 2020 and 2022.

Figure 27 underneath shows the number of deepfake publications in each research category from 2014 and 2023:

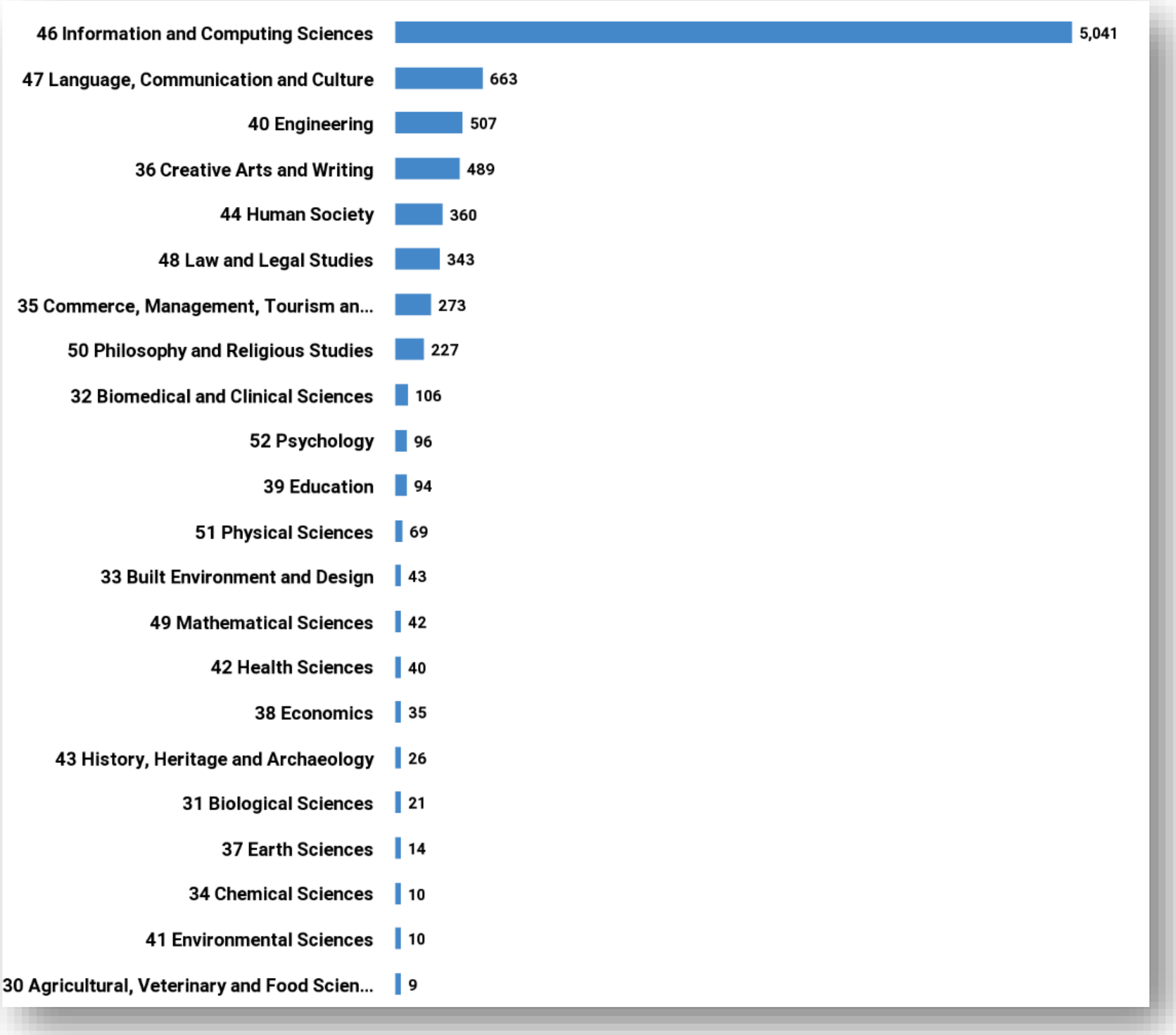


Figure 27: Number of deepfake publications in each research category [41]

Especially in Information and Computing Sciences deepfakes emerged as highly important topic in the last few years as evidenced in above figure in about five thousand various paper publications.

4.2 State-of-the-Art Analysis of Deepfakes

To have a suitable basis of information for the explorative scenario analysis of deepfakes, this chapter describes the currently available technical and organisational measures, applications and tools of deepfakes in more detail. Furthermore, the distinction between the various factors in connection with deepfakes typologies is addressed, although the focus in this thesis will predominantly concentrate on deepfake videos.

4.2.1 Common Application Areas

One of the most crucial points in the creation of deepfakes is that plenty of image or video material of the person to be deepfaked must be available. Since most content is publicly available on the internet from celebrities, they are also predominantly the targets of many deepfakes. [38]

The following areas, as illustrated in the figure below, are those in which deepfakes are frequently applied:

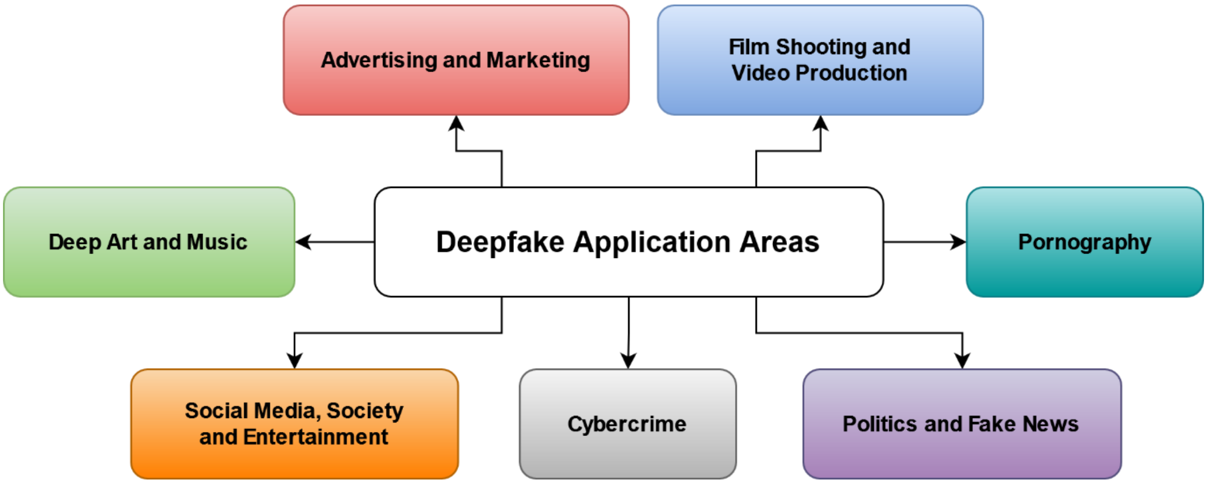

Advertising and Marketing
Film Shooting and Video Production
Deep Art and Music
Deepfake Application Areas
Pornography
Social Media, Society and Entertainment
Cybercrime
Politics and Fake News

Figure 28: Common Deepfake Application Areas [26]

Social Media, Society and Entertainment

Besides deepfake Tom Cruise [42] on Instagram or Youtube or the many new face filters on TikTok [43] and Snapchat, alongside the great and varied entertainment there exist also some drawbacks. Jan Böhmermann has set a great example in the German-speaking region: In 2015, he created a deepfake of the Greek Finance Minister Yanis Varoufakis, who pointed his middle finger at the world in this video. [44] In addition, some funny meme clips exist, mainly related to lip sync deepfakes, such as Barak Obama saying things he have never said before, e.g., calling Donald Trump "[…] a total and complete dipshit […]" [45] [46]. Another example is David Beckham fluently speaking in nine different languages created by the Technical University of Munich, Germany. [47]

Advertising and Marketing

Besides covering social media or the news in general, deepfakes are also used in many other sectors, e.g., for advertising and marketing purposes. In marketing, personalised advertising is very important. With the use of Deep Learning technologies such as deepfakes, it becomes possible to adjust consumer preferences in real time and therefore to conduct marketing even more effectively. [45]

Film Shooting and Video Production

More and more deepfake technologies are also becoming established in the film and video industry. Since 2018, the German company Volucap has specialised in the development of the "deepfake gold standard" in cinema production. The company develops a 3D animated model of the actors with the help of its own trained Neural Network, which can be inserted into the film scenes effortlessly. [36]

Politics and Fake News

In spring 2022, when the war between Russia and Ukraine started, an unprecedented deepfake scandal broke out. A deepfake video of Ukrainian President Volodymyr Zelenskyy was created in which he admitted that Russia had won, and Ukraine had conceded defeat. However, this was merely an attempt by Russia to use this technology to threaten the Ukrainian population and spread war propaganda. After it became known shortly afterwards that it was a deepfake, the video was deleted from some social media platforms such as YouTube, Twitter and Facebook. [48] [49] [50]

Deep Art and Music

Another area of application is found in art and music. With the help of the many different Generative Modelling approaches, it is now easily possible to create deep art and also music. [20]

Pornography

Furthermore, a broad area of application is pornographic videos or photos generated with the help of deepfake mechanisms. Unfortunately, the problematic here mainly impacts people, especially women, who did not explicitly give consent into creating such contents about them. In addition, the number of reports of deepfake porn being created in the course of revenge motives or intended reputational damage is also increasing. [37]

Cybercrime

In terms of cybercrime, deepfakes play a major role. With these technologies and systems, new opportunities arise, especially for attackers and script kiddies - and with hardly any time or effort, but with a mostly very lucrative outcome. One of the most important resulting questions is particularly how this may affect the media and society in the future, but also organizations and private individuals.

4.2.2 Deepfake Requirements

In the previous chapters, mainly the definition, history and known areas of application have been discussed. But what is required to create a deepfake? Consequently, the necessary components and requirements for generating deepfakes are discussed in more detail.

Technology Know-How

First and foremost, it is crucial to understand that while in the beginning a deeper understanding of the underlying technology was required, such as Deep Learning and the workings behind ANNs, GANs and CNNs. With the advancement in technology and the meanwhile relatively wide offer of open source as well as commercial applications, websites and tools, this deep knowledge is nowadays no longer mandatory to create a deepfake. This makes them relatively user-friendly enabling them to be created for any audience with varying levels of computer knowledge and expertise. [1] [39]

Amount of provided Data

To be able to create a deepfake, data on the persons or objects that are supposed to appear in it are also obviously required. Furthermore, for high-quality and authentic results in the creation of deepfakes, masses of raw media material are needed to sufficiently train the Deep Learning algorithms. An approach that is particularly relevant in research lies in the reduction of data dimensions as well as the preservation and conscious mitigation of quality loss while generating the deepfake. [45]

Time and Resources

Even though a variety of tools and programs for creating individual deepfakes are freely available, time is an essential resource requirement when creating and implementing deepfake content. Not only is a model generated and trained, a certain amount of time and resources involved for debugging, testing and potential improvements should be considered.

Tool and Algorithm Selection

Considering that deepfakes require an authentic result, it is also necessary to choose the implementation software and algorithm wisely and according to the preferred requirements.

Software and Hardware Requirements

When generating deepfakes the specific software and hardware requirements for each tool, program or website must be considered. In addition, as described in [51], there are pre-trained GANs and models, which eliminates the computationally intensive part for the end user. Moreover, with the technological and technical opportunities available today, computers with a relatively powerful Graphical Power Unit can generate a deepfake in a matter of minutes to hours, depending on how deceptively real the result should be.

Scope and Purpose

One of the most fundamental requirements for deepfake generation is the purpose behind it, as well as the actual intended use. Furthermore, it is important to consider possible post-production adjustments or improvements after creating a deepfake.

Occlusions

So-called occlusions can help to ensure initial approaches toward correcting errors or problems while training the algorithm. Occlusions describe hair, body or clothing parts as well as other objects, such as glasses or jewellery, which actively cover the necessary parts of the face. [45]

Artefacts

Artifacts also take on an additional function. These occur during the creation process due to neural networks processing and analysing the videos frame by frame and not as a whole video. Noticeable artefacts are flickering or flaring. Researchers are constantly aiming to prevent these anomalies, for example by implementing coherence losses or the deliberate use of RNNs. [45]

In summary, multiple factors and requirements are involved in the process of developing deepfake content. How much importance is attached to the individual requirements depends strongly on how convincing or authentic the result is supposed to be, as well as on the actual purpose of application of the deepfake.

4.2.3 Common Deepfake Applications, Tools and Websites

Meanwhile, numerous possibilities, tools and websites are available on the market to create deepfakes both quickly and easily. According to [37], the two most frequently used tools in practice are predominantly **FaceSwap** and **DeepFaceLab**.

FaceSwap

FaceSwap was already developed in 2017 and is considered a multi-platform deepfake software on the market. The software is based on the well-known Machine Learning framework Tensorflow and uses corresponding Machine Learning libraries, including Keras. Operating systems supported are Windows, Linux and macOS. The technical requirements include the availability of a modern Graphical Power Unit (GPU) supporting CUDA, thus creating the preconditions for ensuring high performance while creating a deepfake with the FaceSwap software. [37]

In addition, FaceSwap offers two ways of application: via GUI or command line. The algorithm works by using Autoencoders, training one Autoencoder for the source and one for the target video. By using the combination of Encoder and Decoder, the faces are replaced and adjusted accordingly. [37]

DeepFaceLab

The second of the most popular and commonly referenced tools is DeepFaceLab, which claims that approximately 95% of all deepfake videos ever created have been made using their software. Furthermore, DeepFaceLab is considered a division of FaceSwap since 2018, but provides certain refinements compared to its predecessor. [37]

The underlying principle is also built on Autoencoders, but a specific type of Neural Network is used - so called Multi-Task Cascaded Convolutional Networks (MTCCNs). These NNs perform the facial recognition in three different steps to identify the individual facial entities, for example, right eye, left eye, right corner of the mouth, left corner of the mouth or the nose. [37]

This type of Neural Network performs so efficiently it is even able to solve occlusions and difficulties with lighting conditions, facial profiles looking sideways, but also variations in facial expressions. [37]

DeepFaceLab is supported on Windows (app installation possible), macOS and Linux by executing via the command line or by running scripts. Due to the special form of the NNs, the training processes often last from a few hours up to several days, which is why powerful computer hardware including adequate cooling is required as a system requirement. [37] [51]

The two tools described in detail above of course do not represent the complete range of tools and software available on the market. To obtain a short overview of other widely available applications, the following table (***Table 1***) lists the best-known open source and commercial applications and websites:

Table 1: Deepfake Applications and Websites Overview [26] [52]

NAME	REQUIRED MEDIA	PURPOSE	APP OR WEBSITE	COMMERCIAL OR OPEN SOURCE	OPERATING SYSTEM
DeepFaceLab [51]	*Videos*	*Deepfake videos with Machine Learning and Image Synthesis*	*App*	*Free (available on GitHub)*	*Windows*
Faceswap [53]	*Videos*	*Multi-platform Deepfake Software*	*App*	*Free*	*Windows, macOS, Linux*
Reface [54]	*Images*	*GIF memes with GANs*	*App*	*Free, but in-app purchases possible*	*Android, iOS*
Lensa AI [55]	*Images*	*Photorealistic AI portraits in different styles*	*App*	*Free, but premium version available*	*Android, iOS*
DeepSwap.ai [56]	*Images, Videos, GIFs*	*Face Swap*	*Website*	*Commercial*	-
Deepfakes Web [57]	*Videos*	*Deepfake videos*	*Website*	*Commercial (3$ / hour of usage)*	-
Wombo [58]	*Single Image*	*Lip Syncing (of singing faces)*	*App*	*Free, but in-app purchases possible*	*Android, iOS*
MyHeritage [59]	*Images*	*Image Animation*	*App* *Website*	*Free*	*Android, iOS*
Deep Art Effects [60]	*Images*	*Deep Art Deepfake Images*	*App*	*Free, but in-app purchases possible*	*Android, iOS*
FaceApp [61]	*Images*	*AI edited Images*	*App*	*Free*	*Android, iOS*
DeepFaceLive [62]	*Videos or Camera input*	*Real-time Face Swap for Streaming or Videos*	*App*	*Free*	*Windows, Linux*
Avatarify [63]	*Images*	*AI Face Animator, Live mode possible*	*App*	*Free, but in-app purchases possible*	*Android, iOS*

4.2.4 Deepfake Categories, Types and Functionalities

In general, deepfakes can be divided into three categories: images, videos and audio files. ***Figure 29*** illustrates this in more detail:

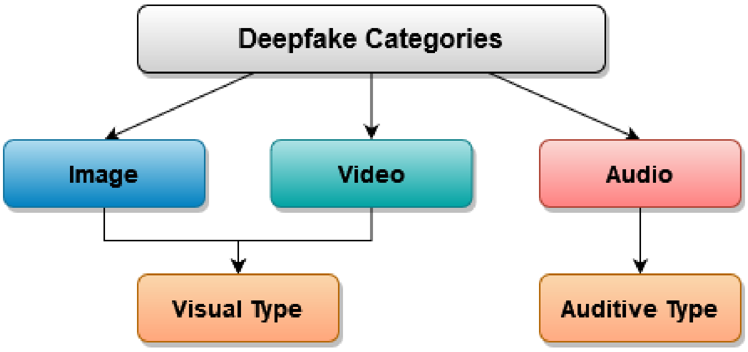


Figure 29: Deepfake Categories [26] [37]

In theory, there are several types and approaches as to what can and cannot be counted as a deepfake. In principle, however, a distinction is made between two main categories: the manipulation of audio files and the manipulation through optical alterations. Hybrid types and combinations may of course be generated and applied as desired.

The following graph (***Figure 30***) lines out the different types and approaches of today's deepfake functionalities:

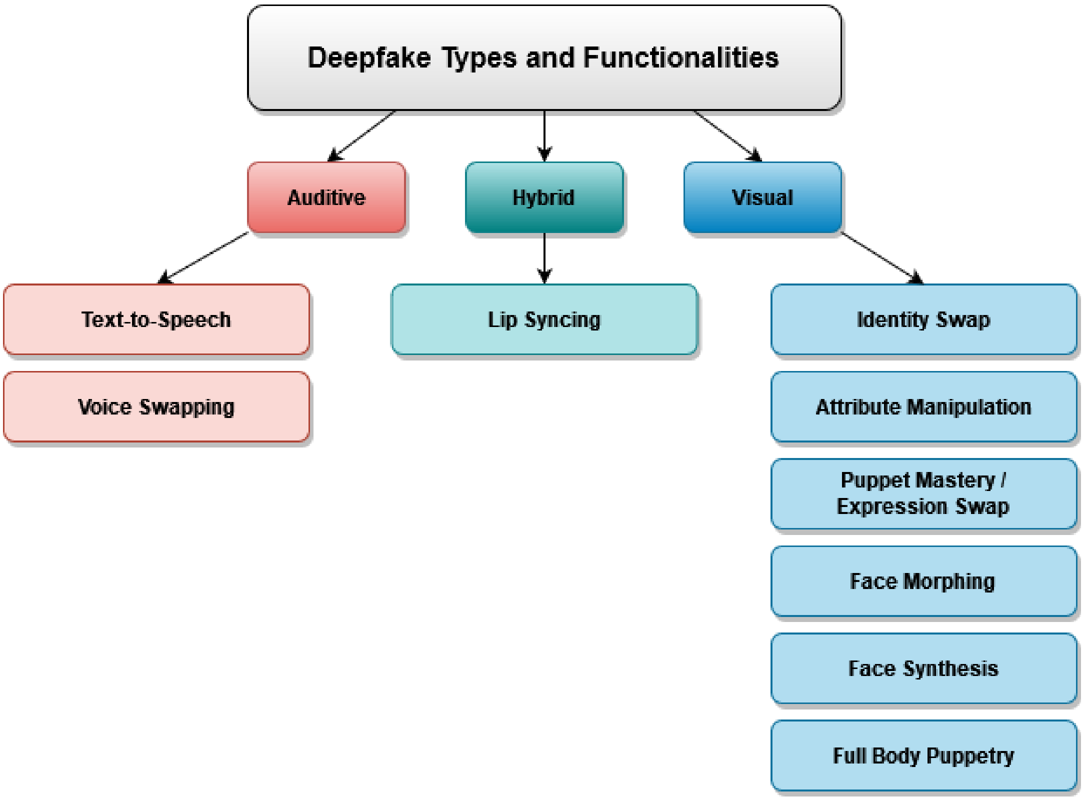


Figure 30: Deepfake Types and Functionalities [26] [38] [45] [64]

Among the **auditory variants**, **Text-to-Speech** mechanisms or even complete **Voice Swapping** are familiar types of deepfakes.

