Radboud University

Is Social Media Well Prepared for the Rise of Deepfakes?

An Effectiveness Analysis of Current Counter Measures

Bachelor's Thesis in Artificial Intelligence

Author:
Felicitas Reddel

Student Number:
s4830717

Supervisor:
Dr. W.F.G. Haselager (Pim)
Associate Principal investigator - Donders Centre for Cognition
Associate Professor - Artificial Intelligence

Second reader:
Dr.H.K. Schraffenberger (Hanna)
Researcher - Interdisciplinary Hub for Security, Privacy and Data Governance

Abstract

We analyzed whether current technical counter measures for deepfakes are less effective in social media environments than in traditional journalism environments. To that end, we compared earlier forgery methods to the possibilities and limitations of various kinds of deepfakes. Our exploration of several deepfake counter measures has enabled the analysis of their effectiveness within the two media environments. Our recommendation is to pay increased attention to the social media environment. We view detection as a decent temporary patch and solutions that combine authentication and provenance as a promising candidate for reliable long-term solutions. Integrating the latter one into mobile phone operating systems could be compelling future work.

12th of January 2020

Contents

List of Figures

1 Introduction

1.1 Background

Deepfakes[1] are ideal candidates for sensational headlines. They are associated with sensitive topics like politics, deception and adult content, mixed with buzzwords such as artificial intelligence (AI) and machine learning (ML). Consequently, various potential risks spring to mind – ranging from defamation, scams, phishing and cyber bullying, over manipulation of elections, up to potentially tricking world leaders into triggering a nuclear first strike. Are deepfakes really as threatening as the headlines suggest?

1.1.1 Impact of Deepfakes

On the one hand, misleading and out-right wrong news has existed for a long time. These range from hear-say over written texts, analog editing of photos and digital editing, to video editing and animation for movies. With many of these, the societal implications and risks had been feared. And yet, with their introduction, counter measures have been developed and applied. On the other hand, one could argue that the implications of deepfakes might be different – or that the known effects of "conventional" media forgery and fake news could reach a degree, that would be challenging to keep control over [1]–[4]. A reason for that position could be that it becomes easier and cheaper to create more realistic deepfakes. Another reason could stem from the speed at which information can spread, and what content is favored in contemporary information environments. A third reason could lie in today's possibility of fake news being targeted much more precisely at very specific audiences or even individuals.

Deepfakes can have different kinds of potential negative effects which we categorize as first-order and second-order effects. The former include already mentioned more direct effects that originate from defamation, scams, cyberbullying, etc. A second-order effect refers to something that has an indirect negative effect. We consider the deterioration of trust to be such a second-order effect. A citizen's trust in other citizens, political institutions, corporations and media could be negatively affected which could lead to societal harm. Intuitively, we consider such a second-order effect as potentially disastrous. While an exploration of this field makes the impression of being fruitful, there seems to be little to no research addressing the influence of deepfakes on society or more concretely trust within society.

1.1.2 Original Plan

Originally, we embarked on the journey to explore the influence of deepfakes on trust and how well various counter measures could maintain aspects of trust. For this purpose, we used and extended the existing literature to chart a space of trust which included eight dimensions. However, we realized that this angle has not been very fruitful. Although the trust dimensions refer to separate concepts, we were unable to infer meaningful differences in influence on trust. For your reference, see appendix A to view our elaborations of the trust space. In the following, we show a slightly different angle that we pursued instead.

1.1.3 Counter Measures

Given the potential harm that deepfakes could bring along, multiple counter measures have been suggested. For example, one solution approach consists in classifying suspicious videos as either deepfake or as real. This can be done by humans or by AI [5]. Unfortunately, deepfakes are constantly improving and are becoming harder and harder to spot – for both human and artificial classifiers. This renders such an approach a cat-and-mouse game which seems to be favoring the offense over the defense.[2] Even if the catch up duration is a small one, this might be sufficient time to allow for substantial damage. Such an approach is just one among many. How should we decide which ones to pursue the most? One thing is certain: resources like time and funding are limited. Which counter measures are most effective and thus should be supported through e.g. funding? It becomes crucial to prioritize wisely between the different measures. With this project, we aim to provide support for decisions on what counter measures might deserve more attention.

1.1.4 Focus on Media Environments

Furthermore, we have the intuition that the power of different counter measures is potentially depending on the kind of media environment which it should be used for. Social Media (SM) differs from Traditional Journalism (TJ) in various aspects. These aspects reach from the kind of content that is published, over how the posted content is moderated, how content spreads and how users tend to consume it. For rough provisory definitions, see the following boxes.

Social Media

Social media refers to websites and applications which can enable a user to share and potentially create content [6]. Examples include but are not limited to Facebook, WhatsApp, Telegram, Twitter, YouTube, Reddit, Quora, Snapchat, WeChat, TikTok and Weibo.

[1] Within this thesis, 'deepfakes' refer to current and future deepfakes, i.e. the capabilities and quality of deepfakes, that might be expected within the next few years.

[2] The offense refers to the (malicious) use of deepfakes and the defense to the classification and potential neutralization of them.

Traditional Journalism

Traditional journalism refers to mainstream media journalism. In TJ, journalistic authority is based on the corresponding institution [7]. Examples include but are again not limited to The Wall Street Journal, The Guardian, The Washington Post, De Telegraaf, De Volkskrant, CBS, BBC and CNN.

We assume that some counter measures will work better for TJ than for SM. For instance, a big media outlet could vow to check all used (video) material for traces of being deepfakes. If they fail in the future and report faulty information through thoughtless use of deepfakes, the media outlet might damage its reputation significantly. The numerous, decentralized, ephemeral and sometimes anonymous SM creators might have not such incentives. While we might expect the same level of scrutiny from some SM users whose social media use is centered around aiming to earn a reputation of sharing truthful information, we would not expect such scrutiny levels from most users. Consequently, insights in differences between SM and TJ in this regard could help to reveal blind spots and to direct the attention to counter measure approaches which could fix them.

1.2 Research Question

In this thesis, we aim to investigate the effectiveness of various deepfake counter measures while we highlight their differences in SM and TJ, respectively. More specifically:

Research Question

Are the current counter measures for deepfakes generally less effective in the SM environment than in the TJ environment?

This research question contains a directed hypothesis. Reasons for this intuition stem from the relative influence of SM and TJ over the attitudes of citizens,[3] the way information is created and spreads via the two media[4] as well as the existing incentives and resources to actively handle deepfakes.[5] Answering this research question could help to effectively (re-)allocate limited resources between approaches. Such a prioritization would hopefully decrease inflicted harm.

Potential answers to this question could include increasing the use of some counter measures or the need for a new, alternative counter measure. And while it is unrealistic to devise detailed new approaches within the scope of this thesis, it might possible to specify properties of such alternative approaches.

[3] E.g. many young people get much of their news from SM [8], [9] – even if the news originated from TJ and are spread via SM.

[4] E.g. considering SM with its higher amount of creators, their knowledge about a given topic and the the faster speed at which SM operates. This combination can lead to "digital wildfires" by SM before TJ can react and debunk them.

[5] E.g. TJ has an advantage to solve the issue through better incentives, less authors, more control, etc.

1.3 Methods & Outline

The aim of this project consists in testing the intuition whether it is wise to focus in the future more on effective counter measures for SM (rather than on those for TJ). To do this, we consider it useful to have an overview over the historical path towards deepfakes which includes aspects of previous photo and video related forgery and counter measures (see section 2), as well as to understand the working and the state of the art of deepfakes (see section 3). We introduce three overall deepfake categories, how they exactly differ and how they are created. Moreover, the comparison between SM and TJ is based on characteristics like the number of involved authors, the incentives that these authors experience and how the published content spreads. These traits have been judged with the help of existing literature in media and communication science. After this, four contemporary main approaches for deepfake counter measures are presented. They play a crucial role in the subsequent section in which we discuss their effectiveness with an eye on the media environment. Finally, in section 7, everything flows together to form an educated guess about whether we should pay increased attention at counter measures in the SM environment and what kind of counter measures could be more effective.

2 The Historical Path Towards Deepfakes

History is filled documents that have long been manipulated. Figure 1 provides an overview over some inventions and the possibilities of forgery. To create a better overview, we have split the timeline into the four eras that are distinguished by shade. One subsection has been dedicated to each of these eras. Era by era, an example of the capabilities of the contemporary forgery methods and their respective efforts have been provided.

2.1 Pre-Photo Editing

In the times before photography had been invented, forgery was mostly found in the realm of legal documents. An example from medieval times can be found in a legal document whose source was allegedly emperor Constantine. This document created the impression that Constantine wanted to transfer control over Italy to the contemporary pope. The stated reason was gratitude for personal god belief and for being cured from leprosy [10]. Many other examples of forgery throughout humanity's history exist. However, the focus of this work revolves around image and sound related forgery which brings us to the other three eras on the timeline.

