Law, Governance and Technology Series 58

Henrique Sousa Antunes ·
Pedro Miguel Freitas · Arlindo L. Oliveira ·
Clara Martins Pereira · Elsa Vaz de Sequeira ·
Luís Barreto Xavier *Editors*

Multidisciplinary Perspectives on Artificial Intelligence and the Law

OPEN ACCESS

Springer

Law, Governance and Technology Series

Volume 58

Series Editors

Pompeu Casanovas, UAB, Institute of Law and Technology UAB, Barcelona, Spain

Giovanni Sartor, University of Bologna and European University Institute of Florence, Florence, Italy

The *Law, Governance and Technology Series* is intended to attract manuscripts arising from an interdisciplinary approach in law, artificial intelligence and information technologies. The idea is to bridge the gap between research in IT law and IT-applications for lawyers developing a unifying techno-legal perspective. The series will welcome proposals that have a fairly specific focus on problems or projects that will lead to innovative research charting the course for new interdisciplinary developments in law, legal theory, and law and society research as well as in computer technologies, artificial intelligence and cognitive sciences. In broad strokes, manuscripts for this series may be mainly located in the fields of the Internet law (data protection, intellectual property, Internet rights, etc.), Computational models of the legal contents and legal reasoning, Legal Information Retrieval, Electronic Data Discovery, Collaborative Tools (e.g. Online Dispute Resolution platforms), Metadata and XML Technologies (for Semantic Web Services), Technologies in Courtrooms and Judicial Offices (E-Court), Technologies for Governments and Administrations (E-Government), Legal Multimedia, and Legal Electronic Institutions (Multi-Agent Systems and Artificial Societies).

Henrique Sousa Antunes • Pedro Miguel Freitas •
Arlindo L. Oliveira • Clara Martins Pereira •
Elsa Vaz de Sequeira • Luís Barreto Xavier
Editors

Multidisciplinary Perspectives on Artificial Intelligence and the Law

Springer

CATÓLICA
RESEARCH CENTRE
FOR THE FUTURE OF LAW
LISBON

Plataforma
de Apoio à
Investigação
em Direito
na Católica

Editors
Henrique Sousa Antunes
Faculty of Law
Universidade Católica Portuguesa
Lisbon, Portugal

Arlindo L. Oliveira
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal

Elsa Vaz de Sequeira
Faculty of Law
Universidade Católica Portuguesa
Lisbon, Portugal

Pedro Miguel Freitas
Faculty of Law
Universidade Católica Portuguesa
Porto, Portugal

Clara Martins Pereira
Durham Law School
Durham, UK

Luís Barreto Xavier
Faculty of Law
Universidade Católica Portuguesa
Lisbon, Portugal

ISSN 2352-1902 ISSN 2352-1910 (electronic)
Law, Governance and Technology Series
ISBN 978-3-031-41263-9 ISBN 978-3-031-41264-6 (eBook)
https://doi.org/10.1007/978-3-031-41264-6

This work was supported by PAIDC - Plataforma de Apoio à Investigação em Direito na Católica

© The Editor(s) (if applicable) and The Author(s) 2024. This is an open access publication.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

The Relevance of Deepfakes in the Administration of Criminal Justice

Dalila Durães, Pedro Miguel Freitas, and Paulo Novais

Abstract Nowadays, it is challenging to distinguish between genuine content created by humans or deepfake created by deepfakes algorithms. Therefore, it is in the interests of society and nations to have systems that can notice and evaluate the content without human intervention. This paper presents the challenges of artificial intelligence, specifically machine learning and deep learning, in the fight against deepfake. In addition, it presents the relevance that deepfakes may have in the administration of criminal justice.

1 Introduction

Deepfakes appeared in late 2017 when an anonymous user on the social network Reddit published videos where pornographic actresses' faces were replaced with celebrities' faces (Europol 2020). The potential of this technique quickly disseminated on the Internet, making it accessible for everyone (Anderson 2018).

Deepfake is one of the most visible malicious uses of artificial intelligence (AI). The name derives from the combination of "deep learning" and "fake media". To create Deepfakes, AI techniques, particularly machine learning (ML), are used to create or manipulate content, which may be audio or video. Content that might be extremely hard for humans or technological solutions alike to distinguish from authentic one (Chawla 2019).

D. Durães (✉) · P. Novais
Algoritmi Centre, School of Engineering, University of Minho, Braga, Portugal
e-mail: dalila.duraes@algoritmi.uminho.pt; pjon@di.uminho.pt

P. M. Freitas
Universidade Católica Portuguesa, Faculty of Law, Porto, Portugal
e-mail: pfreitas@ucp.pt

© The Author(s) 2024

H. Sousa Antunes et al. (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, Law, Governance and Technology Series 58,
https://doi.org/10.1007/978-3-031-41264-6_19

2 Deepfake: Definition and Categories

A deepfake is content (video, audio or otherwise) that was either fully or partially fabricated or manipulated from existing content (video, audio or otherwise). However, the most significant cases of deepfake appear in the video format, whose authentication difficulty allows any information since any audiovisual content can be manufactured. Deepfakes emerged from AI applications that combine, mixture, replace or overlay images and videos, creating fake videos so that they look authentic (Maras and Alexandrou 2019; Europol 2020). Deepfake is a clear example of the intricacy and complexity technology used. However, the available applications allow anyone with low computing power and little knowledge to create fake videos (Figueira and Oliveira 2017).

Deepfakes can be legally created to engage or spark critical observations. Nevertheless, deepfakes are also used to commit fraud, deceive, or intimidate people by releasing images or videos without their consent.

There are different categories of deepfakes: face replacement or body-swapping, face reconstruction, face generation, speech synthesis, and shallowfakes. Face replacement or body-swapping substitutes parts of a person for another. Face reconstruction manipulates a part of a person to assemble it seem as if they are expressing something they are not. Face generation lets create synthetic images of convincing but entirely fictional people. Speech synthesis employs training algorithms to generate a deepfake voice or an artificial audio file. Shallowfakes allow the creation of audiovisual frauds by using elementary or basic editing methods (Europol 2020). Table 1 present examples of different deepfakes.