So-called **Lip Syncing** is often used as a **hybrid** method. The lip movements of a video are adjusted to fit the audio source material. The angles of the mouth and the corresponding movements are modified in this way, so that persons suddenly say or sing words they did not say in the original material. It is particularly important here to ensure realistic results for the lip movements, the mouth, but also the posture of the head and the blinking of the eyes. FACIAL-GAN provides a good example of the technical implementation of Lip Syncing. [37]

Since the focus of this thesis lies on the visual area, i.e., deepfake videos and images, several **visual types** and functionalities based on the previous figure (***Figure 30***) are explained based on [37] [38] and [64] in more detail:

Identity Swap or Face Swap

Defines exchanging faces in videos. In this method, the deepfake is generated in three phases. In the first stage, the source and target faces are recognised in the videos and collected as individual facial frames. Predominantly, similar poses and traits of the face are considered in this stage. The second step involves the algorithm swapping the faces and making necessary adjustments to light and colour, so the swapped face looks genuine within the target video. Finally, the sequence of all images is determined by calculating the respective matching distances between the images to create a continuous transition. [37]

Typically, a further distinction is made between two subcategories: the classical variant using graphical techniques, such as Face Swap, or the application of Deep Learning mechanisms to create deepfakes, e.g., like the ZAO mobile app. Face Swapping was one of the first deepfake technologies on the market, which Snapchat incorporated as a filter in 2017. Face Swap involves exchanging the faces of different people with each other, allowing the people to be represented in a different context. [38] [64]

Attribute Manipulation

Often also known as face editing or retouching, this visual deepfake type describes the change of facial attributes such as hair, hair colour, skin and skin colour, gender, age, beards or inserting glasses and can even change the emotions of the face. Various GAN algorithms are used for implementation, for example StarGAN – a technology based on StyleGAN. However, this variant allows changes to be made only to certain parts of the video through the ability of the generator to learn additional section identifiers. Problems usually only arise with faces inside profile instead of a frontal view. [38] [64]

Puppet Mastery or Expression Swap

Also known as Face Re-enactment, where the aim is to change the facial expression of a person to that of another one, Puppet Mastery or Expression Swap are another deepfake manipulation approach. The most popular algorithms used are Face2Face and Neural Textures. Challenges or problems mostly occur once the head posture does not appear dynamic or unnatural, or when facial expressions change. [38] [64]

Face Morphing

A type of manipulation whereby faces are created that can share the biometric characteristics of several people. This variant is mostly used in animations and movies. In this technique, invisible transitions are responsible for changing one person or object into another. [38] [64]

Face Synthesis

Face Synthesis is a type of deepfake creating completely new faces that does not yet exist with the help of GANs. In chapter *3.4.3 Generative Adversarial Networks (GANs)*, the functioning as well as examples of GANs were described in more detail. In general, these techniques are mainly used in the video game sector and in 3D modelling. [38] [64]

Full Body Puppetry

Thus, a further technique to deepfake videos is the Full Body Puppetry method. This variant reproduces entire movements or sequences of movements and transfers them to another person. A good example are dance apps, such as those from the University of California. [38]

4.2.5 Deepfake Creation Process

At this point, detailed information has been provided in terms of application areas, tools, requirements and the various categories and types of deepfake technologies. This sub-chapter focuses on the actual process of creating deepfakes.

The underlying technologies and algorithms behind deepfakes are primarily Autoencoders, Convolutional Neural Networks and Generative Adversarial Networks. Of course, it is also possible to create deepfakes through post-processing and visual effects. However, these variants are not considered "real" deepfakes. [38] In the chapters ***3.4 Deepfake relevant Neural Networks***, the exact functioning and processes of Autoencoders, Convolutional Neural Networks and Generative Adversarial Networks were identified and described during the Literature Analysis.

The following figure (***Figure 31***) describes the general creation process of deepfakes:

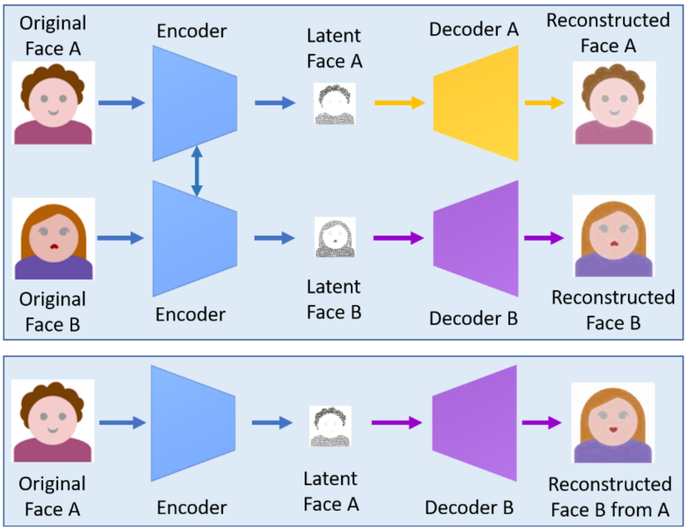


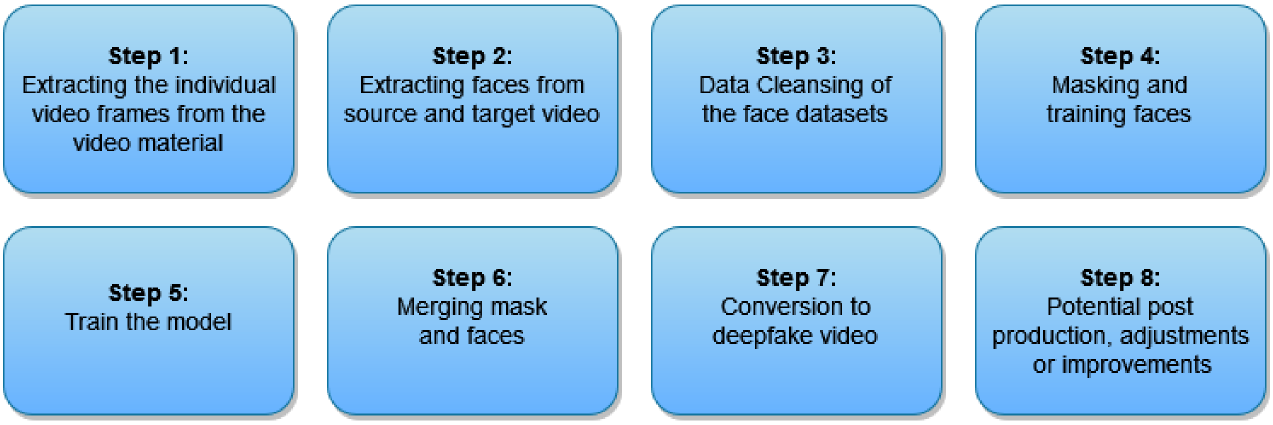
Figure 31: General Deepfake Creation Process [38]

As highlighted in [37] and [38] Autoencoders play an essential role in creating deepfakes:

Autoencoders are used to dimensionally reduce the images while preserving the "meaning" behind them. The Decoder part in an Autoencoder reconstructs an image from the highly compressed latent space as can also be seen in ***Figure 13***Figure 13: Autoencoders Architecture [29]. To further train the Autoencoder to reconstruct images or faces while preserving relevant features and facial details often some noise is added [65]. Since Autoencoders reconstruct frames from latent space and can be trained on video streams rather than still frames they can be used to correct errors. These properties are studied in [66].

Therefore, image reconstruction with Autoencoders can be summarized as training an Encoder and Decoder pair until they are able to identify and reconstruct a certain face. Since the main role of the Encoder is creating the latent space representation of the input a well-performing Encoder can be re-used. Thus, the Decoder's purpose is to create images from that latent space. Using this a face swap can be archived by training a single encoder with multiple decoders in a way that each person (see ***Figure 31***) has a unique Decoder that applies that person's face to the latent space. Swapping the face-applying Decoders a face swap can be achieved. As there are no Discriminators involved a face swap is an Autoencoder and not a GAN as often assumed.

For **creating a deepfake video** the general procedure involves the following eight steps based on [45]:



Step 1: Extracting the individual video frames from the video material

Step 2: Extracting faces from source and target video

Step 3: Data Cleansing of the face datasets

Step 4: Masking and training faces

Step 5: Train the model

Step 6: Merging mask and faces

Step 7: Conversion to deepfake video

Step 8: Potential post production, adjustments or improvements

Figure 32: Creation Steps of a Deepfake Video [26] [45]

Attempting to synthesize a deepfake, certain requirements must be met, as described in ***4.2.2 Deepfake Requirements***, including video material of the individuals intended to be deepfaked.

4.2.6 Deepfake Identification, Detection and Verification

Subsequently to the preceding section on the process for generating deepfakes, the various approaches to detection and verification take on an essential role here. In addition, for the convenience of the reader, the chapter is further divided into the two subchapters on Deepfake Identification Factors and Deepfake Detection and Verification Approaches.

4.2.6.1 Deepfake Identification Factors

Based on the current state-of-the-art, naturally a few technical systems and tools already exist for detecting applied deepfake technologies in images, videos or audio files. Fortunately, not only purely technical features are indicative of the presence of these systems. Several of these factors can already be observed by the human eye.

According [1] [37] to be able to recognize deepfakes – even as user – the following **indications** could speak for the **use of deepfake technology** used in a video, image or audio file:

- Blurring or misalignment of the face or the whole visual
- Unnatural or abnormal eye movements
- No blinking
- Inauthentic facial expressions due to possible use of Face Morphing
- Unnatural shaped body or body parts
- Inconsistent lip movements or bad lip syncing
- Unnatural colours of the skin or face (high saturation, no or misplaced shadows, incorrect light incidence on the face, ...)
- Unnatural lightning
- Inauthentic-looking hair and teeth, as the deepfake algorithms are still not able to efficiently reproduce individual strands of hair or individual teeth
- Weird looking background, e.g. digital background noise
- Strong compression of the audio, image, or video quality
- Robotic-sounding voices
- Pixel errors
- Occlusions (objects or body parts hiding essential parts of the face)
- Flickering images

The following picture underneath **examples a deepfake occlusion** generated by the StyleGAN deepfake algorithm of this-person-does-not-exist [33]:

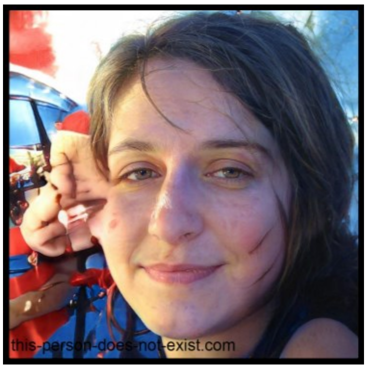


Figure 33: Occlusion Example of StyleGAN [33]

4.2.6.2 Deepfake Detection and Verification Approaches

It is obvious that further development and the continuous improvement of today's technology makes it harder to recognize or identify deepfakes without technical tools or detection and verification systems at some point in the future. Therefore, this section addresses the approaches and nowadays technical implementations for the detection of deepfakes available today. [1]

Although the algorithms and technical capabilities for producing deepfakes have been available on the market since 2017, **no application** has yet been able to **detect a deepfake -** whether in an image, video or audio file **- 100% successfully**. [45] As with all technologies where falsification is generated, there is a race between the methods of falsification and the detection of falsification. This race is described in [64] and also mentions the major research trends and their detection methods and accuracy.

Mostly, the detection of deepfakes is seen as a pure binary classification problem. The first attempts to detect deepfakes were made using manually defined features to pick out inconsistencies within the deepfake videos. [37] However, based on [45], the actual detection approaches can be predominantly divided into two different approach categories - **artefact-based** and through **indirect detection techniques**:

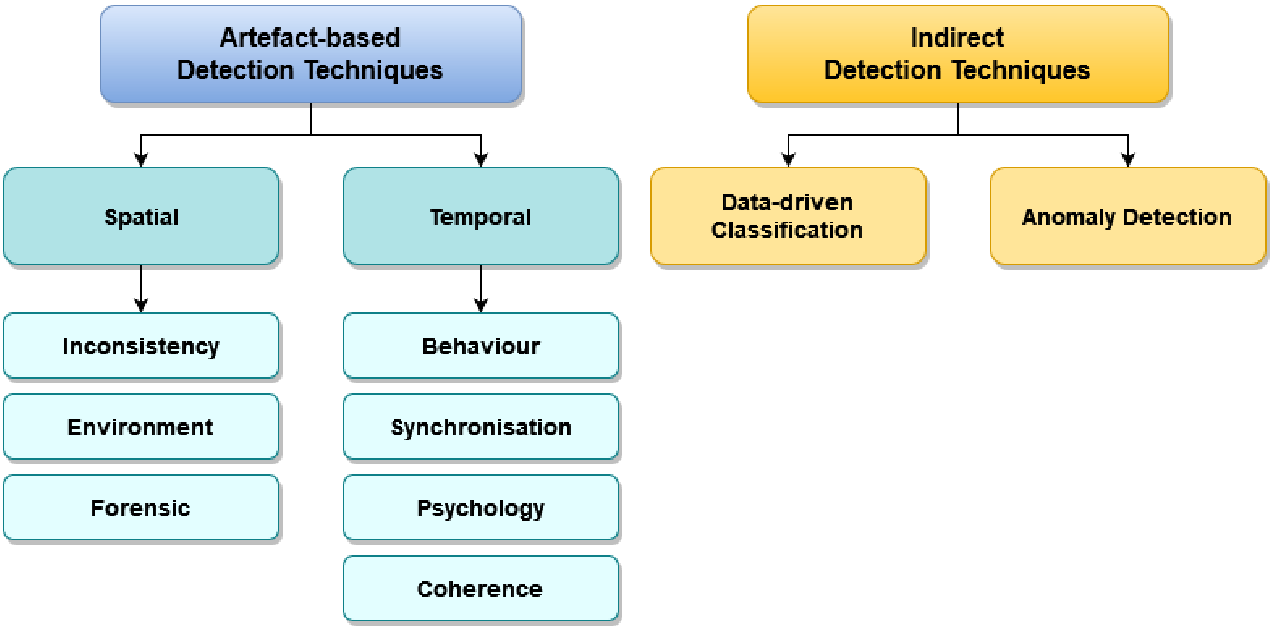


Figure 34: Deepfake Detection Approaches [26] [45]

As illustrated in the graph above, both approaches can be further subdivided. Here, **artefact-based** detection techniques are split into both **spatial** and **temporal dimensions**. Compared to the **indirect detection features**, which can only be further **divided into data-driven classifications and anomaly detections**, the artefact-based detection features are further divided into additional categories. [45]

Firstly, **indirect detection features** based on [45] are listed as following:

Classification

The deepfake classification method involves **CNNs built on data-driven models**. For example, the approach to identifying fake and non-fake is trained to the Neural Network.

Anomaly Detection

Anomaly detection is the **identification and analysis of outliers**. Similarities and dissimilarities in language, face and emotions are analysed to detect anomalies.

Secondly, the following three **spatial artefact-based detection features** [45] are distinguished:

Inconsistency

Inconsistencies in videos include, above all, **conspicuous occurrences** for example, uneven head movements or inappropriate orientation points. Based on the knowledge of always existing small errors in estimating the 3D face pose when placing the fake face, deviations from this can be indicative of a deepfake.

Environment

The environment may also indicate a potentially synthesised video. For example, **abnormal appearances or inappropriate lighting conditions**, e.g., shadows can be considered.

Forensics

The third classification of spatial artefact-based deepfake detection methods is forensics. In the process, **deepfakes** can be **exposed due to traces** left behind, such as patterns or features given during the generation of the deepfake. These identification marks often are compared with a type of **fingerprints**.

Finally, following four **temporal artefact-based detection features** [45] are defined:

Behaviour

In time-based artefact detection methods, behaviour is a key factor. Mostly, noticeable or **unnatural-looking behaviour patterns, movements or** even **facial expressions and gestures** hint towards a deepfake. Dependencies and correlations between facial expressions and the associated head movements are used to distinguish people and identify modified videos.

Synchronization

A further identification attribute are issues or **errors in** the **synchronization of** the individual **image frames** in the videos, for example the matching of the mouth shape with the respective spoken words.

Psychology

People communicate not just through words, but primarily by subconscious gestures. Furthermore, humans have certain biological processes which cannot be controlled or switched off, such as **blinking** several times in a few seconds, **breathing** or **reflexes**.

Coherence

Coherence is the **flickering or jittering between frames** in a video occurring due to the lack of Optical Flow Fields or the presence of visual artifacts. Fake videos generally produce stronger distortion and flickering than original videos and thus can be detected by using CNNs, for example.

To successfully carry out **manipulation detection** for generated **deepfake videos**, **scalable and robust algorithms** are **required** as well as **high-quality data sets** for training the respective models. In this context, indications such as recognition methods for static images do not necessarily also apply for videos, as they only focus on single frames thus not taking the temporal component of videos into consideration, speak for themselves. [45]

According to ***Figure 35***, the following order can be applied concerning deepfake detection in videos:

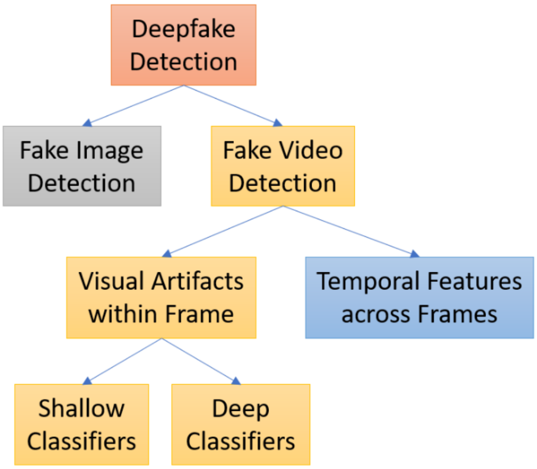


Figure 35: Deepfake Detection Categories [38]

When detecting deepfakes, the first step is to differentiate whether the video is a complete fake or only contains individual fake image frames or sequences. Once this is done, there are two further ways to differentiate further: Visual Artifacts within Frame or Temporal Features across Frames. There are differences between Shallow Classifiers and Deep Classifiers regarding Visual Artefacts within Frames. [38]

The approaches mentioned in [38] and [45] for classifying and recognizing deepfakes, especially in a visual context (deepfake videos), point to the complexity and thus also the resulting challenges of deepfake technology. Furthermore, not only simple methods, which are also used for standard image detection, are also suitable for application to deepfakes. In the further future, therefore, research will continue to be needed to find suitable optimization options in connection with the detection of deepfakes.