2.2 Analog Editing

In the first half of the 19th century, it became possible for the first time to capture reality with the help of light and chemical reactions. Compared to written words or a painting, this was a far more objective way to document events. However, the temptation of forging early-day photographs

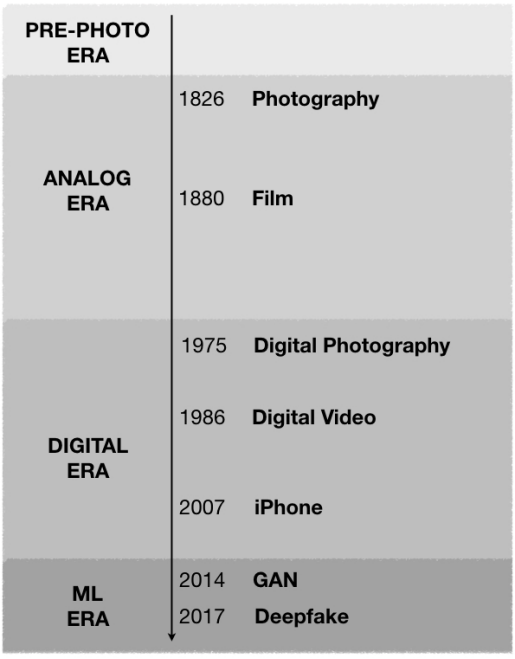


Figure 1: History overview

was lower than one might expect. At that time, it took around eight hours to capture a picture [11]. Furthermore, these pictures were fading when one was looking at them (due to additional light exposure). Thus, image forgery would have been very costly while delivering only short-lived effects.

In the second half of the 19th century, it became easier to manipulate photos. Figure 2 a infamously manipulated image that is supposed to show General (and later president of the USA) Ulysses S. Grant around 1865 [12]. The image is a composite of three other images which can be found in appendix B. The only counter measures of identifying image manipulation consisted in unaided and close observation. For example, the orientation of the general's head does not fully fit the orientation of the body [13]. Furthermore, the knowledgeable observer realizes that the uniform does not fit the time in which this photo was supposedly taken and that the horse's features are not the ones of Grant's favorite horse [13].

Figure 2: A composite picture supposedly showing general Ulysses S. Grant [13]

In the 20th century, Stalin famously removed adversaries from photos (see figure 3) or cut pictures together in order to make crowds look bigger [14]. However, these techniques required a great deal of resources as well as fine artistic and motor skills by the propaganda machinery of the Soviet Union.

Figure 3: Analog editing. Stalin removing people from photos. Left: Before. Right: After [15].

Viewing both images side by side reveals only a few differences such as increased brightness in the manipulated image. But increased brightness is not necessarily a sign of manipulation. Given the quality of these images, it is hard to spot clear traces of manipulation without any expert knowledge. With the increasing resolution and quality of images, it became a bigger artistic challenge to manipulate images in this way as inconsistencies were then easier to spot.

In the 1950s, only rudimentary editing of moving pictures was possible. The physical tape could be cut with a razor blade and taped together [16]. Like this, the order of frames shown could be manipulated. However, one was cutting the original film without being able to know exactly where the cut was in the scene. In the 1960s, electronic editing made it possible to splice almost frame accurate and without the need to risk destroying the original.

2.3 Digital Editing

Until 1975, both pictures and films were exclusively analog. The digitalization of images and the upcoming digital tools for digital editing elicited concerns around not being able to trust one's eyes anymore. Among other digital tools, Photoshop can be used to manipulate images much easier. Like this, we have more tools available, can reverse missteps and try endlessly until a satisfactory version is found.

And the capabilities of Photoshop experienced drastic improvements since entering the market. Around the time when the introduction of the first iPhone in 2007 pushed the wide smartphone adoption, Photoshop was introducing the clone stamp tool [17] which was infamously misused by Hajj to exaggerate smoke (see Figure 4). Hajj was a freelance photographer, that had been working for ten years for Reuters. In 2006, he had taken a photo of Israel bombing of a Lebanese town. Subsequently, it is likely that he used Photoshop's clone stamp tool to make the smoke seem more intense. After this fraud was discovered, Hajj claimed to have only "removed dust specks" to make the image more clearly visible [18]. As a conse-

quence, he was fired and the newspaper officially apologized for this incident.

Figure 4: Digital editing. Copy-pasted parts to give the illusion of more smoke. Left: Before. Right: After. [19]

One can easily spot Hajj's doctoring without any extra tools.[6] However, over time, the tools and techniques to manipulate images have improved. Tutorials are fairly widespread and freely available. The possibilities to create realistic images has become endless. When investing a few hours to get familiar with the tools, one could create most imaginable pictures without leaving such glaring imperfections. Counter measures in this realm include the field *digital image forensics* in which we usually ask two questions: Has the image been taken by the device which it had been claimed to be taken with? And is the depicted scene the original one [12, p. 16]? Any edit of an image leaves subtle traces of manipulation. A cumbersome and long process can provide an analysis of an image's history [12, p. 16-17]. In this context, especially digital watermarking is used. It is a process that consists of embedding information (e.g. about the owner) into a digital image. A successful watermark stays intact even if the watermarked picture is altered [12, p. 18]. This can help in cases in which I take a photo, digitally watermark it, another person steals my image and applies some manipulations and then pretends it is their image. If the watermark remains, I can show that it has been my original image.

2.4 ML Editing

Within the following years, machine learning (ML) started to be more commonly adopted for image manipulation. With the contemporary versions of Adobe's Photoshop, the use of AI makes it possible to select an object with just a few clicks. This works even for very intricate object boundaries, such as hair. Given these advances, one can also remove a whole object such as a person with just a few simple clicks.[7] And Adobe Cloak makes the same action possible for videos (see Figure 5).

Another powerful example tool of Photoshop is the Face Aware Liquify. It automatically detects faces[8] and enables the user to intuitively adjust various facial features.[9] Even videos can be edited in the same manner.

[6] See e.g. the repetition of the patterns in the smoke clouds in the left upper section of the image.

[7] E.g. follow the YouTube tutorial in [20].

[8] This is possible if the face is seen from a frontal perspective.

[9] E.g. follow this tutorial in [22].

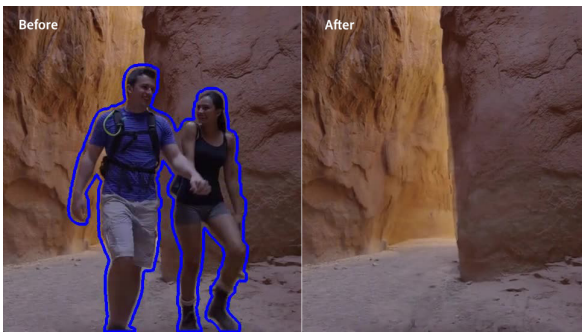


Figure 5: ML editing. Adobe Cloak – content aware eraser with a few clicks [21].

A crucial invention in the ML Editing era are Generative Adversarial Networks (GANs).[10] GANs made it possible to create pictures of people that have never existed [23].[11] Other GAN models have often been the backbone for creating deepfakes. The first deepfakes were made public in the end of 2017 [25].

A question that we could ask is whether we could use ML to detect (ML assisted) manipulation of images. In 2019, Adobe collaborated with UC Berkeley on creating a tool that detects a specific kind of facial manipulation – facial warping [26]. And although this tool can outperform humans in this task, the domain is very limited. It is only about a specific kind of face image manipulation. General solutions are not in sight yet. Moreover, note that this tool and the corresponding research are only happening in 2019 while ML editing is around for over 10 years. Thus, historically, the defense has been significantly lacking behind the offense.

2.5 Overall Trends

To conclude, let us take a zoomed-out perspective on the effort for a given task: e.g. removing a person from a picture. In the epoch of analog editing, this task required expert knowledge, careful handling of the original material, expensive equipment and a bigger time-investment. Furthermore, it could have been the case that the original image might have been destroyed in the process. Already in the early phase of digital editing, steps could easily be reversed and the required expert knowledge could be gathered online or by trial and error. The original picture was now safe from being destroyed because digital copies could be created without any costs and even undoing faulty actions had become possible.

Acquiring Photoshop came (and comes) still with some financial inhibition.[12] However, it can be acquired flexibly by anyone and without attracting much unwanted atten-

[10] A GAN is a class of ML algorithm that uses two neural networks competing in a game against each other. One tries to classify correctly whether an image is real or generated. The other one tries to create better and better images to fool the former network. For a more elaborate explanation of GANs, see section 3.4.1.

[11] There is even a website which you can be refreshed to view examples [24].

[12] See the official pricing in [27].

tion. Furthermore, there are many alternatives to remove people from a photo (see e.g. [28] or [29]). In 2019, the task of removing a person from a picture requires hardly any expert knowledge and can be performed in less than half a minute.[13]

Another aspect in the zoomed-out perspective is how *visceral* a specific media-type is, i.e. how persuasive is a given media type on a gut-level [30]. Viscerality increases from text (pre-photo) to photographs to moving pictures and would be expected to be even higher in realistic virtual reality applications. Photos and videos are very intuitively persuasive [31]. This makes the (expected) effect of deepfakes more potent than the effects of misleading text or still pictures. The costs and time requirements for forgery decreased drastically while the viscerality of the new media trumps those of more traditional media types. Now, let us have a look at how deepfakes make this strong combination possible.

3 Deepfake Categories

As we have seen, humanity has already been facing many forms of disinformation. Currently, deepfakes are entering the picture. To be able to estimate the effectiveness of counter measures against deepfakes, it is crucial to understand how deepfakes differ from previous forms of disinformation. Furthermore, to understand the strengths and weaknesses of contemporary deepfakes, we will also have a look into the creation process and methods behind deepfakes.