3 AI and Deepfake

As we mentioned earlier, deepfakes are created using AI. But how?

One of the great advantages of AI is that it absorbs a large amount of knowledge from the environment in which it is inserted, learning, and improving its responses day by day.

To better understand the concept of AI, it can be divided into two categories: Artificial Strong Intelligence and Artificial Narrow Intelligence.

Artificial Strong Intelligence includes systems that exhibit human intelligence or even superior in all fields. Furthermore, this type of intelligence can share experience from different domains. Several tests are used to indicate if a given system shows Strong AI, but this has not yet happened, what has led experts to suggest that this is merely an aspiration (Muehlhauser 2013).

Narrow AI or Weak AI includes all systems where there are well-defined right or wrong answers, where there are discernible underlying patterns and structures, and where research and computing speed offer advantages over humans. The existing AI systems are not yet designed to apply abstract reasoning, understand concepts,

Table 1 Types and examples of deepfakes (Adapted from (Kietzmann et al. 2020)

Type	Format	Description	Business application
Face replacement or body-swapping	Photo deepfakes	Face and Body-swapping—replacing face or body for someone else.	Cosmetics, eyeglasses, hairstyles or clothes virtuality.
	Video deepfakes	Face-swapping—replacing face to someone else.	Movies, where actor's face are put on body of double.
Face reconstruction (body)	Audio and video deepfakes	Full body puppetry—transporting the body movement from one person to another.	Enterprise directors and athletes can camouflage physical aliments during a video exhibition.
		Lip-synchronism—adjusting the mouse activities and phrases articulated in a talking video.	Institutional videos can be converted into other speeches employing the same spokesperson in the original recording.
Shallowfakes	Video deepfakes	Face-morphing—a face changes into another face through a seamless transition.	Video game participants can introduce their looks to their favorite personalities.
Speech synthesis	Audio deepfakes	Voice-swapping—changing a spokesperson or imitating somebody else.	The spokesperson of narration can be like a young, old, male, or female.
		Text-to-Speech - modifying the audio by writing new content.	Replacing words without the necessity of making a new audio file.
Face generation	Photo deepfakes	Face-generating—generate realistic looking faces.	This can serve as a solution to dispose of photographs without violating the privacy or image rights of any person, since everything has been artificially generated, so no one has been photographed.

or general broad-spectrum problem-solving skills (Krupansky 2017). So, systems in this category have narrow and limited application to solve specific problems.

There are three main different approaches to developing AI systems: Rule-based methods, ML and DL. Of these three approaches, the most used are the last two, which will be addressed in the next section.

4 Machine Learning

ML is a kind of AI that permits software applications to become more trustworthy by improving their capacity to anticipate accurate outcomes (Burns 2021). In ML, the algorithm is presented with a dataset with several data and a tag for the data. In this way, the system will find the common patterns in that dataset. Figure 1 presents the position of ML on AI. The more examples the dataset contains, the more effectively the algorithm responds. In addition, the system learns from its mistakes.

Based on theoretical concepts, ML can be described by a model, which has data as input and prediction as output. Yet, a model is no more than a mathematical formula. A mathematical formula is the result of a ML algorithm implementation. The mathematical formula will measure parameters that can be used for prediction. Models can be trained and learned from training data (Krzyk 2018). Figure 2 presents a diagram visualization of a ML Model.

This is very recurrent in society, allowing companies to investigate trends in the behavior of individuals and patterns in order not only to enhance their product but also to provide a better quality of service to users. Several companies in the world

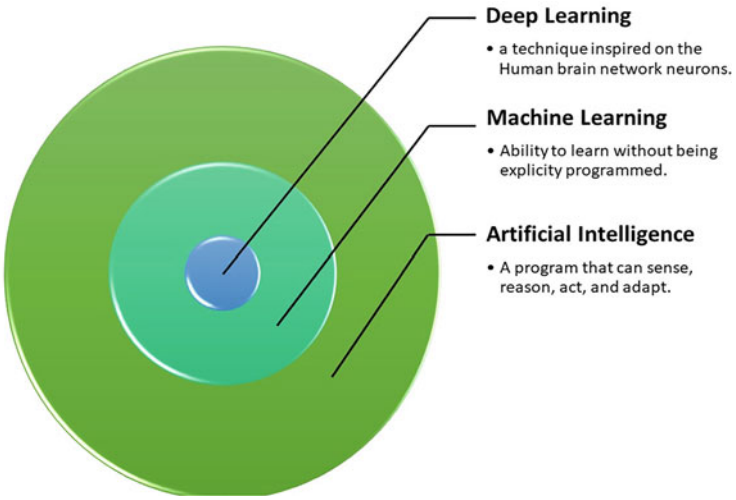

Fig. 1 Position of ML on AI context (adapted from (Cauduro 2018))

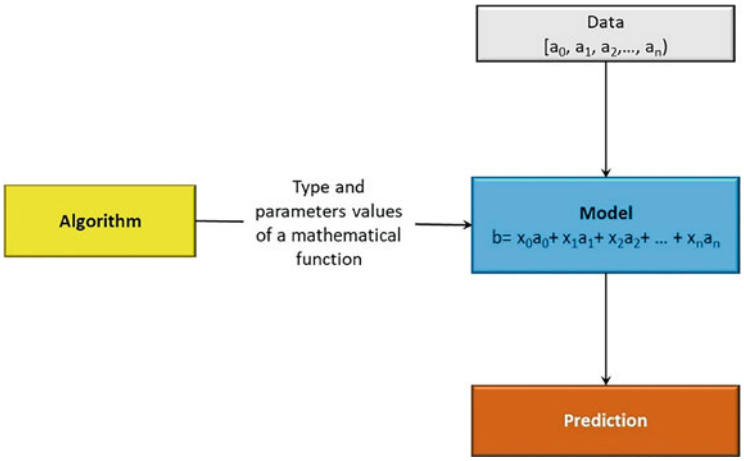

Fig. 2 Diagram visualization of a ML Model

use ML, and it acts as an important differentiating factor between companies. Today, ML is used in many areas and applications. One of the most prominent applications of ML is in recommendation systems, where it allows advising clients on specific topics. Examples are news, games and movies (Burns 2021).