4.2.7 State-of-the-Art Implementations of Deepfake Technologies

The following chapter explains the emerging implementation methods, techniques and technical systems, tools and applications that are nowadays available on the market. The procedure that now continues comprises the explanation and description of various implementation approaches for deepfake detection and verification systems. As listing and precisely explaining all possible state-of-the-art technologies would go beyond the scope of this master's thesis, only a few already established systems will be focused on.

As described in the previous chapters of the state-of-the-art analysis of deepfakes, several different ways and methods exist of deepfake detection and verification systems, which now thus leads to the various implementation techniques and approaches to detect and verify on deepfake videos. Furthermore, it is also evaluated from an information security perspective how resilient these technologies are against deepfake attacks.

4.2.7.1 Technical Implementation Approaches of Deepfake Detection and Verification

Due to the relatively high compression of videos, it is often difficult to apply image recognition methods. In addition, temporary characteristics appear in videos, which make it even trickier for the deepfake detection software to recognize the fakes from the individual real picture frames.

CNNs and Long Short-Term Memory (LSTM)

Due to [38], two different approaches are used to detect deepfake videos: temporal features across video frames and visual artefacts within video frames.

One variation to detect deepfakes is to use CNNs and LSTM procedures to temporarily extract features from a video sequence. These features are interpreted as sequence descriptor. The detection network uses FC layers and the previously extracted sequence descriptor to calculate the probabilities for detecting whether the respective frame from the video sequence is authentic or a deepfake. The graphic underneath (***Figure 36***) illustrates this process:

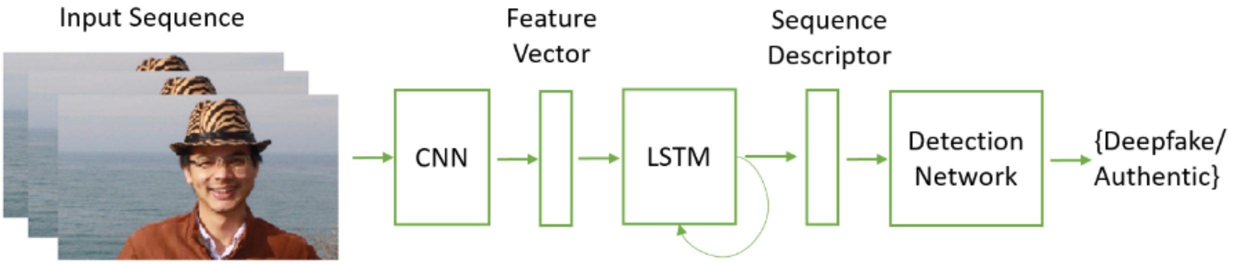


Figure 36: Deepfake Detection using CNNs and LSTM [38]

Face Swapping and Face Re-enactment

Another method to distinguish deepfakes from real videos, according to [67], is Face Swapping and Face Re-enactment. Both approaches give the opportunity to synthesize the destination and source images or videos respecting the given target image or video. While the source image or video provides the necessary background and texture information through Face Swapping or the motion information through Face Re-enactment, the target image or video delivers the necessary identity information.

Figure 37 illustrates the previous described procedure:

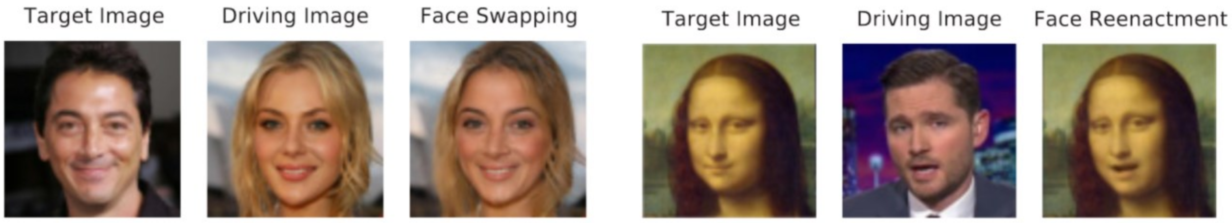


Figure 37: Face Swapping and Face Re-enactment Deepfake Detection Approach [67]

Facial Liveness Verification (FLV)

A promising alternative to using passwords is Face Recognition, which verifies a person's identity based on a face in a picture or video. Due to the use in the identification and authentication of individuals in predominantly security-critical areas, such systems also appear as so-called Platform-as-a-Service (PaaS), playing a major role for cloud applications. [67]

Well-known areas of application include online payments, online banking and government services. An example of a unique security evaluation tool in the deepfake area is the FLV framework LiveBugger. [67]

Adversarial Training and Anti-Deepfake Detection

Since synthetic media technologies, like deepfakes, are rapidly evolving, FLV and other Anti-Deepfake systems face several challenges related to information security in the future. [67]

For example, influencing factors from tools like LiveBugger could be exploited. These include, for example, bias (e.g., gender or ethnicity), which makes it possible to select targeted individuals and attack them with the help of deepfakes. Another factor is that adversarial training techniques (GANs) could bypass deepfake mechanisms of efficient systems like FLV. This also explains another challenge of deepfake detection and verification systems in the future. Because of the special training process, the discriminator learns the difference between real and fake images or videos. However, this is also the goal of anti-deepfake detection systems, where the generator aims to trick the discriminator. [67]

Coherence Detection

This detection method uses the coherence factor to check whether the frames in a video follow each other continuously and without flickering. For security reasons, this information is often not passed on by the operators or manufacturers of such systems. [67]

Lip Language Detection

Another recognized method for detecting deepfakes is Lip Language Detection to identify whether the lip movement in a video fits the corresponding audio. [67]

Artificial Fingerprinting for Generative Models

Based on [68], the artificial inclusion of fingerprinting approaches can also help to identify deepfakes. These fingerprints incorporated into the training data of generative model algorithms for generating deepfakes thus subsequently indicate whether it is a deepfake or not.

From a security point of view, the protection of these artificial fingerprints is essential in order to prevent attackers from misusing them for their own purposes and thus avoid this type of detection approach. [68]

Deepfake Attribution and Network Watermarking

Other recognition mechanisms based on fingerprinting are, for example, extending the fingerprints to individual attributes of the models (Deepfake Attribution) or Network Watermarking, in which the fingerprints via the network and the transmitted data help to identify deepfakes. [68]

Additionally, to all the examples of various implementations for deepfake detection and verification above, following approaches and methods worth mentioning according to [67] [68] [69] are listed as following:

- Deepfake Detection through Speaker and Speech Recognition
- Anti-Deepfake Real-time Detection
- Image Steganography and Watermarking
- Eye Tracking, Blinking and Integrity Verification

4.3 Deepfake Detection and Verification Metric Approaches

In relation to the methods for detecting and verifying deepfakes described in detail above, the following metrics due to [67] can potentially be used to measure and verify deepfake technology from various information security defence perspectives.

4.3.1 Liveness Evasion Rate

This metric captures the rate of images or videos passing the action or voice requirements (if applicable) and Presentation Attack Detection. A higher Liveness Evasion Rate means, e.g., with respect to Facial Liveness Verification, a lower security. [67]

4.3.2 Anti-deepfake Evasion Rate

Some cloud vendors, for example, rely on so-called anti-deepfake detection tools and applications. The collected results from these systems can also be shared with the user and thus provide information about how secure these systems are against deepfake attacks. A higher Evasion Rate implies an increased risk of successful deepfake attacks. [67]

4.3.3 Face Matching Rate

A further feasible metric for deepfake detection is the Face Matching Rate. This measures the number of synthesized media passing through a face recognition process. A higher matching rate in this case implies a better quality of the synthesized media and thus a greater chance of being faked by a deepfake. [67]

4.3.4 General Evasion Rate

Evaluating the total security of a target API by measuring the proportion of synthesized media that simultaneously bypasses Facial Liveness Detection, Deepfake Spoofing and Face Matching Detection, an overall Evasion Rate can be used. A higher rate represents a higher attack effectivity or a decreased security of the targeted API. [67]

The example metrics mentioned above represent only a few of many others that might be used for measuring detection and verification methods from an information security perspective. As technical developments in research continue, international standards or technical requirements catalogues for these types of systems may also play a role in the future. [67]

4.4 Deepfake-influenced Areas in Cybercrime and Cyber Warfare

This chapter discusses several relevant cyberattacks and areas important to cyber warfare reaching a new level of severity when supported by deepfake technologies. These subjects are important to understand as they contribute to a more detailed understanding of the future scenarios in chapter ***5. Explorative Scenario Analysis of Deepfakes in the Future***.

4.4.1 Social Engineering

According to the German Federal Office for Information Security (BSI) [70], social engineering is defined as "the **manipulation of people to perform actions or reveal confidential information**". This type of attack is based on exploiting human psychology instead of technical vulnerabilities to gain access to sensitive information or systems or to manipulate individuals in their doing or opinions. Social engineering attacks can take various forms, such as phishing, false pretences, baiting or even physical break-ins. Tactics such as identity theft, deception and emotional manipulation are often used to gain the victim's trust and extract the desired information.

In recent years, social engineering attacks have become more frequent and sophisticated. [70] This is because traditional technical defences such as firewalls or anti-virus software cannot provide a defence against human error or vulnerability. Via research in social networks, websites, telephone numbers, email addresses or through personal contact, a huge amount of data is collected. This method is called information gathering and is one of the most important steps of social engineering.

In addition, social engineering attacks often go undetected for long periods of time, making them a more effective means of conducting attacks. Therefore, it is important to raise awareness of social engineering attacks and train staff and individuals on how to recognise and respond to them. This includes implementing security policies and procedures, regular security awareness training and promoting a culture of security in organisations and society at large. [70]

4.4.2 Phishing

A **subcategory of social engineering** is phishing. Phishing is one of the most popular attack methods used by cybercriminals for stealing confidential data of their victims or distributing ransomware through malicious email links and attachments. However, phishing also happens via other communication channels, such as SMS or phone calls.

Wang et al [71] defines phishing as **luring or tricking a victim into sharing sensitive information**. The term “phishing” is composed of the two English phrases "password" and "fishing". These refer to the targeted fishing of personal information, e.g., passwords or credit card information. The receiver of phishing, for example via email, is usually asked to click on links, reveal personal data or open file attachments to launch possible hidden malware. In general, cyber criminals try to gain as much data as possible or try to remote control the whole computer of their target. [72] distinguishes between generic, regional reference, company reference as well as individual reference phishing attacks. While generic phishing attacks target the wide mass, regional phishing mails are written in regional language and thus target a smaller group. The specific attacking of companies or individuals is called **spear phishing**. Here, as many details as possible are used about the victims to generate the ultimate deception. They are the most dangerous form of phishing specifically exploiting the needs and trust of the victims. [72]

4.4.3 Remote Identity Fraud

Deepfakes can be used for a wide variety of impersonation attacks, such as Catfishing, CEO Fraud, other forms of violation of the privacy of the victim (e.g., used for extortion) or attacks on the integrity if access management systems. **Impersonating** a CEO (Chief Executive Officer) or whale (meaning a "big fish" like Chief Financial Officer or senior executives) as called by [71] enables attackers to extract money from companies. From an attacker point of view there are multiple options for impersonation ranging from emulating the writing style to generating convincing voice or video calls. In literature examples of attempted CEO fraud with deepfake voice messages can be found [73]. Another example of how much damage can be done is the successful CEO fraud using a voice deception, which resulted in a loss of $243,000. [38] Catfishing can be best described as "having relationships with others online under false pretences", whereas the false impression is that the presented identity is fictional. Deepfakes enable this fraud by providing means to fake images, video and even voice on demand, which make them look more authentic thus harder to recognize. [73]

4.4.5 Lethal Autonomous Weapons (LAWs)

Machines endowed with the capacity to dynamically adapt to their surroundings, smoothly shift between passive observation and active engagement modes, autonomously identify, select, and physically engage with their targets, including the deployment of lethal force, all without human intervention. [74]

4.4.6 Quantum Computing

Quantum computing is a computational paradigm that harnesses the principles of quantum mechanics. This scientific discipline explores the quantum properties of specific subatomic particles, like photons and electrons, with the aim of applying this knowledge to perform calculations and conduct extensive-scale information processing. Quantum computers leverage unique quantum phenomena, such as entanglement and superposition, to achieve computational advantages in terms of both speed and cost-effectiveness compared to classical computers. [75]

Special security challenges as fake news, reputational and financial damage or psychological manipulation can pose a massive threat to private individuals and organizations if deepfakes were used for cybercrime or cyber warfare purposes.

For this reason, recommendations for handling deepfakes in the organisational context are compiled in the following chapter, based on the explorative scenario analysis, providing indications and suggestions on how organisations can prepare and protect against deepfake-based attacks in the future.

5. Explorative Scenario Analysis of Deepfakes in the Future

This chapter is dedicated to the methodology of this thesis: the explorative scenario analysis. Here, the steps to successfully implement this analysis method are applied based on [2] [76] [77] [78].

5.1 Explorative Scenario Analysis

Explorative scenario analysis is a methodology used to explore and understand the potential outcomes and implications of uncertain future events or trends. It involves creating and analyzing a range of plausible scenarios to identify risks, opportunities, and potential strategies for decision-making, especially in the future.

Before going into detail and carrying out an explorative scenario analysis in relation to deepfakes, the approach of scenario analysis in the explorative sense is explained in more depth and the individual implementation steps are explained further.

Regarding [2] [76] [77] [78], the methodology of an explorative scenario analysis typically involves the following steps:

- **Defining Scope and Objectives:**
  Clearly defining the focus and purpose of the analysis and identifying the key questions or issues that need to be addressed through the scenario analysis.

- **Identifying relevant Factors and Enablers:**
  Identifying the key factors and enablers that may influence the future outcomes. These could include technological advancements, social trends, economic factors, policy changes or environmental trends.

- **Developing Scenario Stories per Scenario:**
  Creating a set of plausible and consistent stories describing different future scenarios. These scenarios should capture a wide range of uncertainties and potential outcomes.

- **Determining Scenario Dimensions:**
  Identify the most important key factor categories or variables that differentiate the scenarios from one another. These factors could represent critical uncertainties or critical factors that drive the scenarios.

- **Constructing Scenario Funnels:**
  Developing and creating scenario funnels illustrating the different combinations of scenario dimensions. This helps visualizing the range of the scenarios analysis.

- **Assessing Scenario Plausibility and Coherence:**
  Evaluating each scenario's plausibility based on existing evidence, expert opinions or data ensuring that the scenarios are realistic and internally consistent.

- **Analysing Scenario Implications:**
  Assessing the implications of each scenario on the specific objectives or questions defined at the beginning of the analysis. Identifying potential risks, opportunities and challenges associated with each scenario also plays a role.

- **Identifying Strategies:**
  Identifying strategies or actions that are robust across multiple scenarios, meaning they are effective regardless of the specific scenario that unfolds. Moreover, adaptive strategies that can be adjusted as the future unfolds may be considered.

- **Communication and Results:**
  Communicating the Scenario Analysis findings to relevant stakeholders to get feedback and input for refining and improving the analysis afterwards. If needed, iterating the process can help gain new insights or information.

To sum up, the explorative scenario analysis approach is a flexible and iterative process that encourages creative thinking, challenges assumptions, and helps decision-makers anticipate and navigate an uncertain future. It provides a framework for considering multiple plausible futures and aids in strategic planning and decision-making.

5.1.1 Scenarios

The first term to evaluate when it comes to an explorative scenario analysis is, what is a scenario and how can it be classified. In general, it describes an event occurring at a specific point of time in the future. In addition to the possible future situations, several development paths define the circumstances that cause the exact future event. Furthermore, to identify a scenario specific key factors are taken into consideration. [2]

According to [2], the following steps must be considered to define the scope of scenarios:

- Firstly, a scenario is often mistaken with a complete future vision. But it only approaches to specific sections of reality. Factors are consciously included and excluded and put together in different variations.

- The second step to consider is the choice of key factors. Here, the individual factors are included, excluded or put together in certain combinations.

- Every scenario construction assumes what the future might look like. Another part of it is which trends and assumptions stay constant and which will change in the future.

- The most important thing to understand is that scenarios have no claim of truth. Therefore, they are hypothetical futures based on past and present knowledge.

- Finally, scenarios are no forecasts. They do not predict statistically based developments, because of their hypothetical point of view.

5.1.2 Scenario Methodology

The scenario analysis approach essentially distinguishes between three different approaches [2]:

1. **"The future is predictable."**
2. **"The future is evolutionary."**
3. **"The future is changeable."**

The basis of the first perspective depends on the **present and past knowledge**. The more of today's know-how is used, the more reliable a forecast can be. Statistic also plays a big part in this theory. [2]

"**The future is evolutionary**" defines that today's knowledge is too little because the future is something that is not predictable. This theory states that the future is chaotic, uncontrollable, and random. Therefore, it is not possible to prognose or control future situations. Thus, strategies with "intuitive muddling through" aspects are handling prospects. [2]

The third and last **future vision** is the "**changeable**" future. This means the future is neither predictable nor a complete chaos. It is therefore possible to evoke certain future situations with certain actions. Moreover, the actor's objectives and decisions are emphasized. [2]

5.1.3 The Scenario-Funnel Model

Describing the scenario analysis also includes a model to present the key factors correlating to its scenario. One scenario holds one possible future. This model is called scenario-funnel model illustrated in ***Figure 38***:

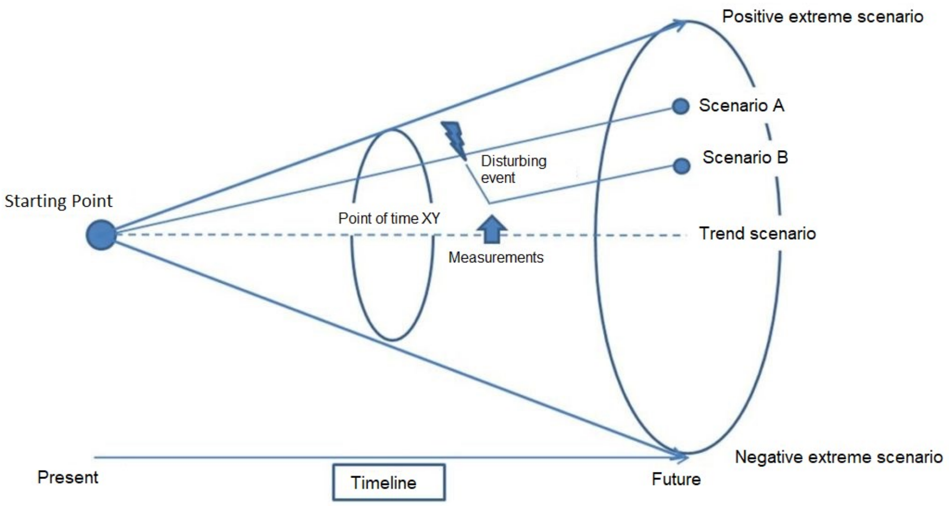


Figure 38: Scenario-Funnel Model Approach based on Fischer Klaus [2]

Per future aspect exists one funnel. It describes the input (key factors) at the starting point and leaves space to interpret the associated future developments. All these funnels connected build the spread for the scenario funnel. [2]

In general, a scenario funnel has three main goal scenarios: **best-case, worst-case and trend scenario.** Beginning at the starting point a key factor occurs with its future developments.