3.1 Deepfakes versus Cheapfakes

Let us first have a look at video manipulations which might be confused with deepfakes. Roughly, if a video has not been edited using machine learning (ML) methods, it is not considered a deepfake. Consequently, manipulations like adjusting the speed of a video[14] do not count as deepfakes. The same holds for videos that are spliced[15] or rely on lookalikes to create a misleading representation. Also videos, which are simply re-used and re-contextualized to give the illusion, that the depicted acts have been happening at a different time or location can not be classified as deepfakes. Rather, the videos that were created with the help of non-ML methods are usually called Cheapfakes[16] or Shallowfakes.[17] Within this thesis, the focus lies on deepfakes. The most widely-known deepfake category is a so-called face-replacement or face-swap. However, there are further forms of deepfakes. In the following subsections, these categories are described and the methods behind them are explained.

[13] Judging from e.g. the video in [20].

[14] E.g. the slurring of someone's speech to make them appear drunk as done in a case of the US politician Nancy Pelosi [32].

[15] I.e. videos, which are cut and added together.

[16] The term *Cheapfakes* is used to indicate that it is easy and cheap to create them.

[17] The term *Shallow*fakes is used to contrast it to *Deep*fakes.

3.2 Replacement

Via *face replacement* (also known as *face-swap*), it is possible to replace the face of a person in a given video, with the face of a different person. A malicious example would be so-called "involuntary porn" in which the victim's face is moved into a video that depicts sexual acts. In applications like the Snapchat face-swap filter, both individuals are in the same picture and both of their faces are swapped.[18] Different aspects of the two individuals are usually combined into one depicted human figure. A harmless example is the trend of the early days of deepfakes in which some people have replaced faces in various movie scenes with the face of Nicolas Cage.[19] Naturally, this method delivers more realistic results if the two individuals share characteristics like skin-color, hair type and head shape [35].

Figure 6: Nicolas Cage face-swapped onto the original face of Amy Adams [36].

Within face-replacement, the target is the person whose body is visible in the final video. Thus, the source's face is transferred onto the target's body. In face reenactment, the target is again the person that should be recognizable in the final video and the source's facial expressions are transferred onto the target's face. By now, there exist already techniques with which it is possible to transfer not only parts of a face, but even a complete (3D) head [37].

3.2.1 Inner Workings

To understand how a face replacement system works, we need to understand its components which include variational autoencoders, backpropagation and gradient descent. These components will partially be also used for the other deepfake categories.

3.2.1.1 Variational Autoencoder

Face replacement usually relies on a variational autoencoder (VAE). Typical for this type of neural network architecture is the goal to reproduce an input.[20] The most

[18] To view examples, see [33].

[19] There is a sub-reddit (r/deepcage) collecting such scenes [34].

[20] By itself, that might not seem to be of any use, but the clever ways in which VAEs are used, make them a powerful architecture.

simple way to achieve that would be to "mindlessly" copy the input to the output. However, like this, the system would not actually learn the relevant features of its training examples and consequently would barely be useful. To avoid this shortcut of transferring all the information, an information bottleneck is created. The system consists of two separate networks – the encoder and the decoder (see figure 7). The function of the encoder is to encode a provided image into fewer, but more abstract features in the latent space. We are forcing the neural network to focus on the higher level features of a face instead of conveying all pixel values. The resulting latent space representation is fed into the decoder. The decoder's function is to decode the latent representation with its more abstract features back to pixel values for the whole size of the image.

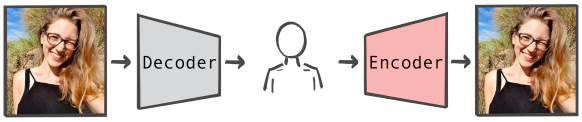


Figure 7: Variational Autoencoder (VAE)

3.2.1.2 Backpropagation and Gradient Descent

As for any neural network, the first batches of training examples through the whole VAE will result in output images that barely resemble the input images. But by implementing a fitting loss function and by propagating the calculated loss of an output (given the input) back through the network, it is possible to use gradient descent to adjust the weights within the network to minimize the losses. In a nutshell, it is calculated how the nodes of the last layer were responsible for the resulting loss and how they would need to be adjusted to reduce the loss. Then it is calculated how each node of the penultimate layer has contributed to the loss and how it should be adjusted. In this fashion, the loss is propagated backwards through the network. Then, the next round of training can begin. By reiterating this process for hundreds of training examples of human faces, the VAE becomes better at grasping the essence and characteristics of human faces.

3.2.1.3 Face Replacement System

For face replacement, usually two VAEs are used in combination (see figure 8). The first VAE is trained on images of the source (upper half of the image), while the second VAE is trained on images of the target (lower half of the image). Crucially, the two VAEs share the same encoder. Once both VAE achieved satisfying results for recreating their inputs, their decoders are exchanged (see figure 9). We now feed an image of the source into the encoder to arrive at a latent representation of the face, but use the decoder of the target to arrive at the final image. Consequently, the target's facial movements are preserved, but the facial features mimic those of the source. For instance, if the original video showed the target smiling, in the new deepfake, the depicted person would resemble the source while still smiling.

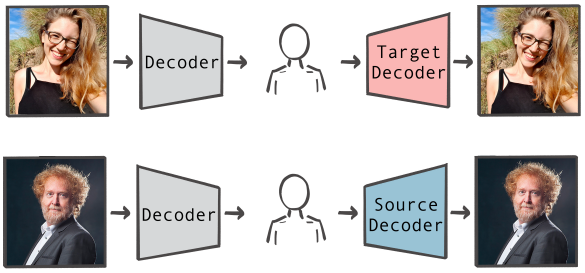


Figure 8: Face replacement, training phase.
(Picture incorporated from [38])

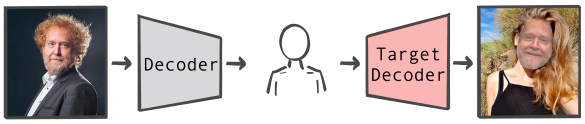


Figure 9: Face replacement.
(Pictures incorporated from [38], [39])

3.2.2 Limitations

A major limitation of replacement systems is that they tend to require a lot of pictures of the source as well as of the target in order to be trained successfully. Furthermore, the accumulated images need to be pre-processed before the actual training can begin. This usually includes to crop the pictures around the face. There is software to help with this step, but it is still an additional complication and results of the cropping should be examined to ensure their success. The actual training of the VAEs tends to require a fair amount of computational resources [40]. After the training, non-ML post-processing needs to take place in order to align the generated images with the position and angle in each video frame. Naturally, many researchers are working on mitigating these limitations. However, currently, they still present a hurdle for creators.

3.3 Reenactment

Lip syncing focuses on the lips of the target and is a common form of face reenactment. Generally, face reenactment can be thought of as a form of puppetry. If one has already a video of the target person (i.e. the puppet), it is possible to change their facial expressions within that video [41]. The facial expressions are transferred from the source to the target. Within reenactment, the source is also sometimes called "the driver" because the source drives the behavior of the target. A harmless recent example is the deepfake in which world leaders are portrayed as singing John Lennon's "Imagine" [42]. Thus, rather than changing who is depicted, the originally depicted person can be represented as "saying" things they never said. We can not only transfer lip-movements and other facial movements of the source, but also their torso and head motion [43] and even full body motion as visible in figure 10.[21] However, these are all only changes for our visual experience. To change the audio of the manipulated video, it is possible to either use a natural voice (e.g. of

[21] Similar examples include [44], [45]

a voice actor) or to artificially generate the voice (see section 3.4.1).

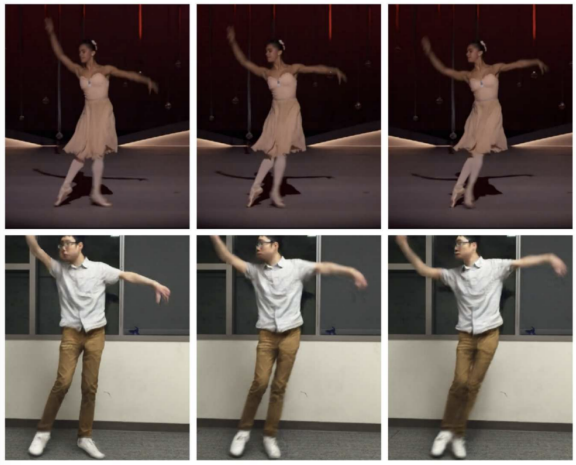


Figure 10: Motion transfer of the full body. From a provided YouTube video of a ballerina (source) and a picture of a student (target), the motion is transferred to the target. [46]

3.3.1 Inner workings

It should be mentioned that there is a great number of approaches to reenactment. For the purposes of this thesis, we do not need to delve into all of them or into great detail. Thus, we introduce only a few of these approaches on a higher level of abstraction. Face reenactment often relies on similar autoencoders as face replacement does. However, reenactment has the advantage of only needing a video of the target rather than needing many preprocessed images [47]. Other approaches track the facial movements and transfer them in real time onto the target. In figure 11, you can see the inputs of the source and the target. The transfer is the tracked and transferred information. Finally, an image of the output video is provided. Furthermore, some approaches achieved lip syncing from only an audio file (rather than from a video). To generate content, that displays temporal consistency, they made use of a (surprisingly simple) recurrent neural network architecture [48].