As mentioned, there are several areas in which ML can be applied, and there are several categories in which the different algorithms are grouped based on their objective. As presented in Fig. 3, the three main categories are supervised, unsupervised, semi-supervised learning, and reinforcement learning (Dey 2016; Krzyk 2018). This figure also presents some applications of each type of ML categories.

4.1 Supervised Learning

Supervised learning can be described as a process that uses algorithms capable of producing patterns and hypotheses from given instances, applying them to predict unknown instances. This type of learning tries to predict dependent variables from a list of independent variables. All ML algorithms follow a similar process: dataset, features, algorithms, evaluation, and training.

To use ML models, it is always needed an input dataset divided into training and testing, which must contain millions of data. The larger the dataset, the better is the algorithm's response. The training includes the variable that will be predicted/classified, and it is with this data the algorithm will learn so that it can apply this knowledge in the test dataset and predict/classify that same variable (Dey 2016).

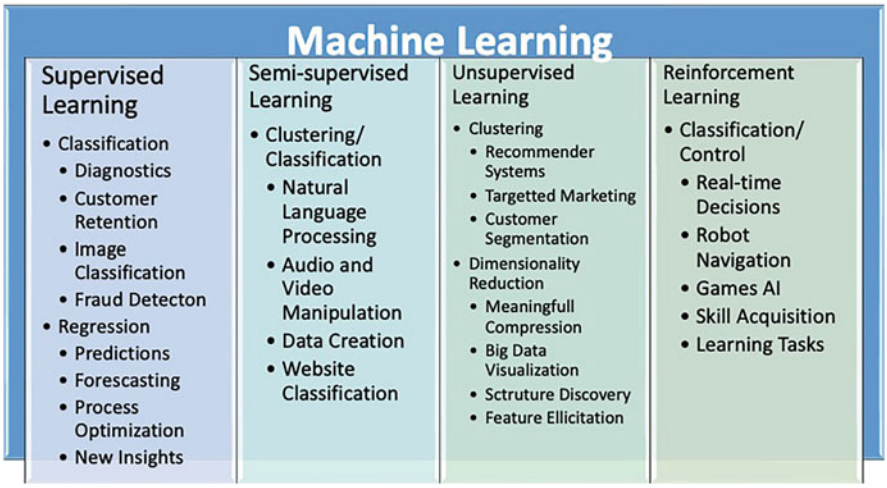

Fig. 3 Machine Learning Methods (Adapted from (Krzyk 2018; Raigon 2020))

Fig. 4 Supervised learning model

Features are data that will be transformed into a numerical representation so that the algorithm can understand it. The algorithm does not use all the data, because it will learn which one is relevant.

On supervised learning the dataset has several data samples, which consist of pairs of input-output examples that trained the model. When the model is prepared, it will predict the expected label outcome. Then prediction is compared with the label. If there is not a match, we have what is called an error. The error is merely feedback in the model, in which it will be updated. Figure 4 represents the supervised learning model.

Different ML algorithms can be used, namely regression or classification algorithms (Table 2).

Regression algorithms aim to know how one variable evolves concerning others. This type of algorithm predicts a continuous value. Examples of regression algorithm applications are predictions, forecasting, or estimating.

Classification algorithms seek to explain a categorical variable with two or more categories, dividing the data into classes, using common features. Some examples of application classification algorithms are fraud detection, image classification, customer retention and person diagnostics.

Table 2 Types of algorithms applications

Type	Algorithm name	Description
Regression	Linear Regression	Compares each feature to the outcome to assist future forecasts.
Regression Classification	Decision Tree	Data resource values are separated into branches at decision nodes by a classification or regression model.
	Naive Bayes	It is a group of basic probabilistic classifiers dependent on the application of Bayes theory with strong self-governance of the Naive Bayes features.
	Random Forest	Applies simples' decision based on approach of the most voted. In regression, the prediction is calculated by the average.
	AdaBoost	Employs numerous models to decide but scale is the important measure to obtain precision in forecasting result.
	Gradient Boosting Trees	It focuses on the mistake made by the previous trees and tries to correct it.
	Support Vector Machine	Used for classification task and locates a hyperplane that separates the classes.
Classification	Logic Regression	Quantifies the connection between the absolute variable and at least one free factor by evaluating probabilities using a logistical capability.

4.2 Unsupervised Learning

Unsupervised learning diverges from supervised learning in that none of the data has defined the label. The way the algorithm will create knowledge from the data is done differently, analyzing the affinity between the analyzed objects to detect similarities/differences between their characteristics; from there, labels will be created and assigned (Kotsiantis et al. 2007). This learning results in finding patterns in data that would otherwise be considered noise, not containing helpful information (Ghahramani 2003; Dey 2016).

An advantage of unsupervised learning is that the data does not need to be categorized, making huge amounts of unstructured data accessible for analysis. Algorithms try to draw assumptions from non-labelled data, finding new data patterns. Figure 5 presents the unsupervised learning model.

An important application of unsupervised learning is anomaly detection. In these methods, networks are trained to discover the composition and overall look of a data stream, settling whether one data point looks different from the rest. The application of these methods allows, for example, the detection of cyber fraud attempts in complex transactions.

Unsupervised learning model can be dividing into clustering and reducing data dimensionality.