After a certain time, it is possible that a disruptive event occurs, which influences the future development of this scenario. Thus, the previous scenario A can become scenario B. [2]

5.1.4 Explorative Approach

A scenario analysis can be conducted in different ways. This thesis uses an explorative approach to perform the scenario analysis of deepfakes focusing on the impact of cybercrime, society and the media.

The basis for this approach is creating “what-if” questions. These serve to systematize and deepen our current level of know-how. Emphasis is set on the individual development paths, characteristics and interactions of key factors.

5.1.5 Scenario Process

The overall procedure of a scenario analysis can be divided into following five main phases based on [2] [79]:

1. **Phase 1: Problem and Environment Analysis**

   The first part serves to illustrate the topic and the associated problem description. It describes where the limits of the analysis lie, defines the scope and which key factors are not considered.

2. **Phase 2: Identification of Key Factors and Effects Analysis – Opportunities, Risks and Solutions**

   Key factors define variables, parameters, trends or events which build the main part in the third phase – the determination of trends and future events. It is important to have basic knowledge about all key factors and to understand the interdependencies to one another. key factors are often called descriptors.

3. **Phase 3: Key Factor Analysis – Determining Trends and Parameters**

   The next phase includes the creation of the scenario funnel. For each key factor a funnel is created with worst-case, trend and best-case scenarios.

4. **Phase 4: Generating Scenarios – Development of Scenario Phases**

   In this stage the key factors are collected and elaborated to a variety of scenarios at a specific point of time in the future.

5. **Phase 5: Scenario Interpretation – Conclusion and Comparison of Scenarios**

   The final phase is the interpretation and recommendation of the individual scenarios thus achieving the best-case scenario. Furthermore, the differences between the scenarios are also discussed here.

5.2 Phase 1: Scenario Environment of Deepfakes

Firstly, a problem description is explained, which outlines the topic, why and for what purpose this explorative scenario analysis should be created and what added value should result from it. Secondly, an environment analysis narrows down and defines the scope and the timeline of the scenario analysis.

5.2.1 Problem Description

The explorative scenario analysis summarizes the issues related to deepfakes with the evaluation of the impact and the future opportunities on cybercrime, cyber warfare, marketing and education as well as to the insights gained from the detailed state-of-the-art analysis of deepfake detection and verification systems, and mechanisms conducted in chapter ***4.2 State-of-the-Art Analysis of Deepfakes***.

5.2.2 Environment Analysis

The Environment Analysis is structured according to the following three different requirements: The scope, the environment of the analysis scenarios and the timeline in which the analysis predicts the future scenarios.

Scope

The scope of this work is determined by the scientific question in this thesis, focusing on topics around deepfakes and their impact mainly on cybercrime, cyber warfare, marketing and education in the future. Furthermore, the ***4. Preliminaries*** section mainly contains the necessary background information and thus forms an additional baseline on which the scenario analysis is focused on.

The results of the analysis are elaborated potential future scenarios with respective associated opportunities and risks. These are subsequently assigned to the key factor areas in section ***5.3 Phase 2: Key Factor Identification, Analysis and Classification.***

Environment

The scenario stories are primarily viewed through an information technology and information security perspective addressing issues and challenges for organizations and society in the areas of cybercrime, cyber warfare, education and marketing in the future.

For further environmental factors about current technical trends and security related issues of deepfakes the sections ***4.2.6 Deepfake Identification, Detection and Verification***, ***4.2.7 State-of-the-Art Implementations of Deepfake Technologies*** and ***4.4 Deepfake-influenced*** provide more information.

Timeline

Further on, the scenario funnels are formed at intervals from today compared to three and five years from now. So, the timeline is divided into two periods **predicting the future of deepfakes in 2026** and **2028**.

5.3 Phase 2: Key Factor Identification, Analysis and Classification

After the exact specification of the scenario environment, a detailed key factor Identification, Analysis and Classification is used to determine which key factors will influence the further steps of the scenario analysis.

5.3.1 Key Factor Identification

The first step is identifying relevant topics and factors around deepfakes, which potentially impact the future on the environment categories as previously mentioned in phase one, ***5.2.2 Environment Analysis*** of the scenario analysis.

Here, to select the most important factors influencing deepfakes today and in the future, the current technologies, systems and tools as well as the geopolitical, media and societal situation are mainly considered based on previous research in the sections ***4.2 State-of-the-Art Analysis of Deepfakes*** and ***4.4 Deepfake-influenced*** .

Following key factors are considered as relevant for deepfakes:

- Remote Identity Fraud
- Social Engineering
- Phishing
- Blackmailing
- Business models of cybercriminals
- Loss of privacy
- Distorted perception of reality
- Copyright infringement
- Cyber mobbing and harassment
- Data breaches and security incidents
- Propaganda
- Fake news
- Pornography
- War
- Terrorism
- Information monitoring and manipulation
- Detection and verification systems
- Quantum Computing
- (Real-time) Monitoring and Surveillance
- Reputational damage
- Economic damage
- Financial damage
- Psychological damage
- Psychological manipulation
- Regulatory and legal requirements

- Home office
- Remote teaching
- State-of-the-art technology
- Research and development in education
- Educational personnel shortage
- Hyperpersonalization in advertising
- Knowledge transfer
- Awareness
- Customer satisfaction
- Technical and organizational measures
- Entertainment
- Virtual reality and gaming
- Social media
- AI chatbots, e.g. ChatGPT
- Art
- Music
- Film shooting and video production
- Medicine and health care

5.3.2 Key Factor Analysis

Based on the findings of the key factor identification, all mentioned key factors are summarized in following classification categories: **cybercrime, cyber warfare, education** and **marketing**.

Usually, individual scenario funnels are co-determined for the individual key factors. In this case, the detailed table in the next section represents the identified key factors assigned to their classification category and additional environmental factors.

5.3.3 Key Factor Classification

During this key factor analysis and classification process, all individual key factors were evaluated and classified in the following table (***Table 2***), also considering their potential influence on additional environmental factors such as **privacy, society, technology, information security** and **law and politics**.

Privacy [P] relates to private individuals, the confidentiality of one's own personal information and the privacy of individuals. **Society [S]** includes all aspects and issues that may have an impact on society, also internationally. **Technology [T]** is a factor referring to all technical systems and technologies available on the market today, which can also be influenced or further developed by deepfake usage. **Information security [IS]** contains all opportunities and impacts on systems, organizations and individuals to protect and maintain information security goals (confidentiality, integrity and availability). **Law and politics [LP]** refer to the impact related to regulatory and legal requirements, international standards and norms as well as authorities and governmental institutions.

Table 2: Key Factor Classification and Impact on additional Environmental Factors [26]

ID	CLASSIFICATION CATEGORY	DESCRIPTION	INFLUENCING KEY FACTORS	IMPACT ON OTHER ENVIRONMENTAL FACTORS				
				P	S	T	IS	LP
1	**Cybercrime**	*Cybercrime describes a constantly evolving criminal phenomenon in which cybercriminals flexibly adapt to technical, organizational, and social developments and aim to cause damage to their victims or enrich themselves through money, data or reputation. Deepfakes have a major impact on this category.*	• Remote Identity Fraud • Social Engineering • Phishing • Blackmailing • Business models of cybercriminals • Loss of privacy • State-of-the-art technology • Detection and verification systems • Quantum Computing • Distorted perception of reality • Copyright infringement • Cyber mobbing and harassment • Data breaches and security incidents • Reputational damage • Financial damage • Psychological damage	x	x	x	x	x
2	**Cyber Warfare**	*Cyber warfare refers to warfare via the internet using information technology methods and systems. Deepfake technologies have a major impact on this category.*	• Propaganda • Fake news • Geopolitical and situational awareness • Information monitoring and manipulation • State-of-the-art technology • Detection and verification systems • Quantum Computing • (Real-time) Monitoring and Surveillance • Terrorism • Reputational damage • Economic damage • Financial damage • Psychological damage • Psychological manipulation • Regulatory and legal requirements		x	x	x	x
3	**Education**	*Considering deepfake aspects to knowledge transfer and development in the educational sector.*	• Fake news • Awareness • Home office • Remote teaching • Knowledge transfer • Educational personnel shortage • Technical and organizational measures • Research and development in education • Distorted perception of reality • Copyright infringement • Cyber mobbing and harassment	x	x		x	x
4	**Marketing**	*Considering deepfake aspects to marketing, media, communication and advertising.*	• Hyperpersonalization in advertising • Customer satisfaction • Fake news • Distorted perception of reality • Copyright infringement • Cyber mobbing and harassment	x	x			x

The following key factors are excluded from the scenario analysis, as they would deviate significantly from the thesis research question:

- Entertainment
- Pornography
- Virtual reality and gaming
- Social media
- Film shooting and video production
- AI chatbots, e.g. ChatGPT
- Art
- Music
- Medicine and health care

5.4 Phase 3: Scenario Funnels

In the third phase of the explorative scenario analysis, the focus is placed on the actual creation and development of certain future scenarios, which evolve from the previously identified key factors and the current state-of-the-art technology opportunities.

The **tool for implementation** is called **draw.io**, an online tool for drawing and flowchart making. [80] The scenarios created are further explained in the next chapter ***5.5 Phase 4: Scenario Generation and Interpretation***.

5.5 Phase 4: Scenario Generation and Interpretation

In the implementation phase of the explorative scenario analysis, a funnel is generated for each possible future scenario examining it from different perspectives. The scenarios are potential future outlooks separated into the four classification categories resulting from the key factor analysis and classification procedure considering the timeline of three and five years.

The scenario funnels include best-case, worst-case, potential trend scenario and other scenario events. Furthermore, both positively and negatively impacting factors as well as the identified key factors of deepfakes are also considered.

In reference to the research question, the individual created scenario funnels were divided into the four domains**: The Future of Deepfakes in Cybercrime, The Future of Deepfakes in Cyber Warfare, The Future of Deepfakes in Education** and **The Future of Deepfakes in Marketing**. This ensures a clearer overview and result understanding in the final phase of the explorative scenario analysis.

5.5.1 The Future of Deepfakes in Cybercrime

The first area on which the future scenarios of deepfakes concentrates concerns cybercrime. Here, the following future scenarios, including risks and opportunities during scenario interpretation, are examined in more detail.

Deep Remote Identity Fraud (2026)

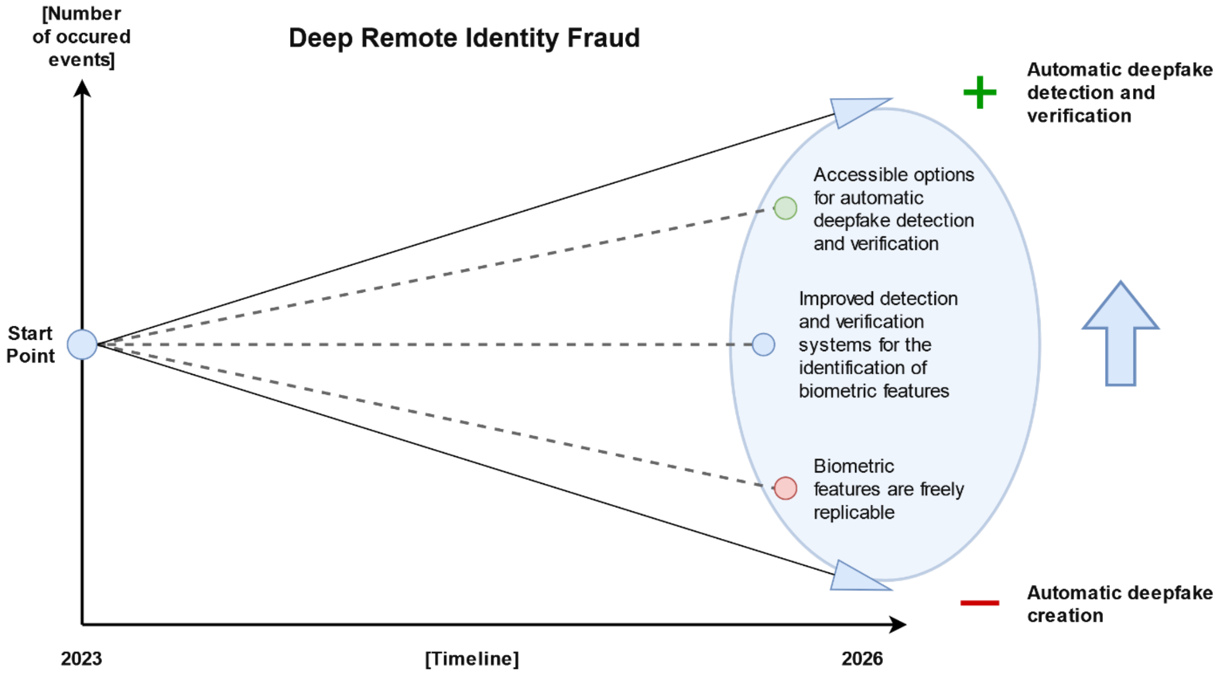
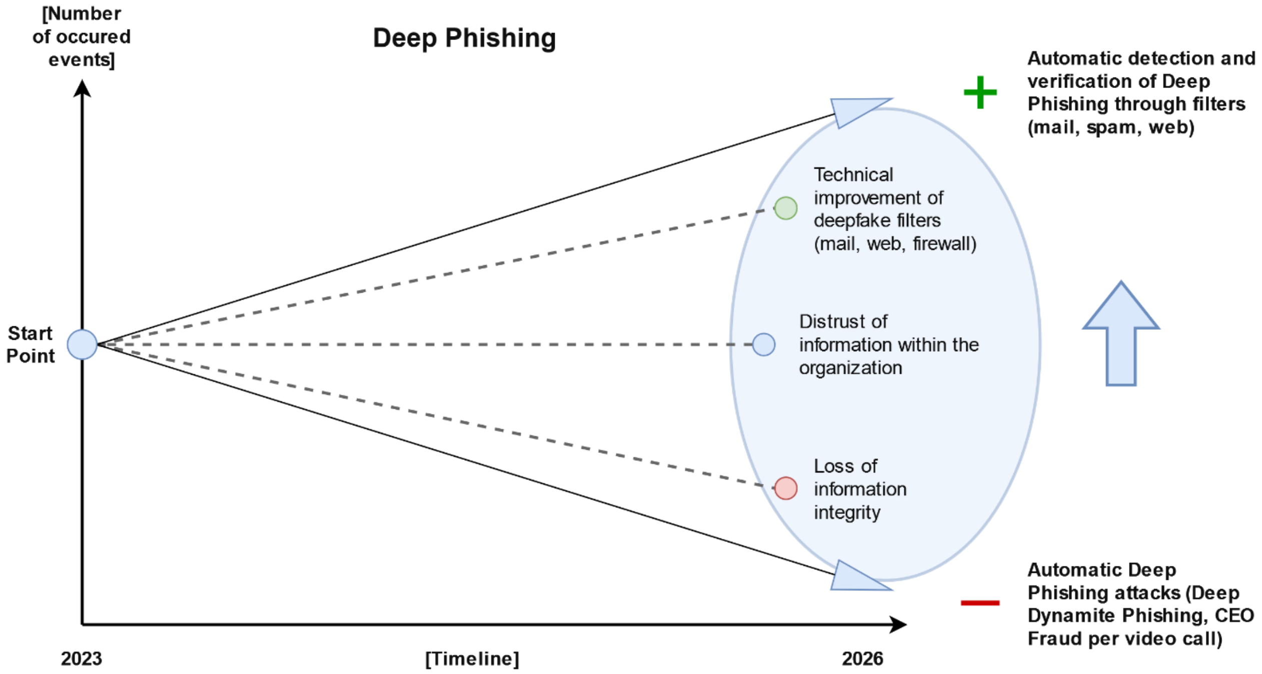

[Number of occured events]
Deep Remote Identity Fraud
+
Automatic deepfake detection and verification
Accessible options for automatic deepfake detection and verification
Start Point
Improved detection and verification systems for the identification of biometric features
Biometric features are freely replicable
−
Automatic deepfake creation
2023
[Timeline]
2026

Figure 39: Deep Remote Identity Fraud [26]

Scenario interpretation of Deep Remote Identity Fraud

Scenario description:	Remote Identity Fraud is one of the techniques used by cybercriminals. The use of deepfake technology provides attackers a greater chance of succeeding in the attack. For example, Open-Source Intelligence (OSINT) could be used to automatically search social media profiles to gather information about the target and enough material about their face, voice, text spellings or behaviour. Based on this, an authentic deepfake could thus be created automatically and without much further effort.
Best-case scenario:	Here, the extremely positive event is the automatic deepfake detection and verification, especially when a Remote Identity Fraud attempt is taking place.
Trend scenario:	The most likely future trend event describes the improvement of deepfake detection and verification systems thus identifying biometric features more efficiently.
Worst-case scenario:	The worst-case scenario includes the automatic creation of deepfakes without any further special requirements needed.
Other scenarios:	Moreover, other occurring scenario events are the breach of biometric security features resulting in free replication by everyone. Another one, but positively influencing, is that accessible options (hardware filters, specific software) for automatic deepfake detection and verification are available for organizations and private individuals.

Deep Phishing (2026)



[Number of occured events]
Deep Phishing
Start Point
Technical improvement of deepfake filters (mail, web, firewall)
Distrust of information within the organization
Loss of information integrity
+
Automatic detection and verification of Deep Phishing through filters (mail, spam, web)
−
Automatic Deep Phishing attacks (Deep Dynamite Phishing, CEO Fraud per video call)
2023
[Timeline]
2026

Figure 40: Deep Phishing [26]

Scenario interpretation of Deep Phishing

Scenario description: Deepfake-enabled phishing attacks make them even more authentic and accurate, with devastating consequences for the victims. For example, cybercriminals might link up immediately after or at events taking place, automatically generating a (live) deepfake and use this for a phishing attack. The information about the events can be found via social media, such as LinkedIn. Also, subcategories of phishing, e.g., dynamite or spear phishing, are more sophisticated.

Best-case scenario: An extremely positive outcome of this scenario is the automatic detection and verification of Deep Phishing attacks through various filters, e.g., web, mail, scam, antivirus or firewall filters.

Trend scenario: A potential trend scenario is the distrust of all available information within an organization leading to further suspicion of employees.

Worst-case scenario: The absolute worst situation that could arise in this scenario is that deepfake-based phishing attacks take place automatically among the broad masses via social media, email, or enterprise communication and collaboration tools such as Microsoft Teams or Skype.

Other scenarios: Another occurring event positively related to Deep Phishing is the technical development and improvement of deepfake filters. On the negative side, it could also lead to a loss of information integrity.