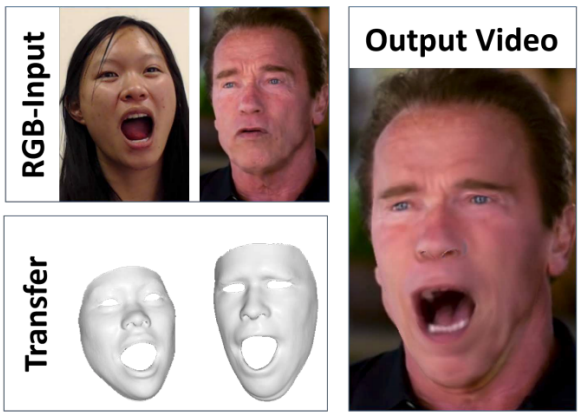


Figure 11: Face reenactment [41]

3.3.2 Limitations

Conventionally, a great number of training material is required to create photo-realistic videos. As mentioned beforehand, this is often a limitation. However, the efforts that have been going into solving this constraint are starting to deliver results. It is already possible to create somewhat realistic reenactment pictures when having only one single photo of the target [49]. While this approach works only for pictures, the authors mention the plan of approaching video transformations in the same manner.

Furthermore, it has often been the case that the identity of the targets has only partially been preserved during the reenactment. However, there are current approaches, which aim to change that while also requiring only a small amount of input images [35]. One last limitation, that is now being tackled, is the difficulty of representing emotional facial expressions. However, significant progress is also being made in this regard (see figure 12 [50]).

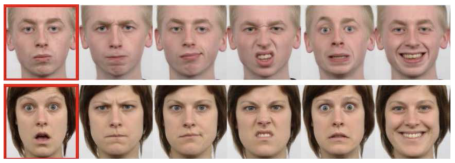


Figure 12: Emotional facial expressions [50]

3.4 Generation

In this section, we delve into methods that generate a complete visual or audio experience of a person, who does or does not exist. With a type of neural network architecture, called GANs, it is possible to generate pictures of people, that have never lived.[22] GANs have seen impressive progress over the last years (see figure 13). Figure 14 shows a higher resolution sample of current possibilities.[23]

Figure 13: GAN progress [51]

For malicious uses, it would probably mostly be profitable if the portrayed person would be an existing one and if only a minimal amount of information about the target would need to be provided. To the best of our knowledge, it is not yet possible to give an AI system a text, a short voice sample and a small number of pictures to arrive at a fully generated video. It is however possible, to input only a few images of the person that should be displayed

[22] An explanation of what GANs are and why they work can be found in the next subsection.

[23] You can find more examples by loading the page *thispersondoesnotexist.com* [24] which will generate such picture for you.

Figure 14: Face generation [24]

(1 to 8 pictures) to seemingly turn a photo into an identity-preserving video (generated as video frames) [52]. Examples of this method can be seen in [53].

Text-to-speech systems are known to have traditionally produced somewhat rough or unnatural speech which has been lacking variation of emphasis within and between sentences. However, the field has recently advanced heavily. So much so, that for the current state-of-the-art synthesized audio, human perception can have a hard time to distinguish artificial from natural human speech. Some such examples can be found in [54] from the Tacotron 2 model or in [55] from Lyrebird AI, a research division of the software company *Descript*.

3.4.1 Inner workings

In 2014, the revolutionary framework of Generative Adversarial Networks (GANs) was proposed [56]. Since then, most completely-generated images are created by GANs. The image of figure 14 is a sample from a state of the art GAN from 2019 [57]. Up until recently, the creation of audio samples had proven itself challenging. Raw speech audio has usually around 16,000 sample points per second [58]. Furthermore, there is the need for temporal consistency on various time scales. Consequently, this subfield was dominated by concatenative text-to-speech systems. Here, a human speaker records all required phonemes. These pre-recorded phonemes can then be concatenated in the required order to form the desired words and sentences [59]. However, this approach left one wishing for more smoothness and more natural intonations. Furthermore, for a new voice, a whole new database of phonemes needed to be recorded. DeepMind's WaveNet made it possible to avoid these downsides and to generate all audio samples from scratch [58]. And recently, the WaveNet and GANs have been combined to synthesize authentic speech [60].[24]

[24]This is the approach of the beforementioned Lyrebird AI.

GANs include two networks – a generator and a discriminator. The generator is generating new content (e.g. images of faces) by capturing the distribution of the training examples. The discriminator is classifying the new content as either an original, true content or as generated content. In such an adversarial setup, the generator is trained to trick the discriminator into mis-classifying the generated output as real content, while the discriminator is trained correctly discern real from generated.

As typical for neural networks, the training proceeds via back propagation of the calculated loss. However, the relation between the two networks makes successfully training a GAN less straight-forward than training other neural network architectures. The training begins with a round of discriminator training (see figure 15) aiming to distinguish real pictures from random noise. During this period, the calculated loss in backpropagated through the discriminator network. Then, the generator is trained (see figure 16) to generate output, that is closer to the real pictures than random noise. To update the weights in the generator network, the loss needs to be back propagated through the discriminator before it can reach the generator network. To make the task not even harder for the generator, the weights of the discriminator are not adjusted in this phase. Each period of training may last one or more epochs before the complementary network is trained further.

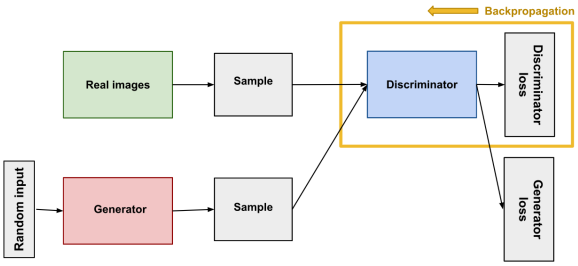


Figure 15: Face generation [61]

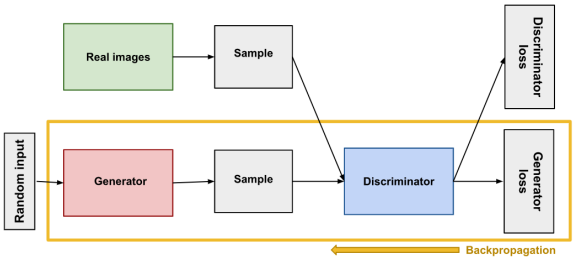


Figure 16: GAN generator training [62]

Training a GAN is a tricky task because the two networks need to be trained separately and need to be in balance – once the generator is too skilled for the discriminator, the discriminator only assign a probability of 50% that the generated output is real. Thus, the generator does not get valuable feedback anymore and cannot improve further.

To conclude, handling GANs successfully can be challenging. However, let us keep in mind that GANs are only one breakthrough in the idea space of neural network architectures and how this new type of architecture has enabled such significant improvements. GANs were not a change in the fundamental principles of how neural networks function, but they nevertheless could remarkably improve the state-of-the-art capabilities. Currently, there are still some limitations (to which we will get next), but we want to keep the potential for such drastic improvements in mind if we want to be well prepared with our proceeding of deepfake counter measures.

3.4.2 Limitations

Currently, these images show remarkable details and look quite photo-realistic. However, it is still difficult to have control over generation of the images. In other words, it is hard to generate an image that fulfills pre-set requirements regarding hair style, pose or the shape of the face. This is being tackled [57], but there is still a lot of progress to be made.

The results for videos are not yet as photo-realistic as those in figure 13. Moreover, the video frames do not extend far around the face region and the background is very similar to the background in the provided photos, but these are fully generated video frames of existing people. Until recently, it was impossible for these methods to generate a subject from angles and with facial expressions, that the model had not seen beforehand. This has changed. From only one input image of the target person, it became possible to generate identity-preserving images on which the person's head is seen from a wide range of angles. Furthermore, the person's facial expressions can convey a wide range of emotions [50]. Because the videos are created by generating video frames, the videos still lack some temporal consistency – face boundaries for example exhibit a lack in stability and are somewhat "trembling". If one looks closely, one can spot some trembling or a lack in stability of boundaries.[25] To generate indistinguishable full videos (including audio) from scratch, visual and auditory consistency would need to be maintained.

Furthermore, text-to-speech synthesis is impressive for calm sentences. A current limitation is that emotionally charged speech (e.g. a sad trembling voice), as well as variations like screaming or whispering are not yet possible to generate with the current combination of model and training data. Overall, it is definitely not yet easily possible for anyone to generate a realistic overall video.

4 Social Media and Traditional Journalism

Social Media (SM) and Traditional Journalism (TJ) can be considered as our main interfaces to news. Section 1.1.4 provided provisory definitions which we want to extend in this section. We need to understand better in what aspects SM and TJ differ, in order to later analyze the effectiveness of various counter measures within the two different media environments. For this purpose, we examine three stages – the creation of content, the consumption of content and the spread of content.

4.1 Content Creation

In this section, we analyze first which kind of platforms are associated with the respective media environment and what goals are related with content creation. Next, we explore the differences in standards and quality. Finally, we inspect what happens when there is suspicion of misconduct with respect to content creation.

4.1.1 Platforms and Goals

As mentioned in the provisory definitions in section 1.1.4, there are different platforms associated with the respective media environments. TJ can be roughly partitioned into TV, radio, print media and their online equivalent.[26] Journalists in each of these partitions have the primary goal of reporting news and inform their audiences.