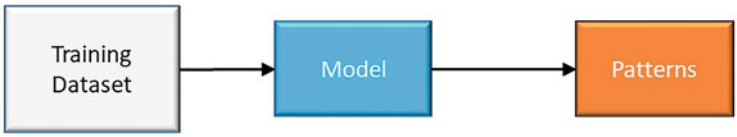

Fig. 5 Unsupervised learning model

Table 3 Algorithms that can be applied for clustering and dimensionality reduction

Type	Algorithm name	Description
Clustering	Gaussian Mixture Model	More flexible in the range and structure of clusters k-means.
	K-Means Clustering	Places the data into a few clusters (k), each having data with equal attributes.
	Hierarchical Clustering	It is a calculation that creates a hierarchy of groups. It begins distributing by groups based on information of each one. Here, two close groups will be in a similar group. This calculation closes when only one group remains.
	Recommender System	Assist to specify the important data for constructing a suggestion.
	K-Nearest Neighbors	It is a direct measure that keeps all available topics and describes new examples dependent on a similitude estimation.
Dimension Reduction	PCA/T-SNE	The processes decrease the number of features to 3 or 4 trajectories with the tallest variances.

Clustering is a technique that splits and groups similar data samples. The groups are called clusters. Examples of clustering are recommended systems, targeted marketing and customer segmentation.

Dimensionality reduction is a method of condensing features into so-called core values that concisely convey similar information. By choosing just a few components, the number of resources is reduced, and a small part of the data is lost.

Table 3 summarizes the algorithms that can be applied for clustering and dimensionality reduction.

4.3 Semi-Supervised Learning

Semi-supervised learning has much use in the digital world, detecting fraud, whether in the news, emails, etc. Algorithms trained in small datasets can learn to label data and be used in translations, allowing algorithms to translate languages using incomplete dictionaries.

The ML model is trained and tested with data present in unequal proportions. The proportion of training data is lower than the ratio of test data. Semi-supervised ML algorithms are located among unsupervised and supervised learning, and unlabeled data can significantly improve learning accuracy (Malapragada et al. 2017).

4.4 Reinforcement Learning

Finally, reinforcement learning consists of teaching a model of ML, defining specific rules that it will have to follow, presenting rewards when the algorithm completes a task well or giving punishments when it behaves wrongly (Burns 2021).

Reinforcement learning differs significantly from the learning as mentioned earlier. In this type of learning, the exchange between the agent and the surroundings in which it works is crucial. In this way, the agent interacts with the environment producing actions that will change the environment causing the machine to receive rewards or penalties (Ghahramani 2003).

It should be noted that the agent is not aware beforehand of the actions needed to take. The decisions it makes will influence future actions (Dey 2016). Figure 6 details the reinforcement learning model.

The machine's goal is to learn to behave in a way that enables rewards (or lessens penalties) over its lifetime. This learning only depends on two criteria: delayed result and trial and error research (Dey 2016; Sutton and Barto 2018). The agent intends to build an optimal policy; however, it solves the problem of studying new states while increasing its overall benefits. This trade-off dilemma is called Exploration versus Exploration. The agent must analyze the two sides of the dilemma and choose the strategy based on the overall results. Hence, to make the best general decision in the future, the agent must preserve information adequately. Examples of reinforcement learning are decisions made in real-time, computer vision, and autonomous driving.

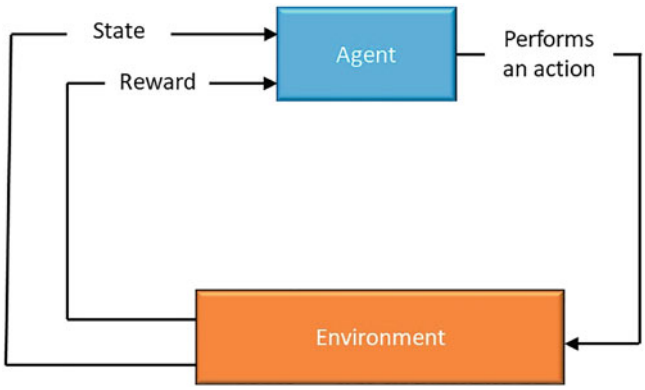

Fig. 6 Reinforcement learning model

There are two different model approach for reinforcement learning, the Markov Decision Processes (MDPs) and the Q-learning model.

An MDP consists of a cluster of finite domain states S, a group of possible actions A(s) in each condition, a tangible reward function R(s) and a transition model P(s', s | a) (Shuweta 2018).

Q-learning is a free of charge method that is applied to make a self-playing PacMan agent. It rotates around revising Q values which represents the value of a behaving a in a state s (Shuweta 2018).

5 Deep Learning

Deep Learning (DL) is a subclass of ML techniques where the systems are comprised of multiple layers to learn representations of data with various levels of conception. Most DL methods use neural network architectures, hence the name deep neural networks. Figure 7 depicts the amount of hidden layers in the neural network. Standard neural networks include 2–3 hidden layers, but deep networks contain around 150 layers (Santos et al. 2021).

The learning of classification tasks from images, text, or sound are performed by DL models (Santos et al. 2021). This learning, on the part of the model, to be done successfully, requires large amounts of labelled data and large computational power, often requiring high-performance GPUs to execute model training (Santos et al. 2021).

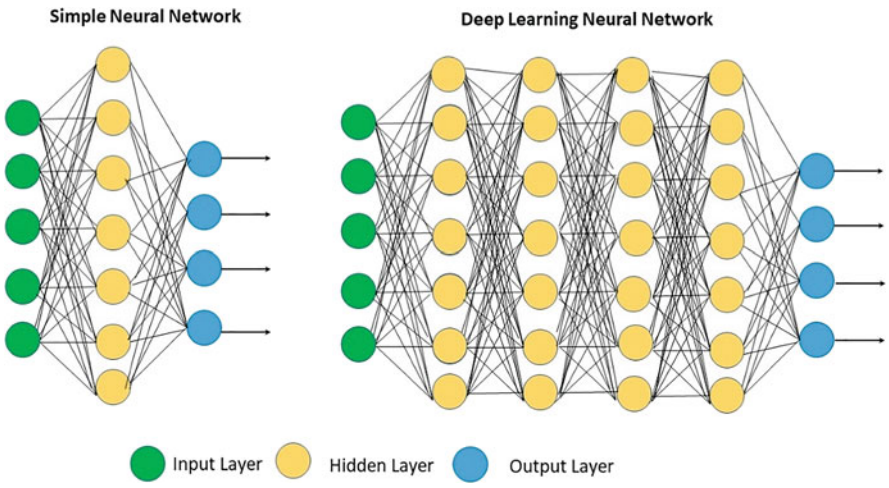

Fig. 7 Comparison between simple neural network and deep learning network (Adapted from (Ceron 2020))