Deepfake Blackmailing (2026)

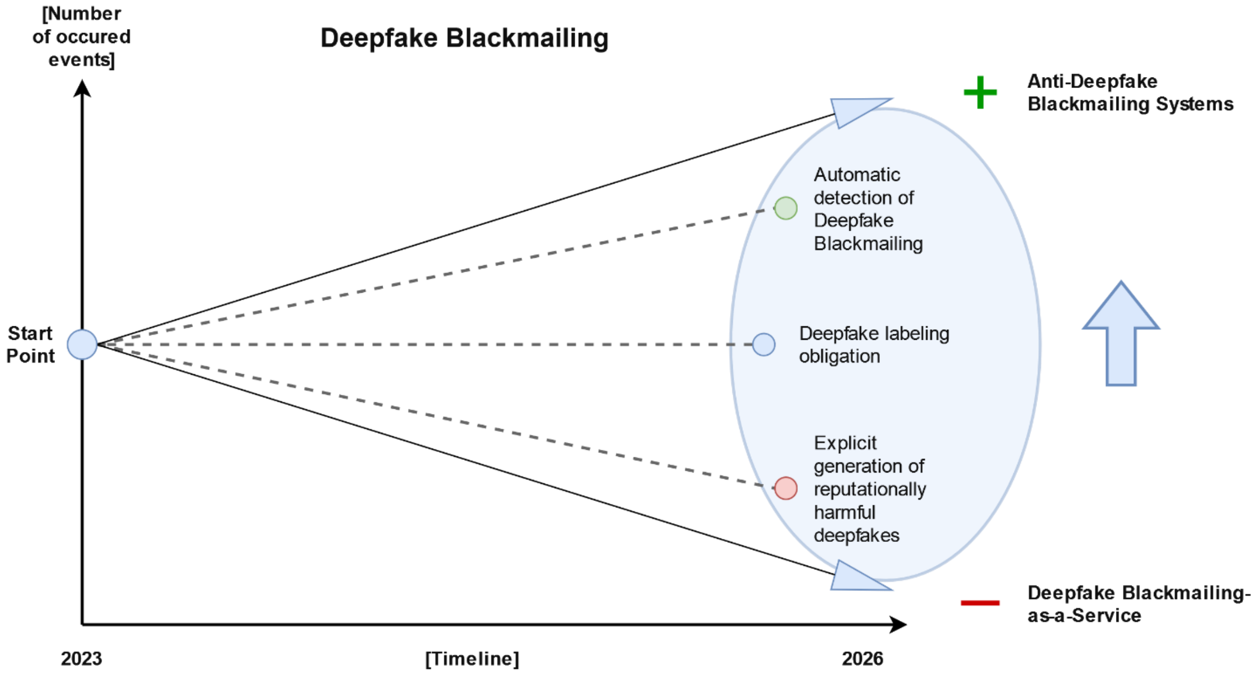

[Number of occured events]
Deepfake Blackmailing
Anti-Deepfake Blackmailing Systems
Automatic detection of Deepfake Blackmailing
Start Point
Deepfake labeling obligation
Explicit generation of reputationally harmful deepfakes
Deepfake Blackmailing-as-a-Service
2023
[Timeline]
2026

Figure 41: Deepfake Blackmailing [26]

Scenario interpretation of Deepfake Blackmailing

Scenario description: When considering classic blackmail attempts, they are already serious enough. However, with the assistance of deepfakes, attempts to blackmail someone are easier and more deceitful.

Best-case scenario: The best-case scenario event here is that systems against deepfake-based blackmailing are available for the public, thus decreasing the success rate of this attacks.

Trend scenario: Most likely, mandatory technical and organizational labelling of deepfakes (in training material or deep learning model) is enforced through regulatory or legal requirements.

Worst-case scenario: An extremely negative outcome is the establishment of deepfake-based Blackmailing-as-a-Service business model of cybercriminals.

Other scenarios: Furthermore, the automatic detection of deepfake blackmailing is an additional positive scenario result. Negatively, authorities or governmental-motivated hacking activists, e.g., Advanced Persistent Threat Groups (APT Groups) could explicitly create reputationally damaging deepfakes.

Deep Social Engineering (2026)

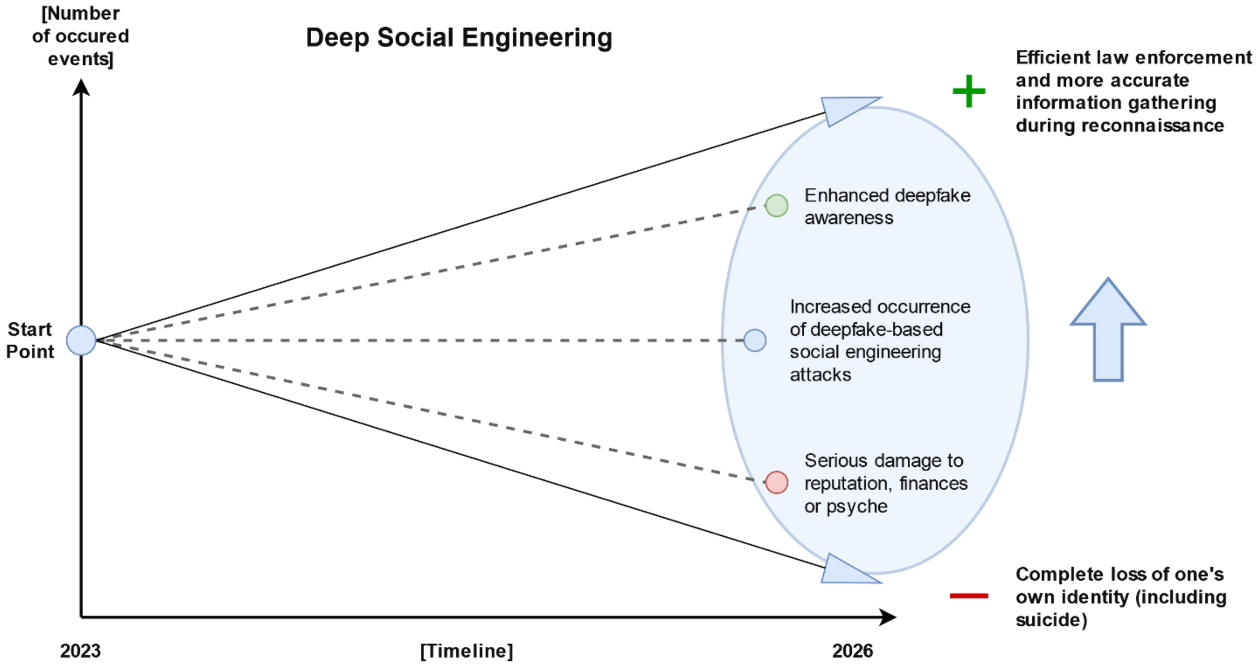

[Number of occured events]
Deep Social Engineering
Efficient law enforcement and more accurate information gathering during reconnaissance
+
Enhanced deepfake awareness
Start Point
Increased occurrence of deepfake-based social engineering attacks
Serious damage to reputation, finances or psyche
Complete loss of one's own identity (including suicide)
−
2023
[Timeline]
2026

Figure 42: Deep Social Engineering [26]

Scenario interpretation of Deep Social Engineering

Scenario description: Social engineering can often have more serious consequences than other attack strategies. Above all, the psychological manipulation factor can cause deepfakes with social engineering to result in the loss of the victim's entire identity. In the worst case, the manipulative toying with the human psyche can even drive the victim to suicide. In addition, the combination with other attack methods, such as CEO fraud, phishing or blackmailing make attacks even easier to execute and increase their probability of succeeding.

Best-case scenario: The best-case scenario leads to efficient law enforcement and even more accurate information gathering during the reconnaissance phase of a cybercrime attempt or hacking attack.

Trend scenario: The most probable trend is the increased occurrence of deepfake-based social engineering attacks.

Worst-case scenario: The complete loss of one’s own identity, which could even lead to suicide due to the psychological and existential damage describes the absolute worst-case scenario outcome.

Other scenarios: Other scenarios are the serious damage to reputation, finances or psyche on the one hand. On the other side, enhanced deepfake awareness helps mitigating the most severe consequences of a social engineering attack.

Deepfake-as-a-Service (DaaS) (2028)

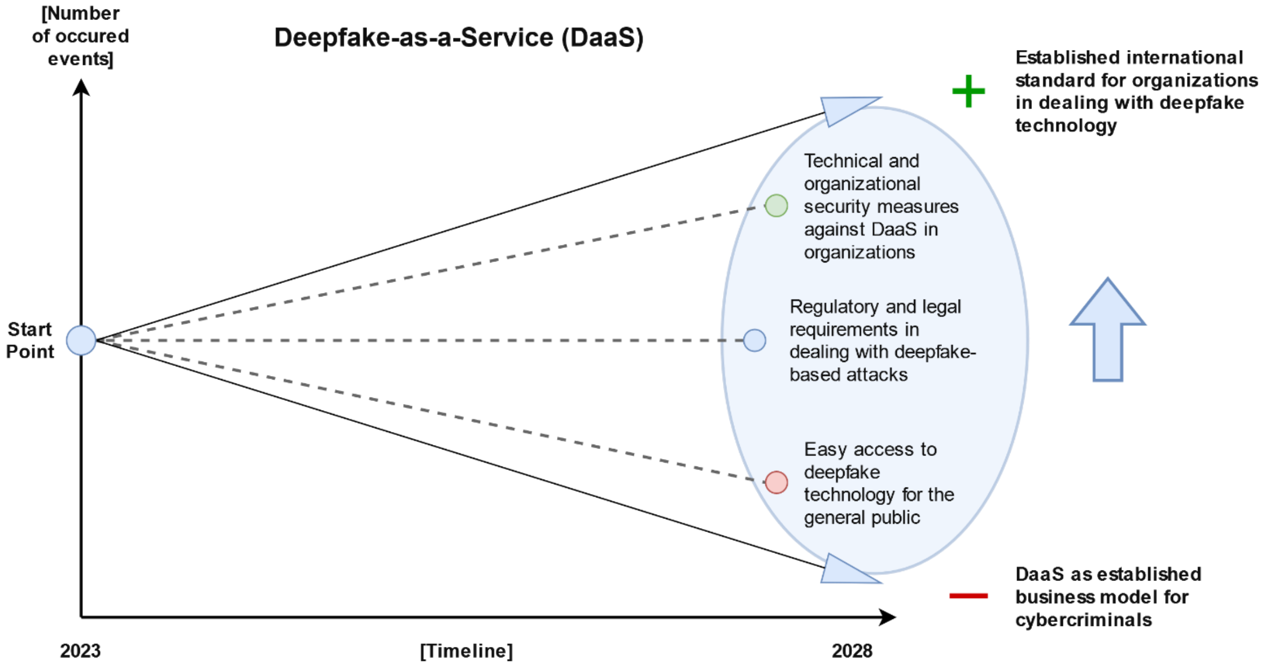

[Number of occured events]
Deepfake-as-a-Service (DaaS)
Established international standard for organizations in dealing with deepfake technology
+
Technical and organizational security measures against DaaS in organizations
Regulatory and legal requirements in dealing with deepfake-based attacks
Easy access to deepfake technology for the general public
Start Point
DaaS as established business model for cybercriminals
−
2023
[Timeline]
2028

Figure 43: Deepfake-as-a-Service [26]

Scenario interpretation of Deepfake-as-a-Service

Scenario description:	Deepfake-as-a-Service, similarly to Cybercrime-as-a-Service, is available as a new business model for cybercriminals with a repertoire of different attack methods and vectors, especially on the dark web. For example, "deepfake" kiddies or people with little or no technical expertise can create deepfakes for their cybercrime objectives by hired attackers, almost like online shopping.
Best-case scenario:	An international standard for organizations dealing with deepfake technology describes the best-case scenario outcome.
Trend scenario:	The creation of regulatory and legal requirements by the state (or even on an international level) in dealing with deepfake-based attacks in organizations shows the trend scenario.
Worst-case scenario:	The established business model Deepfake-as-a-Service for cybercriminals illustrates the worst-case scenario result.
Other scenarios:	According to this future scenario easy access to deepfake technology for the general public and more sophisticated technical an organizational security measures against DaaS are also possible.

Deep Monitoring and Surveillance (2028)

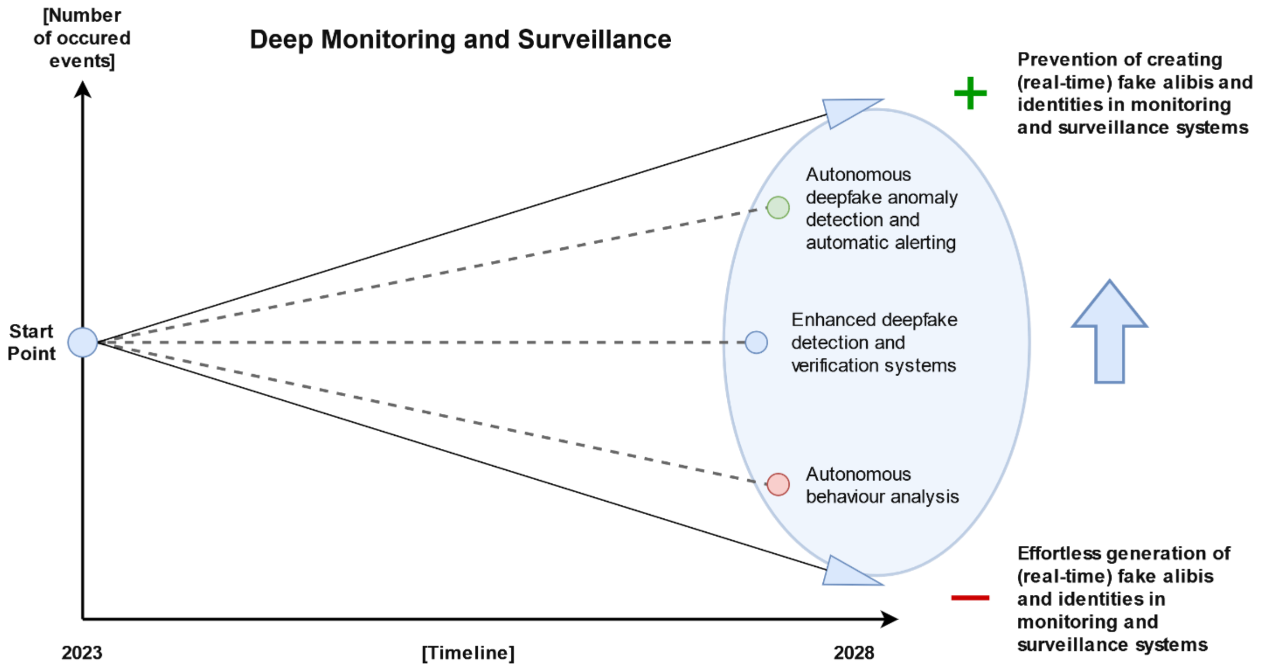

[Number of occured events]
Deep Monitoring and Surveillance
Prevention of creating (real-time) fake alibis and identities in monitoring and surveillance systems
+
Autonomous deepfake anomaly detection and automatic alerting
Start Point
Enhanced deepfake detection and verification systems
Autonomous behaviour analysis
Effortless generation of (real-time) fake alibis and identities in monitoring and surveillance systems
−
2023
[Timeline]
2028

Figure 44: Deep Monitoring and Surveillance [26]

Scenario interpretation of Deep Monitoring and Surveillance

Scenario description:	For monitoring and surveillance systems, deepfakes pose a threat in deceiving security cameras by injecting live deepfakes of cybercriminals to obtain fake identities or provide false alibis when committing a crime. Autonomous behavioral analysis poses another risk and opportunity. Depending on the image source and intention of the individual, the behaviour, facial expressions, gestures, body language or voice of a recorded presentation or live political debate on television can be analyzed. Further on, biometric data for creating deepfakes could be easier accessible thus opening the opportunity for more authentic looking deepfakes.
Best-case scenario:	The best-case result is the prevention of creating (real-time) fake alibis and identities in monitoring and surveillance systems.
Trend scenario:	The trend scenario shows the further enhancement of deepfake detection and verification in monitoring and surveillance systems.
Worst-case scenario:	Attackers can easily generate fake alibis and identities (in real-time) deceiving monitoring and surveillance systems by deepfake technology.
Other scenarios:	Another negative aspect of this scenario is the autonomous behavioural analysis. Moreover, a further positive scenario is the autonomous deepfake anomaly detection including automatic altering in monitoring and surveillance systems.

5.5.2 The Future of Deepfakes in Cyber Warfare

The second domain where deepfakes will have a major impact in the future is cyber warfare.

<u>Deep Lethal Autonomous Weapons (Deep LAWs) (2026)</u>

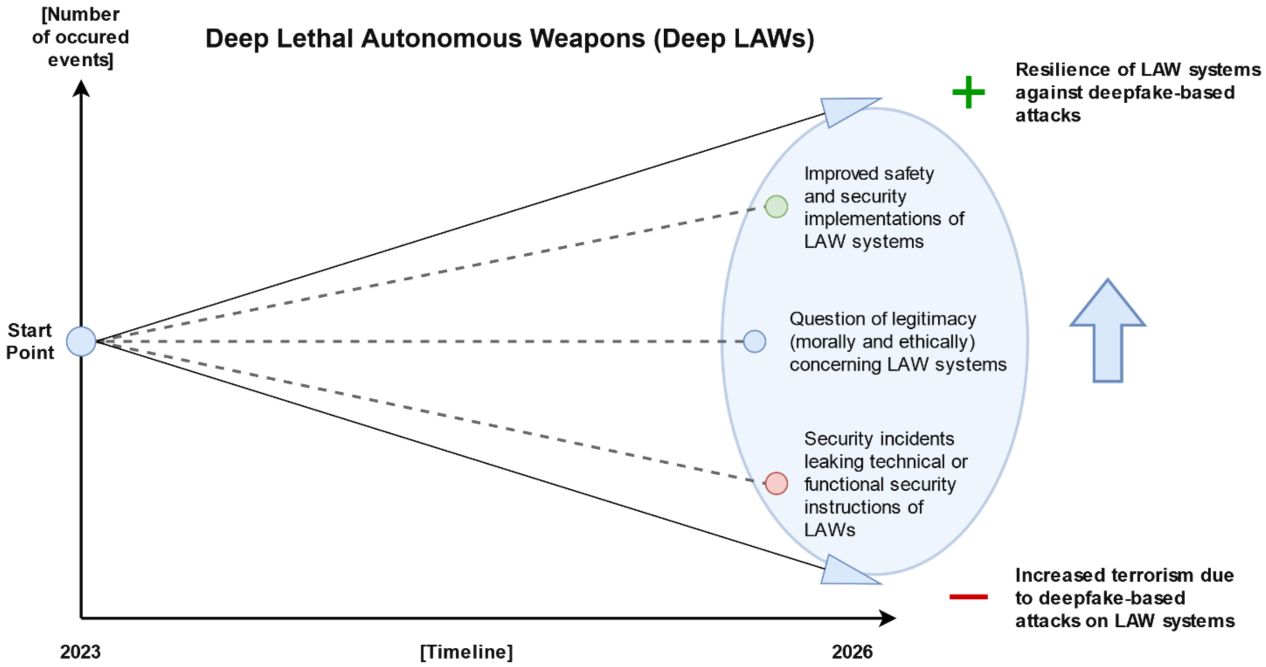

[Number of occured events]
Deep Lethal Autonomous Weapons (Deep LAWs)
+
Resilience of LAW systems against deepfake-based attacks
Improved safety and security implementations of LAW systems
Start Point
Question of legitimacy (morally and ethically) concerning LAW systems
Security incidents leaking technical or functional security instructions of LAWs
—
Increased terrorism due to deepfake-based attacks on LAW systems
2023
[Timeline]
2026

Figure 45: Deep Lethal Autonomous Weapons (Deep LAWs) [26]

<u>Scenario interpretation of Deep Lethal Autonomous Weapons (Deep LAWs)</u>

Scenario description: LAW systems are mainly used for military warfare or surveillance. These weapon systems are available in various versions, e.g., as drones, installed in buildings or on vehicles. AI, "friend-or-foe" recognition and the approval of an individual are used to decide which people or objects are going to be eliminated. Through deepfakes, these systems and the people behind them are fooled, so that friends are now recognized as enemies and vice versa.