SM can be considered as a broader term. There are different ways of how to classify SM. E.g. Andreas Kaplan distinguishes four types of (mobile) social media applications, focusing on location or time sensitive aspects [63].[27] However, we think that it is useful to have a distinction of SM applications on the basis of how they are used – more privately or more publicly. Most SM applications can be assigned to the public category which includes common applications such as Facebook, Twitter, YouTube, Reddit, Quora and Weibo. In general, we can consider all SM applications as public with which you can create and share content in a public manner. E.g. Tweets will usually be made public to the entire world. A post on Reddit is accessible to any user. The same holds for Facebook and the Chinese equivalent Weibo where you have can share details about your life with your friends and the world. The private category of SM applications includes services such as WhatsApp, Telegram, Snapchat and WeChat. These applications work on the basis of sharing personal messages. This distinction will be relevant in section 6.2.1.2 in which we discuss the effectiveness of detection counter measures.

The goals associated with SM are more diverse than the ones in TJ which is used as news source. Next to being used as news source, SM can also be used for interpersonal relationships and self-presentation. Additionally, it can be used for recruitment, law enforcement, pol-

[25] See e.g. [52].

[26] Although all of these partitions exhibit some significance, the most relevant one will be the online version of news channels and former print media.

[27] An application can e.g. be location *and* time sensitive where the users interact with respect to a specific location at a specific time period (e.g. Facebook Places [64] and Foursquare [65]). An application can, however, also be only location sensitive such Yelp [66] or only time sensitive such as Twitter [67] or neither such as YouTube [68] and Wikipedia [69].

itics and education. The interpersonal relationships are strongly connected to private SM applications. The remaining goals are usually associated with the public category. Note that the goal of consuming or reporting news is only one among many. This will become important in our later analysis.

4.1.2 Standards and Quality

TJ content is produced by journalists who have enjoyed professional training in this task. In this highly competitive industry, it is often required to have accomplished the minimum of an undergraduate degree in communication science or journalism [70]. The requirements differ per country but some kind of rigorous and supervised phase has to be gone through. Furthermore, TJ content has to live up to various standards when being proof-read and approved by an editor of the corresponding outlet (see for example [71]). These forms of filtering and improvement are missing within the user generated content (UGC) of social media. On social media, it is very easy to participate.[28] Anyone that has access to the internet, is able to use a platform like Facebook or knows how to use a forum, can create content that is in principle widely accessible. The minimal requirements are limited to a very basic form of media literacy. However, it would be a gross simplification to not elaborate more on the nuances of some platforms. Take for example YouTube. While it is true that even a kid with a phone camera could run a channel where it uploads its content, YouTube's landscape looks more differentiated. Next to the presence of TJ outlets on this platform, we can also find other channels that exhibit a high level of professional journalism and production value. Not rarely, it is the case that such channels have a big production team whose members underwent higher-level education in journalism, media sciences, communication sciences, politics, film making, animation, etc.[29] The bottom line is that the standards and quality claims of SM range from very low to very high – both with respect to inter-platform[30] and intra-platform[31] differ ences. TJ has little intra-platform differences while exhibiting medium inter-platform differences.[32] An additional related factor to the level of requirement, is the number of authors or content creators for the two kinds of media. Within SM, almost every user is or can be a content creator while within TJ only a few selected individuals are doing the same.

[28] For now, we omit the private SM applications and focus on the public ones.

[29] There are countless examples for such channels. Two prominent examples include *Kurzgesagt – in a Nutshell* [72] with over 10 million subscribers and *Veritasium* [73] with more than 6 million subscribers. Both channels create high-quality videos explaining societal or scientific phenomena.

[30] Between different SM platforms such as Facebook and Wikipedia.

[31] Within one platform such as YouTube but between different channels.

[32] Consider for example the differences in journalistic rigor between the TJ institutions of *The New York Times* and tabloid newspapers like *The Sun*.

4.1.3 Misconduct and its Consequences

What are the consequences of reporting false information in SM and in TJ, respectively? We could imagine how journalists themselves as well as their outlets are facing the potential of harsh real-world consequences from sharing false content. The same could, however, be true for content creators in SM. Disappointing the follower base (e.g. on Facebook, Instagram or YouTube) by feeding them wrong information, could also lead to a backlash which would damage their reach.[33] Thus, the difference lies not in what the final consequences will look like but what the exact procedure is that leads to the consequences. In SM, there are usually no proper procedures in place. Wikipedia with its editing and review system might be an exception. Content creators in most other SM applications live from their direct connection to the follower base and try simply avoid to fall into disfavor. Media outlets have a similar possibility of falling into disfavor. However, they have rigorous procedures that are based on national or international codes of practice. To uphold such high standards, a consequence can look like in the case of photographer Hajj who was fired because he manipulated an image (see section 2.3). But in the case of plagiarism by youtuber Siraj Raval, the lack of formal procedures led to a mere loss of followers. Raval is continuing to create his content.

4.2 Content Consumption

In this section, we discuss first how the ratio of SM and TJ use looks like when focusing on online news consumption. Then, we explore how deep the respective media environment consumption is.

4.2.1 Relative Use of SM and TJ platforms

How people access their online news falls into six different categories: direct, search, social media, mobile alerts, aggregators[34] and email [75]. In northern countries like Finland, Norway and Sweden, most (64%) say their main way of a access is direct (i.e. via TJ websites) while the remaining 36% is spread over the rest. In countries like Chile, Brazil and Malaysia, the situation is very different. Here, social media clearly dominates the picture with being the main news source for 42% of the population. The third category uses mainly search and aggregators and is being located in South Korea, Japan and Taiwan. Within the USA, Canada and Australia, the population is mostly split over direct access (27%), social media (25%) and search (20%) [75]. Thus, depending on the country, SM use for news consumption varies from 9% to 42%.

[33] A recent example involves the case of Siraj Raval who runs a YouTube channel on educating people on how to write ML algorithms. Raval plagiarized a paper to boost his credibility. When other researchers started to point it out, Raval confessed and lost a part of his follower base [74].

[34] Aggregators are apps such Apple News+, Google News and Flipboard.

4.2.2 Depth of Consumption

All created content is consumed by a user. As we have seen, this may happen via SM platforms or on dedicated websites of the TJ outlets. An important distinction here is that with the endless stream of posts in SM, the user is nudged to mostly scroll through the headlines without actually reading articles. A study from 2016 showed that only 59% of Twitter users read a news article on this platform before sharing it [76]. If there is a fire-hose of interesting new content, why would one strain oneself by reading full articles? Full articles cost more willpower to read (or maybe even to reflect on or check its sources and reasoning). At the same time, one would feel the "missing out" on the gist of all the following headlines that could skim additionally if one would not be reading full articles. Given limited amounts of time and attention, it does seem to make sense to "first" skim all the headlines to get an overview, before using time to go deeper into any topic [76]. The intent behind headlines is to attract our attention and to make us click to read the full article. Headlines tend to be one of the most catchy statements or questions possible given the well-researched article. This is the case whether articles are shared on the media outlets website or on social media. However, on websites of media outlets, the nudge to mostly scroll through the articles might be less pronounced as some articles are already opened without any extra click and the suggested articles are maybe not corresponding to the general interests of the user, such as an SM feed would [77]. Thus, the tendency to consume mostly headlines might influence, what we finally take away from a given article and might be more extreme when we see that article on social media. A further contributing factor is that social media content often comes with the personal opinion of the person who shares it (e.g. tweets within Twitter), while TJ articles are usually aimed to stay more objective.

4.3 Content Spread

In this section, we discuss how TJ news spread. As already touched upon, the articles from TJ may be shared on their own respective platforms. At the same time, news outlets have been trying to use social media platforms in their favor and have been posting their articles there as well. In comparison, the content produced by social media users, is not (to the same extend) shared on the sites of traditional journalism.

4.3.1 Conventional Spread of News

As mentioned earlier, websites in TJ are sometimes frequented directly by the users. Similar to the print medium, TV and radio, users have their favorites platforms that they visit regularly to consume their news. In this case, the users has to know which of the TJ online outlets they need to visit. In this case, the creator offers their content on their own platform, hoping that a user will find it by usually coming back to the trusted website. This is reminiscent of offline TJ in which users e.g. regularly buy a specific newspaper.

4.3.2 SM Spread

In SM applications like Facebook, Twitter, YouTube and Instagram, content can easily be shared. Whether it is a written status update, a self-taken image, an edited video or a blog post, users can share these on their channel or account to enables others to consume this content. On SM websites, the user is not only the consumer, but also the content creator and the leading force that determines which content is boosted by the underlying platform algorithms to be more prominently presented to other users. Given the nature of the platform, users can like or share content and thus increase the probability that others (e.g. their followers or friends) will access the content as well. In the pure TJ case, users can contribute to the spread of news by recommending them to others directly. This is however a slower process and the reach is quite limited. The main spread happens via the TJ outlet. In SM, users decide what is worthy of spreading. Content can go viral in a short amount of time if users interact heavily with the content. This has been a user-centered view on the spread of in SM. If we take a content-centered view, we can look at memetics. According to Richard Dawkins, memes are selfish replicators on the informational level [78]. For instance, a tweet can be considered a meme which "wants" to replicate and spread.[35] For the meme, it is not relevant whether it is true or useful to a host or not. Current studies suggest that false information tends to spread faster on social media than true content [79]. Furthermore, the veracity of the median true rumour is settled within 2 hours while the median false rumour stays unresolved for 14 hours [80]. Memes want to spread and SM offers a hub in which such memes can replicate very cheaply and quickly – usually via a single click. Sensational and extreme memes can usually spread easier than nuanced and thorough memes.