Several areas use DL, but it has had particular success in autonomous driving. The DL is used to automatically detect objects and people in these systems. Another area where the DL is widely used is the Aerospace and Defense areas, *v.g.* satellites detect objects or identify safe zones for troops. These techniques are also relevant in automatic speech translation and home assistants (Mathworks 2022).

Most DL models use the so-called neural networks, inspired by the connections of neurons in the human brain, and can extract/learn features automatically. For example, you can pass images to a network, and this network can extract features from the image without any human intervention. As these networks are trained, they automatically learn to classify/solve problems. One major advantage they have is the ability to improve their classification capacity as the data increases and time goes by (Mathworks 2022).

Many DL applications use the technique of Transfer Learning (TF). TF is a procedure that requires adjustment a pre-trained prototype. The network training starts with an existing dataset. Then, this network is provided with the new data or dataset that contains unknown classes. From there, the network adjusts, transposing the previous knowledge to this new situation. Computation time is shorter in comparison to an untrained network (Mathworks 2022).

As shown in Fig. 8, DL models can be classified as supervised learning or unsupervised learning. In supervised learning DL convolutional neural networks and recurrent neural networks (RNNs) are employed. The RNNs may also be divided into the gated recurrent unit (GRU) and short-term memory. In unsupervised learning DL we have self-organizing map (SOM) and autoencoders (AE) networks. The AE is a restricted Boltzmann machine (RBM) type.

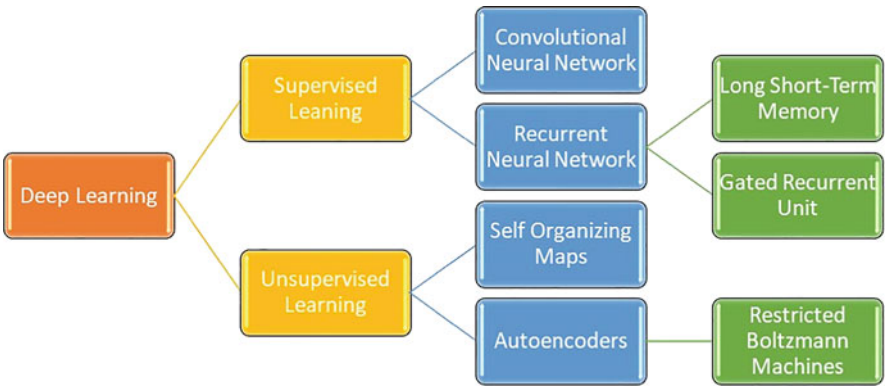

Fig. 8 Deep Learning Models (Adapted from (Madhavan and Jones 2021))

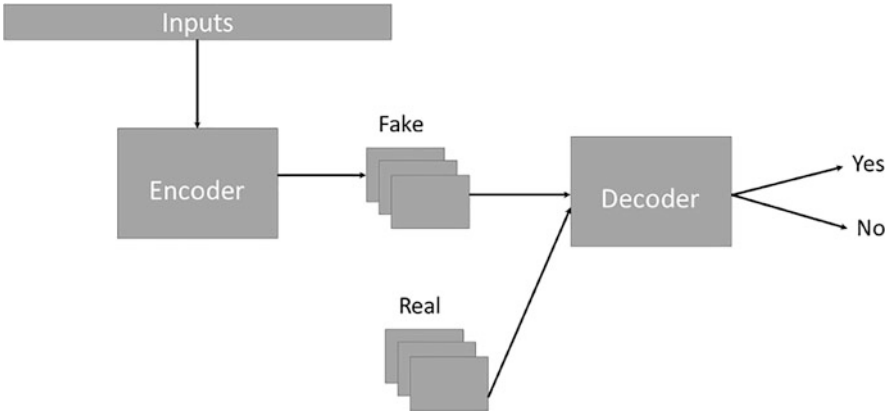

Fig. 9 GAN Architecture

6 Deepfake Generation

Another type of deep neural network are called Generative Adversarial Networks (GANs). These can be used to create deepfakes. They learn from a training dataset and generate a data sample with similar features. GANs have an architecture that is no more than two components of neural networks: an encoder and a decoder. The model uses the encoder to train an extensive dataset in order to generate fake data. The decoder is basically a binary classifier that receives inputs (real or face content) and uses a SoftMax function to identify the authentic data (Fig. 9).

Examples of deepfakes applications are VGGFace (Malli 2017), FakeApp (Malavida 2022), Faceswap (Deepfakes 2022), and CycleGAN (Zhu et al. 2017).

7 Deepfake Detection

Deepfake detection is technology possible, namely in two domains: images and videos.

7.1 Image Detection Models

The literature presents several processes to distinguish the images generated by GAN using deep networks. One of the methods is based on preprocessing procedures to analyze the statistical characteristics of the image and improve the recognition of false appearance pictures made by humans (Li et al. 2018a). Another method is centered on a deep convolutional neural network that identifies

false images generated by GANs (Do et al. 2018). Xuan et al. (2019) revealed a convolutional neural network (CNN) based on Gaussian Blur and Gaussian Noise to identify fake human pictures. A hybrid methodology was established to identify fake pictures effectively (Liu et al. 2019; Faceswap 2022).

7.2 Video Detection Models

The video detection models use one of two approaches: biologic or spatial temporal features analysis.