Best-case scenario: The best-case here describes a high resilience against cyberattacks, especially against deepfake-based attacks on LAW systems.

Trend scenario: The ethical and moral question of whether such systems should be used continues to be recorded as a future trend.

Worst-case scenario: An extremely negative consequence is the increasing terrorism due to deepfake attacks on LAW systems creating destruction and chaos.

Other scenarios: Furthermore, improved security and safety measures can make LAW systems more resilient to cyberattacks. Security incidents or data breaches revealing important details about the LAW systems functionality are also conceivable.

Deep Quantum Computing (2028)

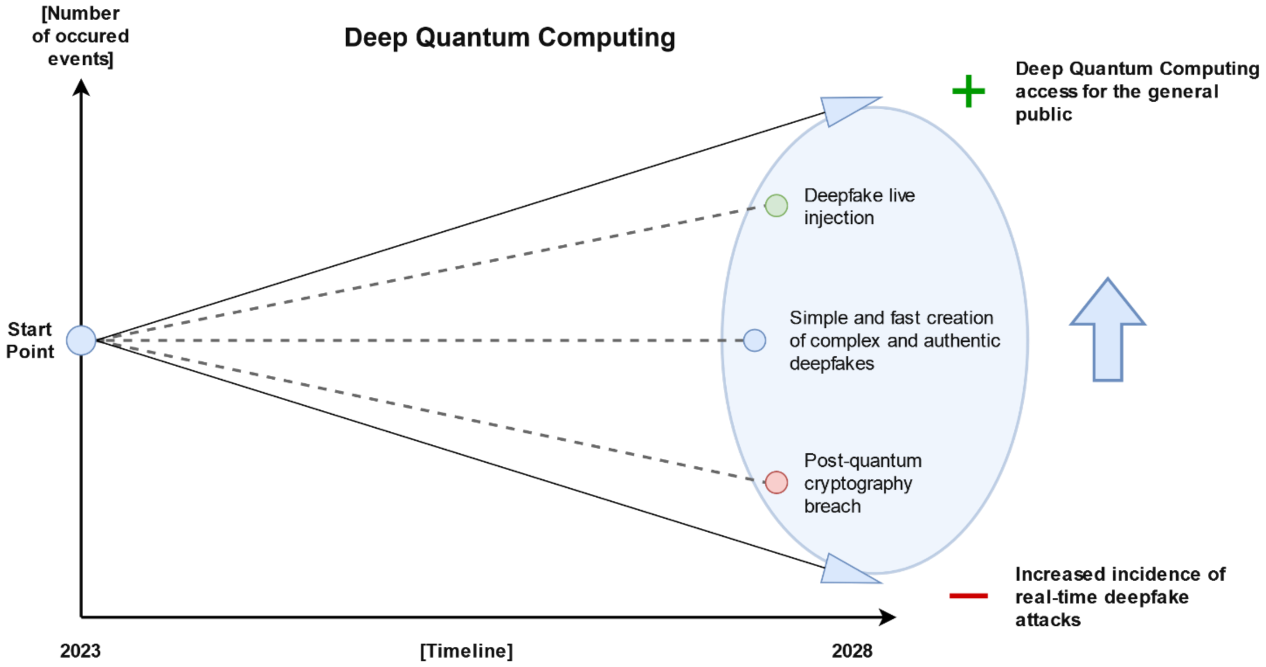

[Number of occured events]
Deep Quantum Computing
+
Deep Quantum Computing access for the general public
Deepfake live injection
Start Point
Simple and fast creation of complex and authentic deepfakes
Post-quantum cryptography breach
−
Increased incidence of real-time deepfake attacks
2023
[Timeline]
2028

Figure 46: Deep Quantum Computing [26]

Scenario interpretation of Deep Quantum Computing

Scenario description: Quantum computers revolutionizing the usage and creation of deepfakes due to their abstract computational methods builds this scenario. Nowadays, computing power and time still pose a considerable obstacle to the generation of especially real-time deepfakes. In combination with quantum computers, it is conceivable that computing power is no longer a problem in the future.

Best-case scenario: The access to deep quantum computing and the underlying algorithms and calculations for the public defines the best-case result of this scenario.

Trend scenario: A trend poses the simple and fast creation of highly complex deepfakes due to the aid of quantum computing technology.

Worst-case scenario: An absolute worst-case is the increased occurrence of real-time deepfake attacks, as computing power no longer pose a hindrance.

Other scenarios: Another chance is the live injection of deepfakes, for example on television, social media or in journalism and reporting. Additionally, the breach of post-quantum cryptography has a devastating consequence for encryption algorithms. Attackers could thus break into encrypted communications and inject their deepfakes, even in real-time.

Deepfake Cyber Information War (2028)

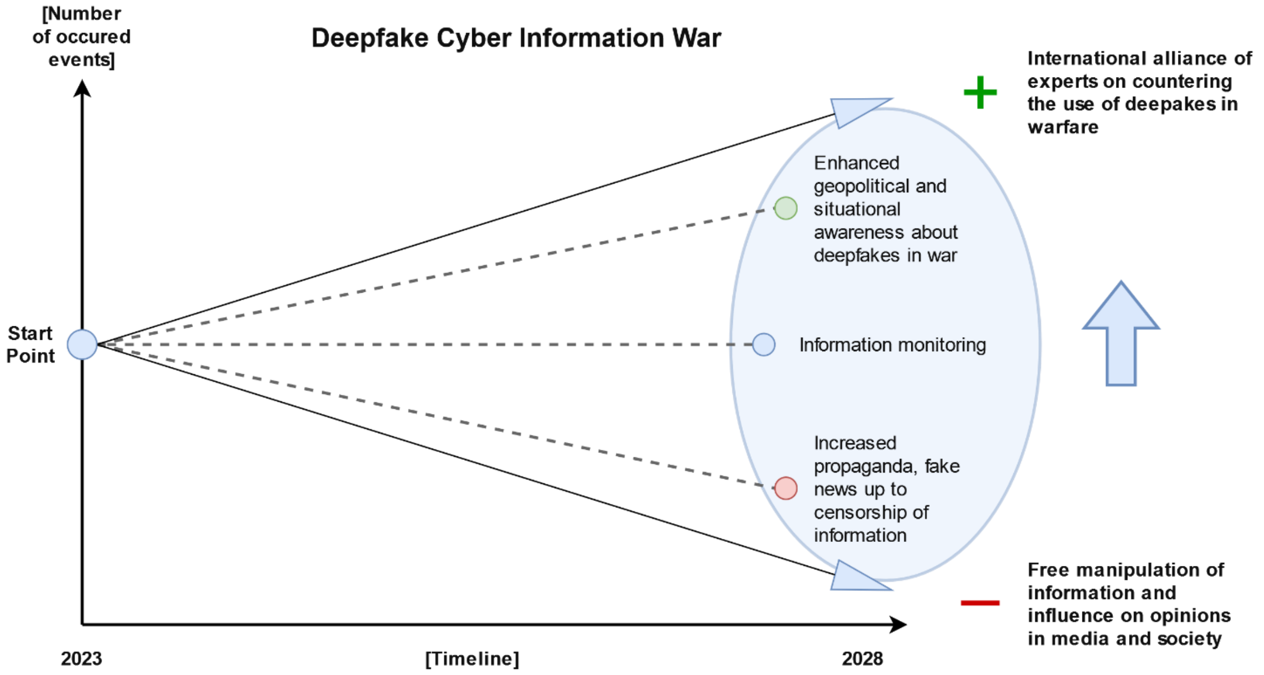

[Number
of occured
events]
Deepfake Cyber Information War
International alliance of experts on countering the use of deepakes in warfare
+
Enhanced geopolitical and situational awareness about deepfakes in war
Start Point
Information monitoring
Increased propaganda, fake news up to censorship of information
Free manipulation of information and influence on opinions in media and society
−
2023
[Timeline]
2028

Figure 47: Deepfake Cyber Information War [26]

Scenario interpretation of Deepfake Cyber Information War

Scenario description: Information wars are highly devastating and characterized by propaganda and fake news. Deepfakes have the effect of also spreading misinformation faster and let the information seem more trustworthily. For instance, a deepfake could be injected during a live news report on war, which then spreads fake news to its viewers. Of course, other information sources like livestreams on social media or television are affected.

Best-case scenario: A best-case opportunity in an information war with deepfake technology is the union of an international experts group counteracting deepfakes and other cyber warfare attack techniques.

Trend scenario: The most likely scenario trend might be the monitoring of all available information, especially during war situations.

Worst-case scenario: The most severe consequence of a cyber information war with deepfakes describes the free manipulation of information at any time from everywhere thus influencing opinions in media and society or leading to censorship.

Other scenarios: As additional positive opportunity, the use of deepfakes during a cyber war leads the broad masses becoming aware of the technology and its consequences. The increased amount of propaganda and fake news is a further risk.

5.5.3 The Future of Deepfakes in Education

In the future, deepfake technologies will not only play a huge role in cybercrime and cyberwarfare. Deepfakes are also affecting domains like education, the exchange of knowledge and information.

<u>Deepfake Educational Personnel (2026)</u>

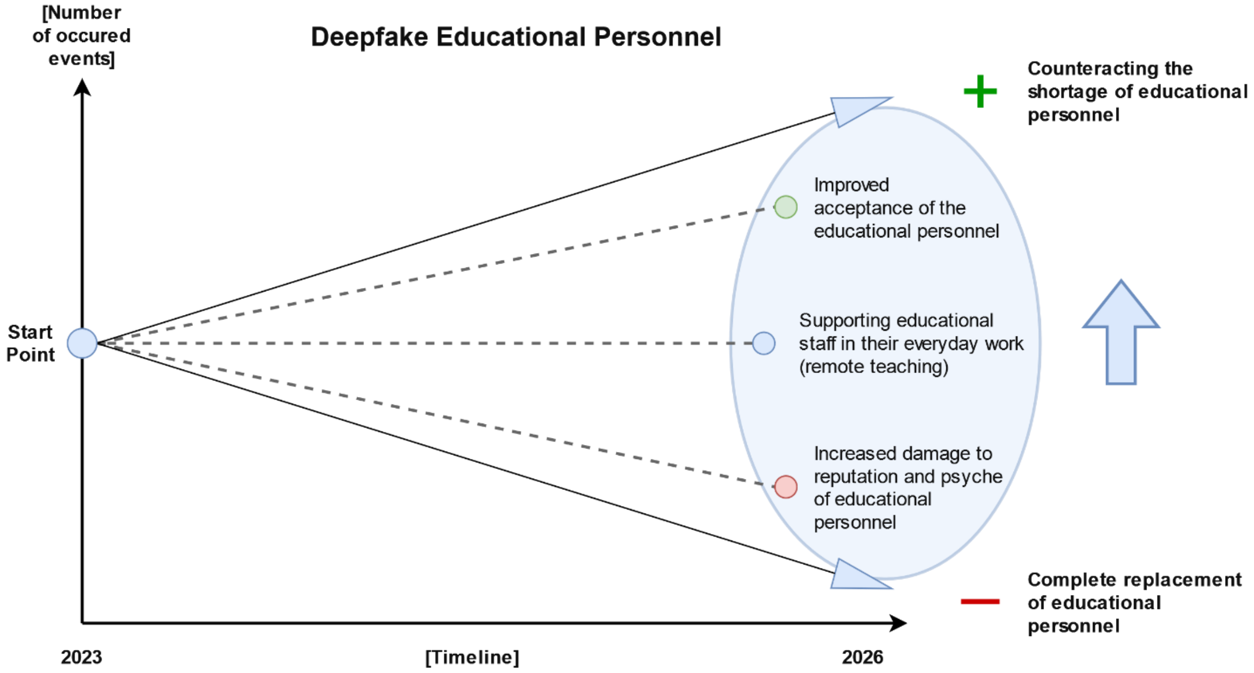

[Number of occured events]
Deepfake Educational Personnel
Counteracting the shortage of educational personnel
Improved acceptance of the educational personnel
Start Point
Supporting educational staff in their everyday work (remote teaching)
Increased damage to reputation and psyche of educational personnel
Complete replacement of educational personnel
2023
[Timeline]
2026

Figure 48: Deepfake Educational Personnel [26]

<u>Scenario interpretation of Deepfake Educational Personnel</u>

Scenario description: Deepfakes are a significantly influencing factor in education and learning. Especially in home offices or during distance learning.

Best-case scenario: The best-case shows teaching with deepfake aid can become a counteract against the educational personnel shortage in the future.

Trend scenario: The support of teaching staff in everyday life by deepfakes or teaching media competence to students is seen as a developing trend.

Worst-case scenario: The absolute worst outcome is the complete replacement of educational staff leading to a rising unemployment rate and unsatisfaction in the educational sector.

Other scenarios: In a negative context, lecturers can also suffer psychological or reputational damage, as the teaching content would be either better or worse received by the students via deepfakes. Additionally, through teaching with deepfakes the acceptance of the educational staff thus the content taught can be better understood.

Deepfake-enriched Information Transfer (2028)

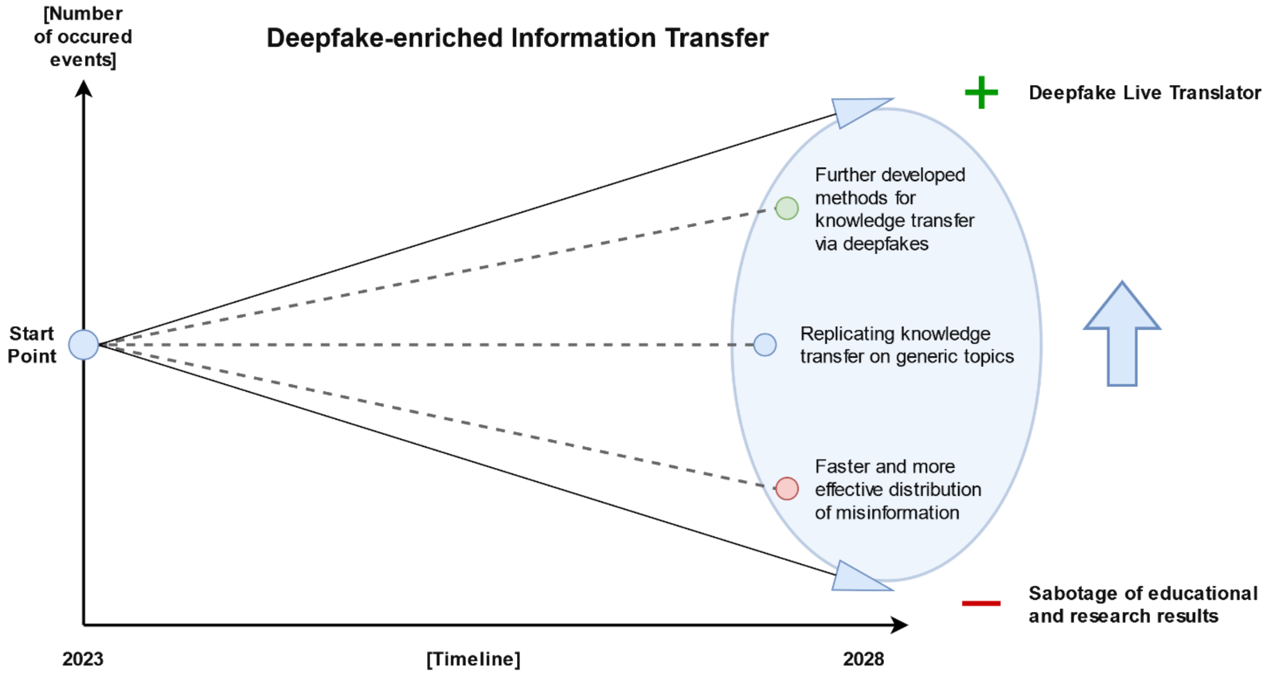

[Number of occured events]
Deepfake-enriched Information Transfer
+
Deepfake Live Translator
Further developed methods for knowledge transfer via deepfakes
Start Point
Replicating knowledge transfer on generic topics
Faster and more effective distribution of misinformation
−
Sabotage of educational and research results
2023
[Timeline]
2028

Figure 49: Deepfake-enriched Information Transfer [26]

Scenario interpretation of Deepfake-enriched Information Transfer

Scenario description: Deepfake technologies are also applied in the transfer of knowledge and the exchange of information. The person presenting is recorded only once. This presentation video can be used as template for further copies in other languages or with other people holding the presentation. This allows more flexibility in sharing knowledge and opens up new possibilities for training courses, workshops or lessons.

Best-case scenario: The best scenario result is the Deepfake Live Translator translating deepfakes in real-time including lip syncing and converting audio. Presentations, livestreams or videos can be translated into other foreign languages in real time, even though the speaker does in fact not speak it.

Trend scenario: The trend in this scenario is the transfer of knowledge on generic topics. For example, for training courses or workshops teaching the same content more than once to their stakeholders.

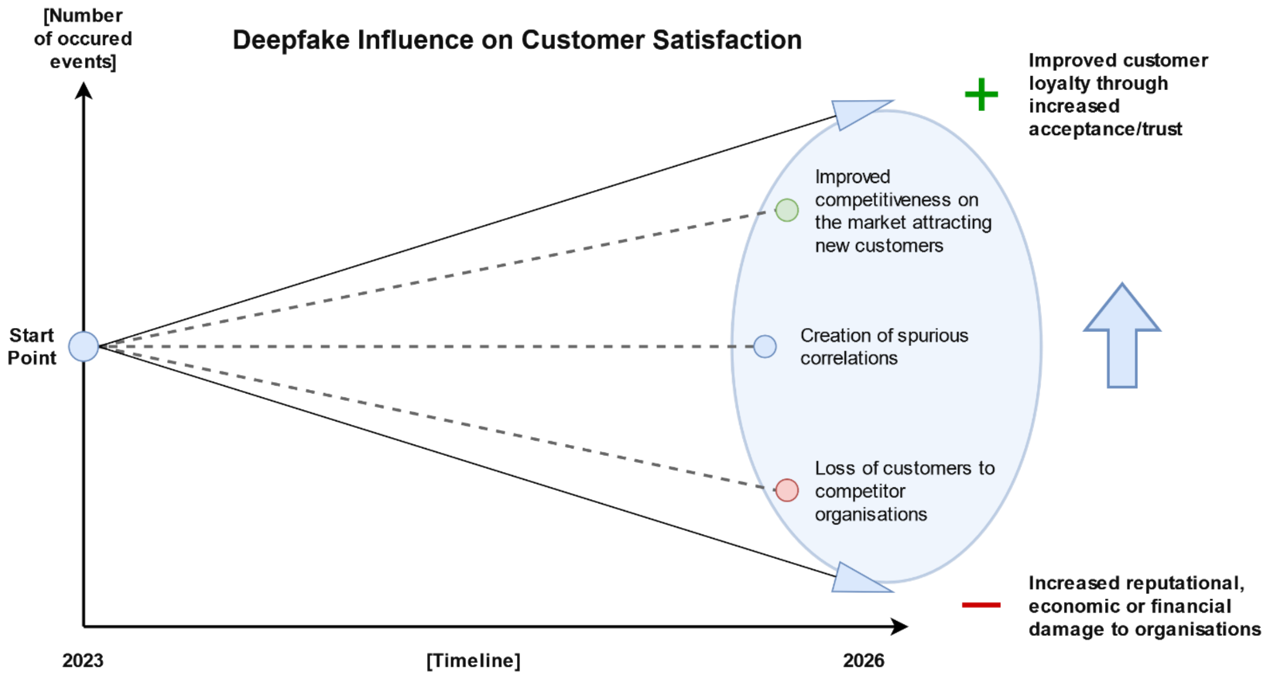
Worst-case scenario: Deepfake attacks on educational or research institutions state a potential risk. Here, the manipulation of research and project results by simplifying the transfer of knowledge is conceivable.