4.3.3 Inter-dependencies of SM and TJ

As touched upon beforehand, there are interactions and inter-dependencies between social media and traditional journalism. TJ outlets are using SM platforms to reach more readers. Furthermore, already in 2009, the rise of social media influenced the focus of TJ from aiming to be the first to report on a given story to attempting on high-quality verification and contextualizing of circulating stories and rumours [81]. Indeed, there are textbooks for modern journalists, which aim to teach about SM-related contents [82]. SJ is widely used to spread TJ content. It is rarely the other way around.

In summary, the content creation process overall favors TJ over SM in terms of promoting veracity. Among factors like the higher number of authors within SM, this is the case because many SM users value personal signalling more than spreading confirmed content. Furthermore, there is often a default nudge towards superficial consumption within SM. This nudge and the recommen-

[35] In reality, memes are not agents. They cannot want anything. But this anthropomorphism is commonly used in the literature to explain memetics in a more intuitive way.

dation algorithms of SM facilitate the viral spread of misinformation and deepfakes. Let us now explore what could be done to hinder such "digital wildfires" of misleading deepfakes.

5 Counter measures

A wide range of counter measures has been suggested and it is easy to get lost in the various approaches. The most comprehensive overview document, that we could find is the work by Sam Gregory for the human rights non-profit organization WITNESS [83]. We tried to build on his work, to improve its overview and to add more up-to-date project instances.

To bring order into the entanglement of methods, we decided to roughly categorize the counter measures by discipline. The four coarse-grained categories of solutions are technical, educational, legal and normative. In the following, relevant information is given for each of these categories separately.

5.1 Technical Solutions

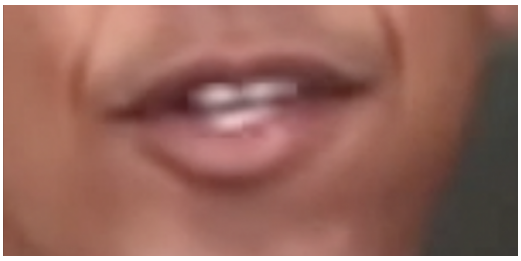
The technical solutions are by far the most developed category and can be further split into detection, authentication, provenance (or phylogeny) and anonymization. Detection is the most straight-forward approach and it probably comes to mind first when contemplating about how to solve the malicious uses of deepfakes. A lion share of the efforts have been going into this approach (including for example the DARPA MediFor project [84]). It makes sense, that this approach is so widely recognized. If one would like to label deepfake videos as such or to take them down, it is an obvious requirement to first be able to detect whether a given video is a deepfake or not.

5.1.1 Detection

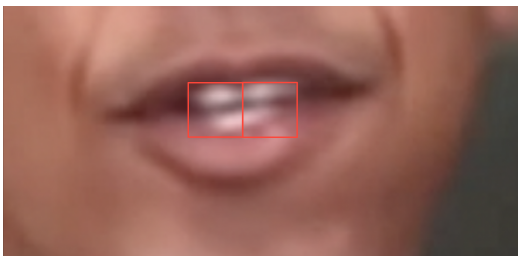
Detection methods generally focus on at least one characteristic in which deepfakes (currently) tend to differ from genuine videos. Such characteristics can be based on what we perceive directly or on details that are not directly perceivable for humans.

5.1.1.1 Directly Perceivable

Aspects that we can perceive directly are the images or pixels themselves, as well as the created sound waves. In deepfakes of lower quality, we might be able to spot that the borders of the face do not align neatly or that there are small artifacts which make the teeth look a bit off such as box artifacts. An example for this can be seen in figure **??**. Furthermore, there might be inconsistencies in the lightning or other ways that seem to "break physics" or we might be able to use biological indicators, such as whether the depicted person is blinking [85]. Such an artifact is due to the nature of the data that the algorithm was trained on. When deepfakes algorithms are trained on a large quantity of images, they usually use images which contain people with open eyes. The algorithm will thus not be very familiar with closed eyes. And this will result in imperfections around the closed eyes. If a human is doing such detection, this is the very close analogy of the historical counter measure where humans tried to find inconsistencies in aspects like angles or lightning. Thus, this historical counter measure still works for deepfakes of lower quality.



(a) Box artifacts raw [13]



(b) Box artifacts pointed out [13] (boxes added by us)

Figure 17: Example of box artifacts [86]

5.1.1.2 Indirectly Perceivable

At the same time, evidence may be found in the pixel or audio information itself, while it is hardly perceivable to the human eye or ear. For instance, biological indicators can be used to detect the slight changes in skin color caused by humans natural pulse [87]. For high-profile individuals with a lot of video material that is publicly available, it may add some protection to let a neural network learn their (often barely noticeable) idiosyncrasies in order to distinguish fake from real video content [88], [89]. Instances of such idiosyncrasies might be how a person pronounces specific words or what their characteristic head movements are when speaking.

In audio material, there can be spectral correlations, which are alien to human speech and which are not directly perceivable to the human ear. However, the information is nevertheless directly in the auditory information and can be picked up and explored digitally. Analogously, there may be sound traces which are above the pitch range that humans produce [90]. This is also an ongoing field of research that is being advanced by Google, for instance in the form of creating databases [91] within the Google News Initiative (GNI) [92] or funding projects under their Digital News Innovation (DNI) Fund [93] (e.g. project *Digger* [94]).

5.1.1.3 Imperceivable

An example for characteristics that are definitely not perceivable when watching a deepfake, is meta-data of the video file, such as the camera identity and fingerprint.[36] The interesting details of how this works are beyond the scope of this thesis, but more can be found in [95] and [96] for example. In a nutshell, the photo-response non-uniformity (PRNU) of the camera sensor leaves a specific noise pattern on the produced photos [95]. Interestingly, GANs do the analogous and add their own fingerprint to a created picture [97]. This method is robust to cropping or compression of an image, but a specific fingerprint can maliciously be added to photographic material [98], [99].

Most current detection methods are very similar in architecture and working to the creation methods – they are also GANs[37] or generally different flavors of neural networks.[38] To reliably discern deepfakes from unaltered videos, these detection methods require vast amounts of training data. Training examples are provided by data sets like FaceForensics [100] for which the FakeApp tool was used to generate fake examples, and via databases that major tech companies like Facebook and Google contributed [91], [101].

5.1.2 Authentication

Another promising method is authentication. If something is authenticated, it is proven, that it is "real, true or what people say it is" [102]. It is often used when people mean that they verified the identity of a person. In the context of potentially doctored images and videos, what is being proven is usually that the given material has not been tampered with as well as the exact time and location of its capture.

5.1.2.1 Current Projects

Multiple apps that use this technology are currently in development and have made prototypes or beta-versions available. Examples of such projects are Serelay, the eWitness tool, Amber, ProofMode, TruePic and Eyewitness to Atrocities. To get an impression, let us briefly have a look at the first two of these. Serelay aims to let you take pictures and videos whose content, location and time of capture are verifiable. This can be useful for example for documenting riots, police violence, corruption, traffic violation or domestic violence. The eWitness tool is an app and has a very similar approach and result. It makes the meta-data of the taken photo immutable by using a blockchain model (a permissioned chain) [103]. Currently, a prototype version is available on the Google Playstore.

5.1.2.2 Blockchains

As mentioned, in some of these applications blockchain technology is being used. For the sake of understanding these counter measures, it is sufficient to keep a few characteristics of this technology in mind – being decentralized and being immutable. Blockchains are a way to store information on a multitude of computers. Thus, once a block of information is added to a blockchain, it is very hard to change any information within the chain because there are many copies of the chain spread over a great number of machines.[39] Furthermore, each block has a unique hash. A hash is a series of letters and digits, which is the result of applying a specific mathematical function on the stored information of the block. We may think of it as the unique name of that block. Thus, each block has its own hash, as well as the hash of the previous block. This hash systems makes it again harder to change information that was added to the blockchain because changing the information within a block would also change its hash [104]. We can thus say, that the blockchain is immutable – it is practically impossible to change information that was added to it. A permissioned blockchain adds a further level of security by only allowing certain individuals to add to it [105].

A somewhat related approach, that may be worthwhile for certain high-stakes individuals is to sacrifice their privacy by setting up an immutable life-log. Chesney and Citron argue that if it would become trivially easy and cheap to create high quality deepfakes that are indistinguishable from untampered videos, individuals whose success is very sensitive to their reputation might opt for services that offer immutable life-logs as a potential alibi service [106]. Such life-logs may include video, audio, location or information from wearables. This could allow these individuals to quickly counter disinformation. But this could be a potentially unsatisfying solution as there is a huge trade-off between *protection from deepfakes* and *privacy*.

5.1.3 Provenance

The goal behind provenance analysis is to discover the origins, the path and the changes of a given piece of media [107] – be it a painting as in the origin of this term, or more currently photos, videos or internet memes. The chronology is often given in a provenance graph.

5.1.3.1 General Principle

To understand the workings of provenance, let us have a look at the example provenance graph in Figure 18. Imagine you found the query picture on the internet and it to be somewhat edited.With provenance analysis, the image would be compared with images of a database. Using the picture, one would for example apply a similarity search within the Internet Archive [108] and find images that are similar to the input picture. The retrieved pictures are ranked in their similarity to the query image. Ideally, this would lead you to separate images that were used for different parts of the picture. From these pictures, a provenance graph is constructed [107]. If you have been com-

[36] This can be considered as some form of digital watermarking.
[37] See section 3.4.1 for more details.
[38] E.g. recurrent, convolutional or capsule neural networks.
[39] E.g. a copy of Bitcoin's public blockchain is present on millions of computers.