The first approach can be applied to three different methods: catch face fakes, timer for deepfakes, and the relationship between audio and video. The first method is based on the physical aspect of eye blinking. So, this method monitors the eye blinking to catch the fake face. A CNN with an RNN and a binary classifier are used to supervise eye blinking (Li et al. 2018b). The second method is based on the physical aspect of the timer (pulsation) to detect fake videos. This method uses a GAN to compare fake to authentic videos (Ciftci et al. 2020). Finally. The third method is based on the relationship between audio and video. This method uses DL models with a triple loss function which detects fake from authentic videos (Mittal et al. 2020).

The second approach is to use a convolutional neural network (CNN) to extract a feature from a frame. Later, these features can be passed through a LSTM that analyzes the temporal sequence in frames. Finally, the video is classified as real or fake with a Softmax function (de Lima et al. 2020). Also, another approach is Recycle-GAN, which employs dependent generative adversarial networks to combine spatial and temporal data. The evaluation results show that spatial and temporal information can produce a good result in detecting deepfakes (Bansal et al. 2018).

8 Deepfakes and the Administration of Criminal Justice

Deepfakes can have profound negatives impacts by targeting the reputation of individuals, creating false events or content that can result in a wrongful conviction or impact lawsuits, decreasing trust in institutions; and threaten national security or harm international relations if misused by governments (Europol 2020; Meskys et al. 2020; Flynn et al. 2021).

From a criminal justice administration perspective, deepfakes can be quite harmful and dangerous as they create a haze between true and false.

Manipulated videos, pictures, audio or documents can be presented as evidence and deceive judges, lawyers and police officers, “casting doubt on audio-visual evidence as an entire category of evidence” (Trend Micro 2020). Moreover, as the processing power of information systems increases and deepfake technology

develops even further, namely in real-time applications, videoconferencing or teleconferencing of witnesses, experts or parties may be manipulated by making them sound, do and act in a manner that did not happen.

Two years ago, the first case of deepfake being submitted as evidence in UK courts made the headlines. In a child's custody case, the mother of the child presented an audio record where the child's father was threatening her (The Telegraph 2020). Later on, the authenticity of the recording was challenged and the court ended up dismissing it. According to the experts that inspected the audio file, it had been tampered with, in order to include words that the father never said (The National 2020).

The court's traditional stance of taking at face value certain type of evidence—video or audio, for example—may have reached its end.

The main question is how the courts and the society in general can adequately protect themselves against the negative effects of deepfake technology.

The suggestion for a Regulation of the European Parliament and of the Council laying down coordinated regulations on artificial intelligence (Artificial Intelligence Act) gives us an idea of what could be the European Union's take on this matter.

Included in its subject matter (article 1(d)) are the rules for AI systems used to "generate or manipulate image, audio or video content". AI systems that are defined in article 3(1) "as software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". Among the referred techniques and approaches we find machine learning and deep learning (annex I(a)).

This proposed framework adopts a risk-based approach to artificial intelligence that distinguishes between four different degrees of risk: unacceptable, high, limited and minimal. To better understand the level of risk posed by deep fakes, it is necessary to briefly define each one of them.

Firstly, in the unacceptable risk lie AI uses that are prohibited by the Artificial Intelligence act, as they fail to comply with the EU's values (*e.g.* fundamental rights). AI used for manipulation of human behavior causing harm physical or psychological harm (article 5, 1(a)(b)), social scoring (article 5, 1(c)), and real-time remote biometric identification in in publicly accessible spaces for law enforcement purposes (article 5, 1(d)) is prohibited. Only in specific cases can real-time remote biometric identification occur, more precisely when it is proportionate and strictly necessary to pursue one of three objectives: targeted search of victims, stop a terrorist attack or imminent threat to life and physical safety, or track a suspect or perpetrator of serious crimes (article 5, 1(d) i,ii,iii).

The categorization of an AI system as high-risk means that it is allowed to use, insofar the requirements laid out in article 8 to 15 are fulfilled (*e.g.* up-to-date technical documentation, logging capabilities, adequate transparency, human oversight, appropriate level of accuracy, robustness and cybersecurity) and the providers, users and other parties comply with the obligations foreseen in article 16 to 29, in particular conformity assessment. The European lawmaker includes in this

category of risk AI systems in the domain of biometric identification, recruitment tools, credit scores, management of emergency services, among others.

Although the European proposal mentions in the proposal a three-tier risk-based approach (unacceptable, high, and low or minimal), it a fourth level is commonly recognized: limited risk. "Certain" AI systems—recalling the expression used in the proposal—are subject to transparency obligations (article 52). Meaning that their use is allowed, but providers and users must inform the end-users that they are interacting with an AI system or content artificially generated. The range of AI systems covered in this risk level is quite large as it encompasses all AI systems that interact with natural persons, detects emotions, or categorizes based on biometric data, or generates deepfakes.

The AI systems that pose no or low risk fall into the last category of risk. No obligations are imposed to the providers or users of such AI systems. Nevertheless, providers may decide, on voluntary basis, to create codes of conduct and ensure compliance with the requirements set out for high-risk systems (article 69). It is however a decision of the providers of non-high-risk AI systems to apply these requirements, as they are only mandatory for high-risk systems.

Drawing from the aforementioned classification of risk, we can safely affirm the use of deep fakes falls unequivocally into the third category: limited risk. Accordingly, deep fakes would not be prohibited nor deemed as high-risk, but they wouldn't also be classified as low or minimum risk. So, users "who use an AI system to generate or manipulate image, audio or video content that appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic, should disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin" (recital 70 of the proposal). Users of AI systems that produce deep fakes must therefore disclose the artificial origin and nature of the content generated (article 52(3)).

The transparency obligation has however some exceptions.