Other scenarios: Moreover, faster and more efficient distribution of misinformation or the development of knowledge transfer methods using deepfakes mentions potential outcomes for deepfake-enriched information transfer.

5.5.4 The Future of Deepfakes in Marketing

The last main sector in which deepfakes pose an important impact factor concerns the sector of marketing, advertising and customer satisfaction.

<u>Deepfake Influence on Customer Satisfaction (2026)</u>



[Number of occured events]
Deepfake Influence on Customer Satisfaction
Improved customer loyalty through increased acceptance/trust
+
Improved competitiveness on the market attracting new customers
Start Point
Creation of spurious correlations
Loss of customers to competitor organisations
–
Increased reputational, economic or financial damage to organisations
2023
[Timeline]
2026

Figure 50: Deepfake Influence on Customer Satisfaction [26]

<u>Scenario interpretation of Deepfake Influence on Customer Satisfaction</u>

Scenario description: Deepfakes in marketing and advertising influence customers and therefore having a massive impact on their satisfaction or opinion to products and services, which can lead to reputational, financial or economical damage of competitor organizations.

Best-case scenario: Through deepfake technology in marketing and advertising customers become more loyal through the higher acceptance and trust in the company and their products or services.

Trend scenario: Every company wants to retain as many customers as possible. Thus, the generation of spurious correlations to damage the competing companies or boost the own organization's reputation represents the trend scenario result.

Worst-case scenario: The worst-case scenario is the reputational, financial or economic damage to organizations through the distribution of deepfakes spreading misinformation of the organization's products, services, organizational culture or employees.

Other scenarios: Losing loyal customers to competitor companies poses another negative scenario result. Furthermore, deepfakes can also lead to an improved competitiveness on the market thus attracting new customers.

Deep Marketing and Hyperpersonalization (2026)

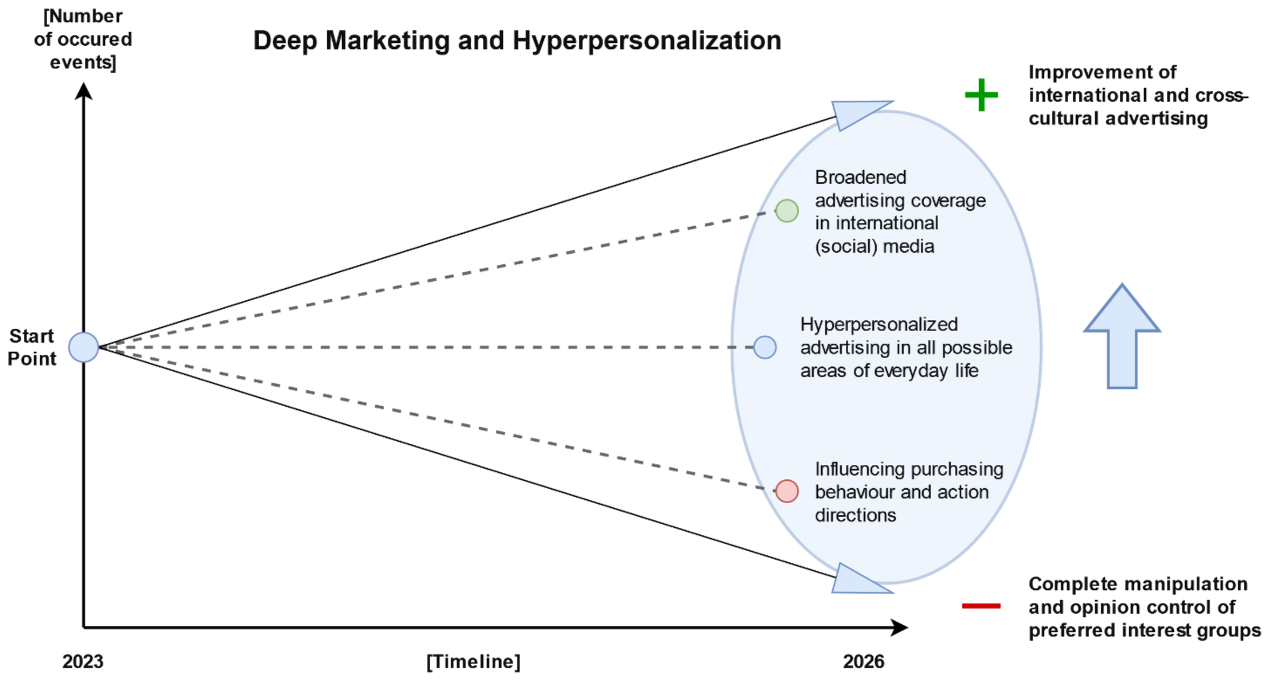

[Number of occured events]
Deep Marketing and Hyperpersonalization
Improvement of international and cross-cultural advertising
Broadened advertising coverage in international (social) media
Start Point
Hyperpersonalized advertising in all possible areas of everyday life
Influencing purchasing behaviour and action directions
Complete manipulation and opinion control of preferred interest groups
2023
[Timeline]
2026

Figure 51: Deep Marketing and Hyperpersonalization [26]

Scenario interpretation of Deep Marketing and Hyperpersonalization

Scenario description: Due to the high credibility of deepfakes, new opportunities arise in the marketing and advertising industry.

Best-case scenario: International and cross-cultural advertising can be further enhanced by (real-time) customization of facial features, hair, clothing or languages, making advertising more attractive to people in other countries.

Trend scenario: The most likely emerging future trend in this scenario is the strong individualization of advertising with the support of deepfakes including all individual and target group preferences and biases.

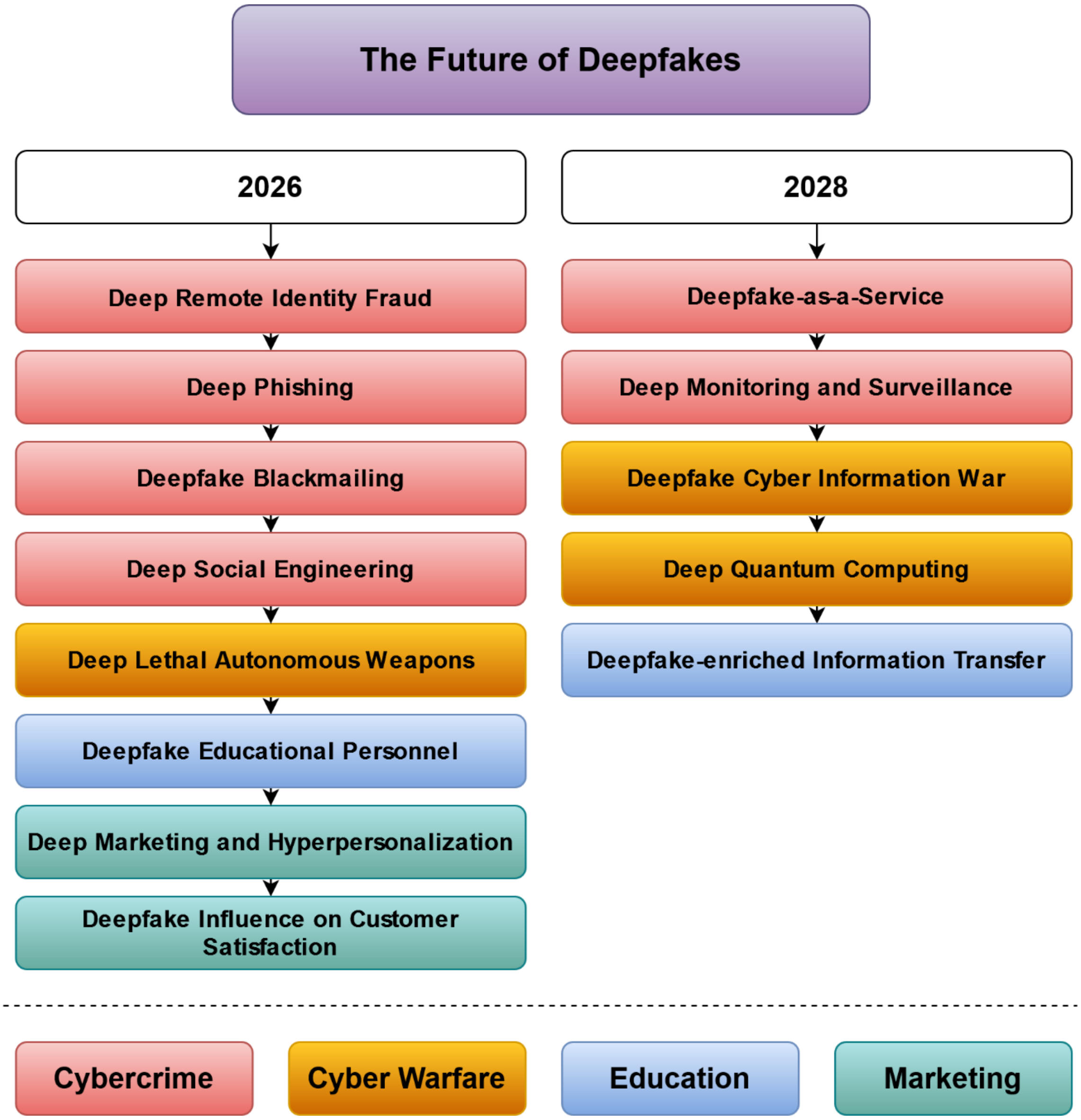
Worst-case scenario: Complete manipulation or control of the target group’s opinions state worst threat.

Other scenarios: Deepfakes improving the coverage of international advertisements, e.g., in social media, and better reflect the requirements of different countries and cultures. Another probable risk is the impact of deepfakes on purchasing behaviour and actions directions.

5.6 Phase 5: Scenario Analysis Results

Finally, the last phase of the explorative scenario analysis represents the results from the previous phases of the **potential impact of deepfakes in the years 2026 and 2028**. As carried out in the previous analysis stages, all built future scenarios are divided into cybercrime, cyber warfare, marketing and education thus reflecting the potential future opportunities and risks of each scenario.

The following figure sums up all scenarios of deepfakes in the future:



The Future of Deepfakes
2026
Deep Remote Identity Fraud
Deep Phishing
Deepfake Blackmailing
Deep Social Engineering
Deep Lethal Autonomous Weapons
Deepfake Educational Personnel
Deep Marketing and Hyperpersonalization
Deepfake Influence on Customer Satisfaction
2028
Deepfake-as-a-Service
Deep Monitoring and Surveillance
Deepfake Cyber Information War
Deep Quantum Computing
Deepfake-enriched Information Transfer
Cybercrime
Cyber Warfare
Education
Marketing

Figure 52: Future Scenarios of Deepfakes in 2026 and 2028 [26]

6. Deepfake Countermeasures and Recommendations for Organizations

Following the explorative scenario analysis, countermeasures and recommendations for business organizations are derived in this chapter of the thesis. A catalogue of technical and organizational recommendations for security measures was compiled. The following chapters describe the scope, target objectives and non-target objectives as well as the retrieved countermeasures.

6.1 Scope, Target and Non-Target Objectives

Scope

The catalogue of recommended countermeasures mainly refers to technical and organisational security measures for business organizations. The recommended measures are described and examined in more detail, primarily from an IT security and information security perspective.

Target Objectives

The catalogue pursues the following target objectives:

- Security recommendations for dealing with deepfakes in the business and corporate context.
- Technical implementation approaches for deepfake detection and verification.
- Recommendations for integrating security measures for dealing with deepfakes into existing organisational structures.

Non-target Objectives

The catalogue pursues the following non-target objectives:

- Security recommendations for individuals or other areas outside the corporate and business context.
- Fully developed processes for integrating the measures into the existing organizational infrastructure.

6.2 Organizational Deepfake Countermeasures and Recommendations

The following table describes the **organizational security measures and recommendations** for organizations in handling deepfakes.

Table 3: Organizational Deepfake Security Measures and Recommendations [26]

ID	ORGANIZATIONAL SECURITY MEASURE	DESCRIPTION
1	**Deepfake policy and requirements**	*Creation of a policy for handling artificial intelligence including deepfake technologies and deriving requirements for the most important business processes in the company. Based on the policy, further security requirements, strategies and measures for internal organisational structures result.*
2	**Integration of deepfake factors into other essential organizational structures**	*Including risks and opportunities of deepfake technologies in the following essential areas of information security management:* • *Information Security Management System* • *Risk and Opportunity Management* • *Business Impact Analyses* • *Incident Management* • *Security Information and Event Management* • *Vulnerability and Patch Management* • *Business Continuity Management* • *Disaster Recovery Management*
3	**Expanding roles and responsibilities**	*Expansion of the roles and responsibilities in the company to include one or more individuals, who explicitly deal with matters relating to AI and deepfakes.*
4	**Expand internal audit and control structures and procedures**	*Extend internal control structures and audit systems to include deepfake technologies and systems used within the organisation.*
5	**Deepfake awareness campaign**	*Enhancing deepfake awareness through regular annual workshops, trainings and deepfake simulations for all employees. At the same time, content related to the use of social media and artificial intelligence is taught.*
6	**Derivation of requirements from laws and regulatory requirements**	*Including legal and regulatory requirements, e.g. from the EU AI Act, into internal organizational structures and procedures.*
7	**Controlling, Documentation and Reporting**	*Inclusion of deepfakes and its security in internal reporting, documentation and controlling structures.*
8	**Deepfake-based security incidents**	*Definition of procedures and processes in the event of deepfake-based cyberattacks. Inclusion in the emergency handbook, which steps must be taken, when and in communication responsible persons.*
9	**Deepfake auditing**	*The systems, tools or methods used in the organisation to detect and verify deepfakes (in real time) are regularly checked for their hardware and software security requirements and patch status.*
10	**Hardening of Deepfake Detection and Verification Systems**	*Enhancement and further development of security implementations, hardening and resilience approaches of deepfake detection and verification systems, applications and technologies in standard and real-time environments of the organization.*
11	**(Real-time) Monitoring and Surveillance Hardening**	*Enhancing resilience on (real-time) monitoring and surveillance systems against deepfake-based cyberattacks.*

12	**Penetration and stress testing of organizational deepfake systems, software and applications**	*Regularly carrying out penetration tests on deepfake technologies used in the company. The security vulnerabilities identified are then remedied during the patch process in Vulnerability Management.*
13	**Deepfake based Cyber Insurances**	*Enterprises have the possibility to insure themselves against cyberattacks through so-called cyber insurances. Here, deepfake-based attacks and the resulting consequences for companies are included.*
14	**Geopolitical and situational awareness**	*Regularly informing about the geopolitical situation and international security incidents related to deepfakes.*

6.3 Technical Deepfake Countermeasures and Recommendations

The following table describes the **technical countermeasures and recommendations** for organizations, which can be implemented in deepfake detection and verification systems to enhance resilience of these systems against deepfake-based attacks.

Table 4: Technical Deepfake Detection and Verification Measures [26]

ID	TECHNICAL SECURITY MEASURE
1	**Coherence Detection using CNNS and LSTM**
2	**Deepfake Detection through Speaker and Speech Recognition**
3	**Lip Language Detection**
4	**Deepfake Labeling**
5	**Image Steganography and Watermarking of media**
6	**Identity Authentication and Verification using Face Swapping and Face Re-enactment**
6	**Eye Tracking, Blinking and Integrity Verification**
7	**Browser Plugins identifying deepfaked media**
8	**Deepfake Filter as Hardware Component**
9	**Artificial Fingerprinting for Generative Models**
10	**Live Anomaly Detection of Deepfakes**
11	**Live Detection of Face and Voice Recognition**

The exact functionality of the individual technical implementations for deepfake detection and verification is described in the state-of-the-art analysis in chapter ***4.2.7 State-of-the-Art Implementations of Deepfake Technologies***.

7. Conclusion and Future Prospects

Taking all mentioned points above into consideration, the following important aspects of this thesis are once again referred to.

In the development of a comprehensive Literature Analysis (see ***3. Literature Analysis***), the knowledge base on the background technologies, methodologies and algorithms of Supervised and Deep Learning was introduced and further explained. This knowledge is an essential prerequisite for understanding the preliminaries section in ***4. Preliminaries***. Starting with the definition and a short historical overview, it continues to the current state of research regarding deepfake technologies in chapter ***4.2 State-of-the-Art Analysis of Deepfakes***. Based on the literature review and the state-of-the-art analysis, potential future scenarios could be created using the scientific method of an explorative scenario analysis (see ***5. Explorative Scenario Analysis of Deepfakes in the Future***), which illustrates the positive as well as negative opportunities and effects of deepfakes in the future years 2026 and 2028. For this time interval, key factors were identified, analyzed and divided into five categories, which can be viewed under ***5.2.2 Environment Analysis*** and ***5.3 Phase 2: Key Factor Identification, Analysis and Classification***. These categories were used for classification purposes when creating the scenario funnels for the respective year in phase 4 of the scenario analysis. In addition, the scenario funnels were further divided into the three main areas of the research question: media impact, societal impact and cybercrime impact. The exact events and effects of the future scenarios created a trend, best-case and worst-case scenario and can be viewed in ***5.5 Phase 4: Scenario Generation and Interpretation***. Following this chapter, all the results of the scenario analysis can be found in ***5.6 Phase 5: Scenario Analysis Results***. Based on the previous scenario analysis results, chapter ***6. Deepfake Countermeasures and Recommendations for Organizations*** mentions feasible security countermeasures in relation to the technical challenges with the deepfake technologies, systems and tools currently available on the market.

The following conclusions can be drawn in connection with how deepfakes can be dealt with in the future:

As mentioned in this thesis, deepfake technology has the potential to change our society and hence has a severe impact on future developments concerning the evolutions of society, economics, the media, cybercrime, cyber warfare, marketing and education, even manipulating the behaviour of people. The desirable trend of the arms race will stay shifted toward the detection of deepfakes in a way that it cannot be used to create propaganda. Deepfakes have **the potential to enhance our privacy** but require an **increased consciousness and awareness** regarding the upload of media depicting ourselves.