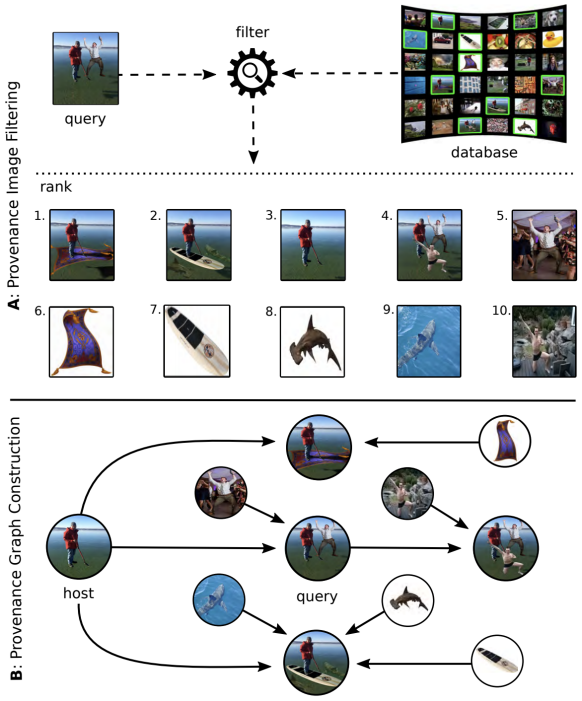


Figure 18: Example of a provenance graph.
Image credit: [107]

ing across this picture somewhere online, you might also be interested in its path (or the path or its original parts) since the first upload. It becomes possible to trace back the path of a given image or video, when also meta-data such as camera ID, geotags and timestamps are being used [109]. With the help of provenance, it becomes possible to check whether a given video can be traced back to reputable, trusted sources.

An example of an undertaking, that also makes use of provenance is the European InVID project [110]. InVid is aiming to provide verification services that are easily available like their browser plugin [111] to debunk or verify images and videos.

5.1.3.2 Potential Application to Deepfakes

Especially in the case of the deepfake categories of replacement and reenactment, provenance might become useful. As a deepfake needs in these cases at least one source video, a similarity search can be applied to find the original source video. Imagine for example, a deepfake that shows some person T saying something that would damage his reputation. The deepfake at hand is a face replacement. Applying a similarity search could yield the original video showing person S's face. The body, clothes and background can be identified in the original picture as they are they are identical in the deepfake. Even the spoken text is the same, however, with another voice. Such a provenance approach could help finding the origins of a deepfakes.

However, this approach does not seem to work when we look at generation. If e.g. a GAN is trained on a multitude of data, it can then generate a new face image. Here, provenance can not be used as the similarity of the generated content is not close enough to the original ones. The GAN abstracts from its training input and generates new images. No provenance technique could direct you to the used training images. This approach would probably be futile in the generation category.

5.1.4 Anonymization

Anonymization of images and videos (including deepfakes) can either make the depicted person only unrecognizable by facial recognition systems, or it can go further and make the person also unrecognizable by humans. These different amounts of anonymization are useful for different situations.

5.1.4.1 Partial anonymization

If a deepfake of a person exists, the harm is done because people still recognize the depicted person even though the video was edited. So, how could it be beneficial to have partial anonymization, which only hinders facial recognition systems (i.e. humans are still able to recognize the depicted person)? This type of anonymization might be useful to hinder the creation of deepfakes in the first place (rather than its effects once it exists). If someone is trying to create a deepfake of you, they will likely first search for photos of you to train their model with it (see section 3 for details). Thus, they are probably searching online to find many pictures of you. To do this, they might use facial recognition systems in order to quickly find many pictures of your face. However, if facial recognition systems fail to recognize you in pictures, you could force them to manually do this step. The hope is that you could prevent attackers from creating realistic deepfakes of you in a quick way.

How could one hinder facial recognition systems while not bothering the human eye? This could be done via so called adversarial attacks. Facial recognition systems, and neural networks in general, can still be mislead by pixel perturbations, which are not perceivable by the human eye [112]. Figure 19 provides such an example. An image classifier is supposed to classify what object the input image is depicting (left most image in figure 19). In the given example, the classifier predicts "panda" with a confidence of 57.7%. If the adversarial input (middle image in figure 19) is added to the original image, the classifier has suddenly a confidence of 99.3% that the image depicts a "gibbon". While the artificial classifier is fooled, the human classifier is not. For the human eye, nothing changed and the panda is still a panda.

5.1.4.2 Complete Anonymization

Complete anonymization stops computer vision systems as well as humans from recognizing you in a video. One possibility to achieve this is to replace the original face with the face of another person, who gave informed consent – this is similar to the common face-swapping used in

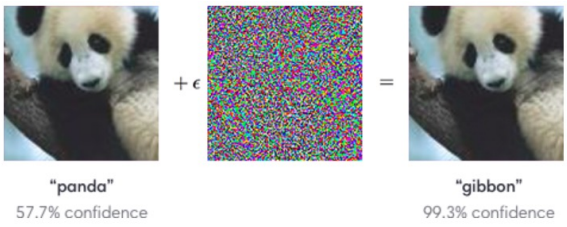


Figure 19: Adversarial input [112]

many deepfakes as well as famously in Snapchat filters.[40] Another possibility is to replace the original face with one, that was completely generated – thus, with a face that has never existed. There are however powerful benefits to be reaped from this technology (e.g. see [113]). The current examples (such as figure 20) are definitely still recognizable as being deepfakes.

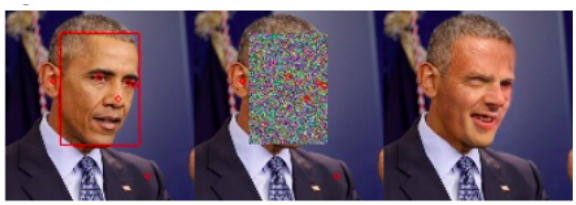


Figure 20: DeepPrivacy example [113]

However, highest quality is not necessary to fulfil the function of hiding the original face. With such complete anonymization, it would for instance be possible to use recorded videos of crimes while granting more privacy to the victims by fully replacing their faces. Another example would be recording interviews with whistle-blowers, human right activists or other individuals that want to get a message across while not exposing their own identity. Like this, complete anonymization can be a very valuable tool. However, according to our current knowledge and ideas, this form of anonymization can do little to nothing about the risks from deepfakes.[41]

5.2 Educational Solutions

After having introduced some technical solutions, let us face another potentially promising approach – educational solutions. For this purpose, we talk about the medium of delivery and the content separately.

5.2.1 Medium of Delivery

Given the literature, we identified three media of delivery: articles, games and workshops. In the following, we will discuss them one by one.

[40] See in [33] for a few failed examples

[41] People that became victims of a deepfake creation, might decide against suing the responsible person because they do not want to attract more attention to the deepfake. This is also called *Streisand effect*. Here, anonymization might help by decreasing this personal inhibition. However, we do not see anonymization applications to prevent harmful deepfakes.

5.2.1.1 Articles

Most educational resources so far seem to be in the form of regular articles or websites. Here anyone interested can read, try to improve their skill of discerning real from fake and maybe even test themselves. Some example resources are from BuzzFeed [114], The Washington Post's guide [115], as well as the websites of NiemanLab [86] and the Oxford Internet Institute [116]. Currently, a downside of such valuable resources is that internet users only rarely stumble upon them. This might be improved by linking a similar resource right where users are confronted with potential disinformation. If SM platforms would integrate labels to indicate the estimated probability that a given article or video is an instance of disinformation, such a resource could be available with the label.

5.2.1.2 Games

Google takes a somewhat different approach and tries to teach children in an online game *Interland* [117] about various (fundamental) skills like handling phishing attacks, passwords, responsible sharing behaviour and recognising fake and dubious content. Such games can e.g. be used within the classroom. However, there is still significant uncertainty around how deepfake related media literacy should be communicated [83].

5.2.1.3 Workshops

A further approach could be to organize workshops for specific target groups. Example target groups may be journalists[42], editors or SM content moderators. The workshops could be initialized by organizations like WITNESS or by TJ or SM platforms themselves. More general deepfake workshops have already been hold by WITNESS [119], the United Nation's center fro AI and Robotics (UNICRI) [120], [121] and university institutes (e.g. [122]).

5.2.2 Content

One crucial point to consider is also the kind of advice that is given. More specifically, we can differentiate between tactics, which are useful in the short-run (say, in the next few months) or more long-lived value that is useful for years to come. Tactics for the short-run focus on the current "Achilles heel" of deepfake creation – for example on a lack of blinking. More sustainable tactics include for example checking the source of a given video. Focusing on short-lived characteristics of deepfakes, can easily be harmful in the long-run as people are likely to be lulled into a false sense of security. One feels like one has the tools to recognize fakes, while that is not anymore the case. Consequently, one would be *more* likely to believe and share false content. Furthermore, it might be helpful to use some of the before-mentioned medium to raise general awareness among the population that such technology is already out there.

[42] Like the new online course by Reuters and Facebook [118]

5.3 Legal Solutions

Another way of avoiding harmful consequences of deepfakes is via official regulations and laws. Can we make the creation or the spread of harmful deepfakes so unattractive, that a sufficient part of the attackers would refrain from malicious actions? In this context, we discuss the aspects: liability, administrative agencies and bans.