The European lawmaker tried to reach a balance between the protection of very different values and goals: development, marketing and use of AI systems (recital 1); economic growth and social development (recital 2 and 3); human dignity, health, safety, freedom, equality, democracy, the rule of law, the right to non-discrimination, protection of personal data and privacy, and the rights of children (recital 1 and 15). In doing so, the proposal allows for the dismissal of the obligation of transparency if the deep fakes are used for legitimate reasons. Article 52(3) specifically declares the artificial origin of deep fakes may not be disclosed if it is used with the purpose of detecting, preventing, investigating and prosecuting criminal offences or it is a legitimate exercise of the right to freedom of expression and the right to freedom of the arts and sciences.

Understanding whether deep fakes must be announced or marked as such, or even if they are legal, implies an analysis done case by case. There may be cases where this type of artificial content amounts to a breach of fundamental rights of third parties prescribed as criminal (*e.g.* defamation, extortion, child pornography). In others however it may simply be an expression of creativity. The borders of

the freedom of expression, arts and sciences are somewhat difficult to define and curbing the risk of mass manipulation with deep fakes may prove to be a challenge too difficult for law to overcome.

Technological content control or identification tools could be a valuable tool for governments, businesses, and persons alike.

9 Conclusion

The big problem that the administration of criminal justice must solve is enforceability. The current legal framework is unable to tackle deepfakes.

The problem of applying existing legal rules in the case of deepfakes can be summarily described as follows. First, the rapid evolution of technology makes any legal norm quickly out of date. Second, there is an urgent need to define what and how technology should be used. Third, due to the cross-border nature of the technologies, it might be extremely complex to identify the rules that these technologies must comply with; hence they are usually registered in countries with more lenient rules. Fourth, it is difficult to enforce legislation when technology development and usage is not confined to a single country. Fifth, the duties of the parties interested in deepfakes are frequently partial. Sixth, it is possible to circumvent the rules of a given jurisdiction easily (van der Sloot et al. 2021).

AI should assist in identifying and removing problematic deepfakes. In the same way that ML and DL provide the problem, they also have to be part of the solution. Nevertheless, digital literacy should not be neglected, as the pace of technological solutions employed in deepfake creation are frequently ahead of their detection counterparts. There is no real substitute for a critical stance with regard to digital content.[1]

[1] *See* generally, on the different applications of Machine Learning and AI, in this book A Oliveira and M A T Figueiredo - Artificial intelligence - historical context and state of the art; I Trancoso, N Mamede, B Martins, H S Pinto and R Ribeiro - The impact of language technologies in the legal domain; J Gonçalves-Sá and F L Pinheiro - Societal Implications of Recommendation Systems - A Technical Perspective; A T Freitas - Data-driven approaches in healthcare - challenges and emerging trends; M Correia and L Rodrigues - Security and Privacy; E Magrani and P G F Silva - The Ethical and Legal Challenges of Recommender Systems Driven by Artificial Intelligence; M Lanz and S Mijic - Risks associated with the use of natural language generation - Swiss civil liability law perspective; M S Fernandes and J R Goldim - Artificial Intelligence and Decision Making in Health - Risks and Opportunities; and W Gravett - Judicial Decision-making in the Age of Artificial Intelligence. *See* also, on the AI Act, in this book P U Lima and A Paiva - Autonomous and Intelligent Robots - Social, Legal and Ethical Issues; A T Fonseca, E V Sequeira and L B Xavier - Liability for AI Driven Systems; M N Duffourc and D S Giovanniello - The Autonomous AI Physician - Medical Ethics and Legal Liability; A Keller, C Martins Pereira and M Lucas Pires - The European Union's approach to Artificial Intelligence and the Challenge of Financial Systemic Risk; J C Abreu - The "Artificial Intelligence Act" Proposal on European e-Justice Domains Through the Lens of User-focused, User-friendly and Effective Judicial Protection Principles. *See* also, on biases, in this book P G Marques - AI Instruments for Risk of Recidivism

References

Anderson KE (2018) Getting acquainted with social networks and apps: combating fake news on social media. Libr Hi Tech News 35:1–6

Bansal A, Ma S, Ramanan D, Sheikh Y (2018) Recycle-GAN: unsupervised video retargeting. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018. Springer, Cham, pp 122–138

Burns E (2021) Tech Accelerator. In: In-depth guide to machine learning in the enterprise. TechTarget. Available via TechTarget. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML Accessed 14 Feb 2022

Cauduro A (2018) Medium. In: Deep Learning: o motor dos negócios na era da inteligência artificial. Stay curious. Available via Huia. https://medium.com/huia/intelig%C3%AAncia-artificial-uma-corrida-desleal-80bfa53075ed. Accessed 14 Feb 2022

Ceron R (2020) Blog de Infraestruturas de TI. In: A Inteligência Artificial hoje: dados, treinamento e inferência. IBM. Available by IBM. https://www.ibm.com/blogs/systems/br-pt/2020/01/a-inteligencia-artificial-hoje-dados-treinamento-e-inferencia/. Accessed 14 Feb 2022

Chawla R (2019) Deepfakes: how a pervert shook the world. Int J Adv Res Dev 4:4–8

Ciftci UA, Demir I, Yin L (2020) How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In: 2020 IEEE international joint conference on biometrics (IJCB). IEEE, Houston, pp 1–10

de Lima O, Franklin S, Basu S, Karwoski B, George A (2020) Deepfake detection using spatiotemporal convolutional networks. arXiv:2006.14749

Dey A (2016) Machine learning algorithms: a review. Int J Comput Sci Inf Technol 7:1174–1179

Do NT, Na IS, Kim SH (2018) Forensics face detection from GANs using convolutional neural network. ISITC 2018:376–379

Europol (2020) In: Malicious uses and abuses of artificial intelligence. Available via Europol. https://www.europol.europa.eu/cms/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf. Accessed 15 Feb 2022

Faceswap (2022). DeepFakes. In: Deepfakes – faceswap. Gihub. Available by Github. https://github.com/deepfakes/faceswap. Accessed 15 Feb 2022

Figueira Á, Oliveira L (2017) The current state of fake news: challenges and opportunities. Procedia Comput Sci 121:817–825