Social media will be shaped in foreseeable future by the uprising of TikTok's or Instagram's AI-filter trends, which can be viewed as synonymous with deepfakes. The need for **labelling the usage of deepfakes** just like nowadays marking of advertisements on most social media platforms should be mandatory. Also, the potential to **anonymize non-participating bystanders** and therefore enable freedom is welcomed. **De-identification** usage for deepfakes as a desirable tendency and seeing the normalization of V-Tubing as the final form of De-Identification on social media platforms, like TikTok or Youtube.

According to **cybercrime** deepfakes are already having a vast impact on cyberattacks due to highly realistic fake videos, images, or audios to trick people in an even vicious and more targeted way to deceive and manipulate others. With deepfake technology, an attacker could create a video or audio recording that appears to be of a trusted authority figure, such as a CEO or government official, and use it to deceive employees or the public into taking a certain action or revealing sensitive information. Overall, the use of deepfakes in cybercrime could greatly increase the effectiveness of these attacks by making it harder for individuals to **distinguish between real and fake information** thus increasing the level of trust in the attacker's messages.

Enhancement and further **development** of **security implementations** as well as **hardening and resilience approaches** of deepfake detection and verification systems, applications and technologies in standard and real-time environments will be one key criteria.

In the context of media popularity, headlines, advertising, marketing and the other application areas (see chapter ***4.2.1 Common Application Areas***), deepfakes are gaining more reach than ever, even for people who do not have much contact with technology and research in general. **Deepfake awareness, education in media usage and sensible exchange of information** of further developments are essential for the future to be able to take proactive action against the misuse of this technology by cybercriminals.

New **approaches to identification and authentication systems**, as the biometric factor in technical systems is increasingly exploited as a security measure, should be considered in the future. Implementing **steganographic fingerprinting, watermarking or labelling techniques** into deepfake algorithms or software can help identifying deepfakes and enhance anomaly detection and verification of deepfake technologies. These suggestions for addressing potential countermeasures to deepfake attacks may assist in keeping these associated technologies, methods and approaches more secure and in using them for the **benefit of research and further development** in the future. Deepfakes are not only used for progress in research or for entertainment reasons. Unfortunately, there exist many approaches of deepfakes causing harm and thus largely encourage cybercrime and cyber warfare in the following points:

Deepfake porn currently represents 96% of all deepfakes created and available on the internet. The majority of these are women or celebrities, with an increasing number of deepfake pornographies created purely for hate or revenge. [37] Furthermore, deepfakes not only **damage the reputation** of a person or organisation, but can also be used for **cyberbullying** purposes, **spreading misinformation** about a person, such as committing crimes that were never committed, but leaving the victim responsible for the damage. In combination with social engineering and the psychological manipulation, victims might even be driven to suicide. As with other hacking attacks, deepfake attacks lead to **financial damage**. However, the lack of hardening and implementing security measures in models and tools used to create deepfakes may also allow attackers to exploit them more easily. **Cyber mobbing and harassment**, especially on the internet, often leads to severe consequences and causes **psychological damage** to the victims. [37]

Cybercriminals do not “ask“ their victims in advance whether they are allowed to attack them. The same principle is also followed when using the data and personal information of people available on the Internet. It is becoming more and more common to realize that systems, applications or technologies are susceptible to **security or privacy breaches** not protecting confidential data correctly to prevent such violations.

Deepfakes remain with us in the future and will cause further change, turbulence and progress in the various fields of cybercrime, cyber warfare, marketing, education, healthcare, society, media, music, art, technology and research. However, it takes patience, time and further research to really discover all the possibilities of this trending technology in the future.

References

[1] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor, and A. M. Momani, ‘Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security’, in *2022 International Conference on Cyber Resilience (ICCR)*, Oct. 2022, pp. 1–5. doi: 10.1109/ICCR56254.2022.9995821.

[2] H. Kosow and R. Gaßner, *Methoden der Zukunfts-und Szenarioanalyse Überblick, Bewertung und Auswahlkriterien*. 2008.

[3] P. P. Shinde and S. Shah, ‘A Review of Machine Learning and Deep Learning Applications’, in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Aug. 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697857.

[4] J. Alzubi, A. Nayyar, and A. Kumar, ‘Machine Learning from Theory to Algorithms: An Overview’, *J. Phys. Conf. Ser.*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[5] M. Pirker, ‘MIS PMML Introduction & Basics’, p. 373.

[6] M. G. M. de Souza, ‘Machine Learning: a brief history, definitions, and impacts.’, Medium. Accessed: Dec. 30, 2022. [Online]. Available: https://medium.com/@mateusguil.melo/machine-learning-a-brief-history-definitions-and-impacts-34a4e0739f5c

[7] ‘Deep learning Machine learning Artificial intelligence Data mining ... png for Free Download | DLPNG’. Accessed: May 17, 2022. [Online]. Available: https://dlpng.com/png/6707472

[8] Z. D. R. / Analytics, ‘Dimensionality Reduction (Part 2): Why do it?’, Medium. Accessed: Oct. 09, 2022. [Online]. Available: https://zdfdigital-rd.medium.com/dimensionality-reduction-part-2-why-do-it-89a6abfb96a4

[9] S. Kb, ‘Clustering: If you had to explain it.’, Medium. Accessed: Oct. 09, 2022. [Online]. Available: https://medium.com/@srikanth_kb/clustering-if-you-had-to-explain-it-befe97cb8a24

[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018.

[11] T. Richter, ‘Data Noise and Label Noise in Machine Learning’, Medium. Accessed: Nov. 07, 2022. [Online]. Available: https://towardsdatascience.com/data-noise-and-label-noise-in-machine-learning-98c8a3c8322e

[12] Explorium, ‘Data Bias and What it Means for Your Machine Learning Models’, Explorium. Accessed: Nov. 08, 2022. [Online]. Available: https://exploriumprod.wpengine.com/blog/data-bias-and-what-it-means-for-your-machine-learning-models/

[13] S. A. Metwalli, ‘5 Types of Machine Learning Bias Every Data Scientist Should Know’, Medium. Accessed: Nov. 08, 2022. [Online]. Available: https://towardsdatascience.com/5-types-of-machine-learning-bias-every-data-science-should-know-efab28041d3f

[14] B. Ghojogh and M. Crowley, ‘The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial’. arXiv, May 28, 2019. Accessed: Nov. 08, 2022. [Online]. Available: http://arxiv.org/abs/1905.12787

[15] ‘What Is Data Labelling? - Definition, How It Works & More | Proofpoint UK’, Proofpoint. Accessed: Dec. 16, 2022. [Online]. Available: https://www.proofpoint.com/uk/threat-reference/data-labeling

[16] ‘What is data labeling?’, Amazon Web Services, Inc. Accessed: Dec. 16, 2022. [Online]. Available: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/

[17] J. Bullock, ‘Machine Learning Foundations: Features and Similarity’, Medium. Accessed: Dec. 16, 2022. [Online]. Available: https://towardsdatascience.com/machine-learning-foundations-features-and-similarity-a6ef2901f09f

[18] A. Géron, ‘Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow’, p. 851.

[19] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, ‘A comprehensive survey on support vector machine classification: Applications, challenges and trends’, *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[20] David Foster, *Generative Deep Learning*, 1st ed. O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[21] ‘Alexander Adrowitzer, Thomas Delissen - 2022 - Artificial Intelligence.pdf’.

[22] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[23] Dr. Hassoun, ‘Fundamentals of Artificial Neural Networks’. Accessed: Jan. 07, 2023. [Online]. Available: https://neuron.eng.wayne.edu/tarek/MITbook/chap1/ch1-15.html

[24] A. Dertat, ‘Applied Deep Learning - Part 1: Artificial Neural Networks’, Medium. Accessed: Jan. 09, 2023. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[25] ‘FIG. 3. Commonly used activation functions: (a) Sigmoid, (b) Tanh, (c)...’, ResearchGate. Accessed: Jan. 14, 2023. [Online]. Available: https://www.researchgate.net/figure/Commonly-used-activation-functions-a-Sigmoid-b-Tanh-c-ReLU-and-d-LReLU_fig3_335845675

[26] Schneller Kathrin, ‘Self-created Figures and Graphs’.

[27] C. Mack, ‘Machine learning fundamentals (I): Cost functions and gradient descent’, Medium. Accessed: Jan. 14, 2023. [Online]. Available: https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220

[28] D. Kapil, ‘Stochastic vs Batch Gradient Descent’, Medium. Accessed: Jan. 15, 2023. [Online]. Available: https://medium.com/@divakar_239/stochastic-vs-batch-gradient-descent-8820568eada1

[29] A. Dertat, ‘Applied Deep Learning - Part 3: Autoencoders’, Medium. Accessed: Jan. 09, 2023. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

[30] A. Dertat, ‘Applied Deep Learning - Part 4: Convolutional Neural Networks’, Medium. Accessed: Jan. 09, 2023. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

[31] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, ‘Generative Adversarial Networks: An Overview’, *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[32] A. O. J. Kwok and S. G. M. Koh, ‘Deepfake: a social construction of technology perspective’, *Curr. Issues Tour.*, vol. 24, no. 13, pp. 1798–1802, Jul. 2021, doi: 10.1080/13683500.2020.1738357.

[33] ‘ThisPersonDoesNotExist - Random AI Generated Photos of Fake Persons’. Accessed: May 26, 2023. [Online]. Available: https://this-person-does-not-exist.com/en

[34] ‘thesecatsdonotexist.com/’. Accessed: May 28, 2023. [Online]. Available: https://thesecatsdonotexist.com/

[35] G. Pichler, ‘Ludwig und der falsche Klitschko: Die Deepfake-Gefahr droht real zu werden’, DER STANDARD. Accessed: Jan. 20, 2023. [Online]. Available: https://www.derstandard.at/story/2000136887385/ludwig-und-der-falsche-klitschko-die-deepfake-gefahr-droht-real

[36] ‘“Matrix Resurrections” und Elon Musk: Deutsche Firma arbeitet am “Deepfake-Goldstandard”’, DER STANDARD. Accessed: Jan. 20, 2023. [Online]. Available: https://www.derstandard.at/story/2000141125948/matrix-resurrections-und-elon-musk-deutsche-firma-arbeitet-am-deepfake

[37] J.-T. Kötke, ‘Deepfake - Eine kurze Einleitung’, p. 8.

[38] T. T. Nguyen, Q. V. H. Nguyen, C. M. Nguyen, D. Nguyen, D. T. Nguyen, and S. Nahavandi, ‘Deep Learning for Deepfakes Creation and Detection: A Survey’, *ArXiv190911573 Cs Eess*, Apr. 2021, Accessed: Nov. 10, 2021. [Online]. Available: http://arxiv.org/abs/1909.11573

[39] N. Giansiracusa, ‘Deepfake Deception’, in *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*, N. Giansiracusa, Ed., Berkeley, CA: Apress, 2021, pp. 41–66. doi: 10.1007/978-1-4842-7155-1_3.

[40] R. Delfino, ‘Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act’, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3341593, Feb. 2019. doi: 10.2139/ssrn.3341593.

[41] ‘deepfake papers in Publications - Dimensions’. Accessed: Jun. 18, 2023. [Online]. Available: https://app.dimensions.ai/discover/publication?search_mode=content&search_text=deepfake%20papers&search_type=kws&search_field=full_search

[42] ‘deeptomcruise - YouTube’. Accessed: Jan. 20, 2023. [Online]. Available: https://www.youtube.com/@deepcomtruise

[43] A. Neyaz, A. Kumar, S. Krishnan, J. Placker, and Q. Liu, ‘Security, privacy and steganographic analysis of FaceApp and TikTok’, *Int. J. Comput. Sci. Secur.*, vol. 14, no. 2, pp. 38–59, 2020.

[44] *Varoufakis and the fake finger #varoufake | NEO MAGAZIN ROYALE mit Jan Böhmermann - ZDFneo*, (Mar. 18, 2015). Accessed: Jan. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=Vx-1LQu6mAE

[45] P. Selent, ‘Deepfakes im Marketing: Eine Analyse von Deepfake-Marketing mit exemplarischer Ausarbeitung eines eigenen Deepfake-Modells zur Überprüfung von Effekten bei Rezipienten’, Technische Hochschule Brandenburg, 2022. doi: 10.25933/opus4-2839.

[46] *Fake Obama created using AI video tool - BBC News*, (Jul. 19, 2017). Accessed: Jan. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=AmUC4m6w1wo

[47] *David Beckham speaks nine languages to launch Malaria Must Die Voice Petition*, (Apr. 09, 2019). Accessed: Jan. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=QiiSAvKJIHo

[48] ‘Deepfake presidents used in Russia-Ukraine war’, *BBC News*, Mar. 18, 2022. Accessed: Jan. 20, 2023. [Online]. Available: https://www.bbc.com/news/technology-60780142

[49] M. Holroyd and F. Olorunselu, ‘Deepfake video shows Zelenskyy’s false call for Ukraine to surrender’, euronews. Accessed: Jan. 20, 2023. [Online]. Available: https://www.euronews.com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war

[50] ‘The deepfakes in the disinformation war – DW – 03/18/2022’, dw.com. Accessed: Jan. 20, 2023. [Online]. Available: https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433

[51] iperov, ‘DeepFaceLab’. Jan. 20, 2023. Accessed: Jan. 20, 2023. [Online]. Available: https://github.com/iperov/DeepFaceLab

[52] B. Staff, ‘12 Best Deepfake Apps and Websites You Can Try for Fun’, Beebom. Accessed: Jan. 15, 2023. [Online]. Available: https://beebom.com/best-deepfake-apps-websites/

[53] ‘Faceswap’, Faceswap. Accessed: May 30, 2023. [Online]. Available: https://faceswap.dev/

[54] ‘Reface. Face swap videos’. Accessed: Jun. 18, 2023. [Online]. Available: https://reface.ai/

[55] ‘Lensa - Prisma Labs’. Accessed: Jun. 18, 2023. [Online]. Available: https://prisma-ai.com/lensa

[56] ‘Deepswap - AI Face Swap Online für Gesichtertausch Video’, DeepSwap. Accessed: Jan. 20, 2023. [Online]. Available: https://www.deepswap.ai/de

[57] ‘Make Your Own Deepfakes [Online App]’, Deepfakes Web. Accessed: Jan. 20, 2023. [Online]. Available: https://deepfakesweb.com/

[58] S. B, ‘This AI Lip Sync App Can Make Your Favorite Character Sing & It’s Hilarious’, Beebom. Accessed: Jan. 20, 2023. [Online]. Available: https://beebom.com/ai-lip-sync-app-wombo-make-your-favorite-character-sing/

[59]‘AI Time Machine^{TM}: Erstellen Sie AI-Avatare & reisen Sie durch die Zeit’, MyHeritage. Accessed: Jun. 18, 2023. [Online]. Available: https://www.myheritage.at/ai-time-machine

[60]‘Deep Art Effects: Sei ein Künstler. Verwandle Bilder in atemberaubende Kunstwerke’. Accessed: Jun. 18, 2023. [Online]. Available: https://www.deeparteffects.com/

[61]‘FaceApp: Face Editor’. Accessed: Jun. 18, 2023. [Online]. Available: https://faceapp.com/

[62]‘DeepFaceLive’, Easy With AI. Accessed: May 30, 2023. [Online]. Available: https://easywithai.com/ai-video-tools/deepfacelive/

[63]‘Avatarify — Bring your photos to life’. Accessed: Jun. 18, 2023. [Online]. Available: https://avatarify.ai

[64]R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, ‘Deepfakes and beyond: A Survey of face manipulation and fake detection’, *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014.

[65]T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, ‘Analyzing and Improving the Image Quality of StyleGAN’. arXiv, Mar. 23, 2020. Accessed: Jan. 15, 2023. [Online]. Available: http://arxiv.org/abs/1912.04958

[66]Y. Cheng *et al.*, ‘GRACE: Loss-Resilient Real-Time Video Communication Using Data-Scalable Autoencoder’. arXiv, Oct. 29, 2022. doi: 10.48550/arXiv.2210.16639.

[67]C. Li *et al.*, ‘Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era’, presented at the 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 2673–2690. Accessed: May 30, 2023. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/li-changjiang

[68]N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, ‘Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data’, presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14448–14457. Accessed: May 30, 2023. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Yu_Artificial_Fingerprinting_for_Generative_Models_Rooting_Deepfake_Attribution_in_Training_ICCV_2021_paper.html

[69]T. Jung, S. Kim, and K. Kim, ‘DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern’, *IEEE Access*, vol. 8, pp. 83144–83154, 2020, doi: 10.1109/ACCESS.2020.2988660.

[70]‘Social Engineering – der Mensch als Schwachstelle’, Bundesamt für Sicherheit in der Informationstechnik. Accessed: Jul. 08, 2023. [Online]. Available: https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Cyber-Sicherheitslage/Methoden-der-Cyber-Kriminalitaet/Social-Engineering/social_engineering.html?nn=132176

[71]Z. Wang, L. Sun, and H. Zhu, ‘Defining Social Engineering in Cybersecurity’, *IEEE Access*, vol. 8, pp. 85094–85115, 2020, doi: 10.1109/ACCESS.2020.2992807.

[72]A. Keser, ‘Phishing - die beliebteste Waffe der Cyberkriminellen’, Confare. Accessed: Aug. 05, 2020. [Online]. Available: https://confare.at/phishing-die-waffe-der-cyberkriminellen/

[73] M. P. Nolan, ‘Learning to circumvent the limitations of the written-self: The rhetorical benefits of poetic fragmentation and internet “catfishing”’, *Pers. Stud.*, vol. 1, no. 1, pp. 53–64, doi: 10.3316/informit.967696119864418.

[74] M. Taddeo and A. Blanchard, ‘A Comparative Analysis of the Definitions of Autonomous Weapons Systems’, *Sci. Eng. Ethics*, vol. 28, no. 5, p. 37, Aug. 2022, doi: 10.1007/s11948-022-00392-3.

[75] D. P. García, J. Cruz-Benito, and F. J. García-Peñalvo, ‘Systematic Literature Review: Quantum Machine Learning and its applications’. arXiv, Jan. 11, 2022. doi: 10.48550/arXiv.2201.04093.

[76] Y. Tourki, J. Keisler, and I. Linkov, ‘Scenario analysis: a review of methods and applications for engineering and environmental systems’, *Environ. Syst. Decis.*, vol. 33, no. 1, pp. 3–20, Mar. 2013, doi: 10.1007/s10669-013-9437-6.

[77] K. Fischer, ‘Blick in die Zukunft – Arbeiten mit Szenario-Techniken’, Coverdale Austria. Accessed: Aug. 07, 2020. [Online]. Available: https://www.coverdale.at/blick-in-die-zukunft/

[78] P. N. Duinker and L. A. Greig, ‘Scenario analysis in environmental impact assessment: Improving explorations of the future’, *Environ. Impact Assess. Rev.*, vol. 27, no. 3, pp. 206–219, Apr. 2007, doi: 10.1016/j.eiar.2006.11.001.

[79] ‘Die Szenariotechnik - Methode, Schritte, Tipps’, Christian H. Meyer. Accessed: Aug. 02, 2020. [Online]. Available: https://www.christianhmeyer.de/die-szenariotechnik-methode-schritte-tipps/

[80] ‘draw.io’. Accessed: Sep. 09, 2023. [Online]. Available: https://app.diagrams.net/