Flynn A, Clough J, Cooke T (2021) Disrupting and preventing deepfake abuse: Exploring criminal law responses to AI-facilitated abuse. In: The palgrave handbook of gendered violence and technology. Palgrave Macmillan, Cham, pp 583–603

Ghahramani Z (2003, February) Unsupervised learning. In: Summer school on machine learning. Springer, Berlin, pp 72–112

Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC (2020) Deepfakes: trick or treat? Bus Horiz 63:135–146

Kotsiantis SB, Zaharakis ID, Pintelas PE (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160:3–24

Prediction and the Possibility of Criminal Adjudication Deprived of Person-al Moral Recognition Standards – Sparse Notes from a Layman; W Gravett - Judicial Decision-making in the Age of Artificial Intelligence. *See* finally, on AI and judicial reasoning, in this book P G Marques - AI Instruments for Risk of Recidivism Prediction and the Possibility of Criminal Adjudication Deprived of Per-sonal Moral Recognition Standards – Sparse Notes from a Layman; L M Pereira, F C Santos and A B Lopes - AI Modelling of Counterfactual Thinking for Judicial Reasoning and Governance of Law; W Gravett - Judicial Decision-making in the Age of Artificial Intelligence; and J C Abreu - The "Artificial Intelligence Act" Proposal on European e-Justice Domains Through the Lens of User-focused, User-friendly and Effective Judicial Protection Principles.

Krupansky J (2017) Untangling the definitions of artificial intelligence, machine intelligence, and machine learning. https://perma.cc/RVZ4-88NP. Accessed 3 Feb 2022

Krzyk K (2018) Towards Data Science. In: Coding deep learning for beginners. Towards Data Science. Available by Medium. https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d. Accessed 17 Feb 2022

Li H, Li B, Tan S, Huang J. (2018a) Detection of deep network generated images using disparities in color components. arXiv preprint arXiv:1808.07276

Li Y, Chang MC, Lyu S (2018b). In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877

Liu F, Jiao L, Tang X (2019) Task-oriented GAN for PolSAR image classification and clustering. IEEE Trans Neural Netw Learn Syst 30:2707–2719

Madhavan S, Jones T. (2021) Deep learning architectures: the rise of Artificial Intelligence In: Artificial Intelligence. https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/ Available by IBM. Accessed 21 Mar 2022

Malapragada P, Jain R, Liu Y (2017) Applying reinforcement learning and supervised learning techniques to play hearthstone. In: 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, Cancun, Mexico, pp 1145–1148

Malavida (2022) FakeApp 2.2.0. In: Malavita. https://www.malavida.com/en/soft/fakeapp/. Available by Malavita. Accessed 28 Feb 2022

Malli RC (2017) Keras-vggface. In: Github. https://github.com/rcmalli/keras-vggface. Available by Github. Accessed 28 Feb 2022

Maras MH, Alexandrou A (2019) Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. Int J Evid Proof 23:255–262

Mathworks (2022) What is deep learning? 3 things you nedd to know. In: Deep Learning. https://www.mathworks.com/discovery/deep-learning.html. Available by MathWorks. Accessed 28 Feb 2022

Meskys E, Liaudanskas A, Kalpokiene J, Jurcys P (2020) Regulating deep fakes: legal and ethical considerations. J Intellect Prop Law Pract 15:24–31

Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions don't lie: an audio-visual deepfake detection method using affective cues. In: Proceedings of the 28th ACM international conference on multimedia. ACM, New York, NY, USA, pp 2823–2832

Muehlhauser L (2013) The privacy expert's guide to artificial intelligence and machine learning (future of privacy forum, 2018) at 5; "what is AGI? https://intelligence.org/2013/08/11/what-is-agi/. Accessed 2 Feb 2022

Raigon J (2020) Machine learning for fraud prevention keeps TrafficGuard agile. In: trafficguard. https://www.trafficguard.ai/resources/machine-learning-for-fraud-prevention-keeps-trafficguard-agile. Available by trafficguard. Accessed 28 Feb 2022

Santos F, Durães D, Marcondes F, Gomes M, Gonçalves F, Fonseca J, Wingbermuehle J, Machado J, Novais P (2021) Modelling a deep learning framework for recognition of human actions on video. In: Rocha A, Adeli H, Dzemyda G, Moreira F, Correia AMR (eds) Trends and applications in information systems and technologies. Springer International Publishing, Cham, pp 104–112

Shuweta B (2018) Reinforcement learning 101. In: Medium. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292. Available by Toward Data Science. Accessed 20 Mar 2022

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. A Bradford Book, Cambridge

The National (2020) Deep fake audio evidence used in UK court to discredit Dubai dad. In: The National Newspaper. https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764. Accessed 22 Mar 2022

The Telegraph (2020) Doctor Audio Evidence used to damn Father in Custody Battle. In: The Telegraph Newspaper. https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/. Accessed 22 Mar 2022

Trend Micro (2020) Trend Micro Security Predictions for 2020. In: Trend Micro Research. Available by Trend Micro. https://documents.trendmicro.com/assets/rpt/rpt-the-new-norm-trend-micro-security-predictions-for-2020.pdf. Accessed 21 Feb. 2022

van der Sloot B, Wagensveld Y, Koops BJ (2021) Summary Deepfakes: the legal challenges of a synthetic society. In: Tilburg Institute for Law, Technology, and Society. https://www.tilburguniversity.edu/sites/default/files/download/Deepfake%20EN.pdf. Availablen by Tilburg University. Accessed 3 Mar 2022

Xuan X, Peng B, Wang W, Dong J (2019) On the generalization of GAN image forensics. In: Sun Z, He R, Feng J, Shan S, Guo Z (eds) Chinese conference on biometric recognition. Springer International Publishing, Cham, pp 134–141

Zhu JY, Taesung P, Isola P, Efros AA (2017) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: CycleGAN Project Page. https://junyanz.github.io/CycleGAN/. Available by github. Accessed 28 Feb 2022

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.