5.3.1 Liability

One option is to consider liability law. There is civil liability law and criminal liability law. Civil liability law is applied when one citizen sues another one, for example for not acting according to a specific contract or for caused injury or loss. Criminal law on the other hand covers cases in which a person acts against the criminal code, for example forced entry in a house or murder. In such cases, the government is prosecuting that person [123]. Chesney and Citron wrote a thorough review of various such options within the US [106] – from laws around copy-right, defamation as well as privacy-focused laws and criminal liability law. They concluded that most options are not fitting in their applicability or face practical hurdles. An example of a practical hurdle is that legal cases tend to attract attention, while many victims want to avoid more publicity around videos that represent them in disapproved ways. Another one can be the costs of the potential trial and the burden of proof [106]. Furthermore, there is the tension with the right to freedom of speech and there have been precedence cases in which courts have abolished attempts of prohibiting election-related lies in order to protect free speech (e.g. [124]). While some individual cases might benefit from liability laws around deepfakes, the options within this domain fail to address systemic and more indirect consequences [106].

Within Europe, Germany has recently been passing the Network Enforcement Law ("Netzwerkdurchsetzungsgesetz"). This law is focused on limiting hate speech and obliges social media platforms to take down reported hate speech within twenty-four hours or face fines. The reasoning for this law also mentions the mitigation of fake news [125]. Consequently, an extension of it or an analogous law for deepfakes or for fake news in general seem possible.

5.3.2 Administrative Agencies

Another option is to turn to administrative agencies like the European Commission. Within the US, this would be agencies like the Federal Trade Commission, the Federal Communication Commission and the Federal Election Commission. However, the responsibilities, goals and possibilities of these agencies do not fit well with the required functions regarding deepfakes. Thus, Chesney and Citron judge this route as "quite limited" [106].

5.3.3 Ban

A last option might be banning deepfakes that have been created without informed consent. This is the path that China plans to pursue [126]. Also California aims for a more limited law into the same direction – here, it is becoming a criminal act to share audio or video, which promotes a false, damaging impression of an election candidate within 60 days of an election [127], [128].[43] However, there have problems been voiced regarding the enforcement ability of this law [129]. Also New York state has been aiming to go into this direction [130] and Virginia has been pushing back against AI generated non-consensual porn [131].

While banning non-consensual deepfakes might seem to elegantly separate most benign uses from malicious ones, it comes with risks, that many in the Western world will want to avoid. Importantly, it might make it easier to censor voices, that are uncomfortable or challenging views for certain high-profile officials [31]. Along similar lines it would be in tension with the right to freedom of speech and would thus face notable barriers in being passed and implemented. According to a BBC World Service survey in 2017, out of the eighteen polled countries, only the UK and China preferred governmental internet regulation [132].

5.4 Normative Solutions

In contrast to laws and regulation, norms are "an accepted standard or a way of behaving or doing things that most people agree with" [133]. As such, norms are usually not legally binding. Furthermore, norms might not even be explicitly written down anywhere. While one might fear that these characteristics render norms toothless, they can be more powerful than one might expect at first glance. Humans are particularly social animals that tend to assign high values to signal cherished characteristics – be it about their own skills, status, knowledge or commitment to a group.[44] For this reason, the often subconscious desire to signal group commitment or other valued characteristics can be a powerful force.

5.4.1 Speed and Ease of Creation

Furthermore, norms can be set up in a brief amount of time. This is a powerful advantage of norms over legally binding regulations. For instance, a handful of people could organize a workshop for experts and various other stakeholders. The goal of such a workshop might be the creation of norms and principles for one specific sub-field (e.g. for one actor-category or more general norms overall). Examples for this approach are the general AI principles like the Asilomar AI Principles of 2017 [135], the principles and norms of companies like Microsoft [136], Google [137] and IBM [138] or of countries (e.g. the OECD Principles [139]). Examples of such norms with a more restricted focus are codes of conduct for AI in the health sector (e.g. in the UK [9]) or for deepfakes specifically (for example by the WITNESS Media Lab [119] and

[43] Content, that falls within TJ, parody or satire is excluded from the law.

[44] For a thorough analysis of such signalling and hidden motives, we recommend the book *The Elephant in the Brain* as an overview [134].

by the UN's UNICRI [120]). One might imagine even a more narrow focus – for example what the creators of commercial creation tools could do to contribute to the cause.

Furthermore, before the above mentioned Network Enforcement Law in Germany has been passed, some social networks pledged voluntary obligation to take down hate speech. This can be considered the attempted starting point of a wider-reaching norm. However, this did not result in sufficient action [140] and led to the passing of the Network Enforcement Law. And we have seen a similar development on a wider scale within Europe. In 2017, the European Commission has been collaborating with some tech giants like Facebook, Twitter, YouTube and Microsoft to agree to a code of conduct in order to mitigate hate speech [141]. Here again, one might expect similar plans regarding deepfakes, or fake news more generally. Additionally, such norms might take again the role of a precursor for a legally binding regulation.

5.4.2 Range of Applicability

The range in which norms could be applied is broad and ranges from the creators of commercial editing tools to the wider public. Some have suggested that strong shame should be used as a social tool to incentivize careful sharing and to avoid sharing falsehoods [142]. However, depending on the amount of people for whom a norm should change and by how much these norms should change, it may take a long time before the respective social circle has robustly adapted. This brings us to the associated downsides of relying on norms.

Would it be needed to discourage all attackers? Or would it be sufficient to achieve robust beneficial norms among citizens, journalists, platforms and creators for example? One might argue that such changes would grant already a sturdy protection. However, for an individual the signal might be more important than the veracity of that very signal. In other words, from a short-term individual perspective, it can be possible to have the cake and to eat it, too. This can be the case if one seems as if one is acting in favor of the group (i.e. virtuously) while actually opting for the selfish option. Thus overall, it might be hard to guarantee wide-enough adoption and coherence with the norms. Furthermore, if even just one skillfully planned and executed attack may lead to extreme consequences, norms might not be the tool, that should be prioritized.

5.4.3 Summary

To summarize, norms can be a powerful tool with a wide range of possible applicabilities. On the one hand, they can be created swiftly. On the other hand, it may take long until they are robustly spread. As potentially even only very few skillfully planned deepfakes could have substantial consequences, adjusting norms can be one valuable tool, but probably not one that is to be used in absence of other ones.

6 Effectiveness Analysis

In this crucial section, we discuss how effectively the explored solution approaches might be for SM and TJ, respectively.

6.1 Choice of Focus

We focus mostly on the technical counter measures. The reasons for this focus are two-fold. Firstly, the B.Sc. program in AI has prepared us best to understand this branch of counter measure. Secondly, the technical counter measures come with a multitude of promising attributes. Technical measures can be implemented in a rather quick and uncomplicated fashion. They do not require a widespread behavioural change of societies. Rather, successful collaboration of a fewer individuals can suffice to find and implement beneficial decisions. Furthermore, technical counter measures can be prerequisite for other counter measure categories. For example legal solutions require robust solutions to identify which videos are deepfakes and which ones are not. In other words, reliable technical solutions need to be in place. Last but not least, technical counter measures may deliver tools for many different stakeholders (citizens, SM users, SM platforms, TJ outlets, courts, etc.) and technical insights can be used to inform decisions around educational counter measures.

Now let us get to the center of the effectiveness analysis – making an educated guess about the effectiveness of technical counter measures to social media and traditional journalism.

6.2 Effectiveness of Technical Solutions

We examine each of the technical counter measures in turn. Overall, please note that their effectiveness additionally depends on how much people trust in the effectiveness of these measures. Imagine for example a warning label is displayed below a shared article, but the person seeing the label does not put much certainty on its reliability. In such case, the warning label has little effect – independently of whether it was correctly there or not and independently of the method that was used in the background to gauge whether the video was a deepfake.

6.2.1 Effectiveness of Detection

The introduced detection methods have some general limitations and some that are more specific for SM or TJ, respectively. They and the benefits of detection, are examined in this section.

6.2.1.1 General Limitations

The two main general challenges of detection methods are the endless back and forth between offense and defense (i.e. the cat and mouse game) and the lack of explainability.

The Cat and Mouse Game
A major drawback of detection methods is that they depend on the still flawed aspects of the deepfake creation methods to detect whether a given video is a deepfake. As the creation methods are improving, the detection methods that once could pride themselves with high accuracy scores, become outdated and loose their effectiveness. As a consequence, the protection that detection methods provide is short-lived and new improved detection methods need to be invented continuously. In other words, the creation and detection methods are in a cat and mouse game. On the upside, is it also hard for the manipulators to find enough training data? Yes, maybe. However, if we would plan to label deepfakes online for users to see, this tells the manipulators where the current detection methods stand and hands training data to them on a silver platter. Even if we would not label them directly online, but make detection systems widely available, not only users with benign intentions may use them. Also deepfake creators are general citizens and may use these services to find out about the current strengths and weaknesses of detection systems.[45]

Many more people are investing leisure time on the side of creation than on the side of detection. According to the digital forensics expert Hany Farid, for every one person working on detection, there are 100 to 1000 people working on creation [89], [143]. Thus, this game seems to favor the side of the manipulators (i.e. the side of the creation).

Explainability and Bias
We think that another challenge and open question is how explainable these detection tools need to be. Like other applications of neural networks, also these detection tools are inherently opaque – it is hard to interpret why they classified a given input how they did (i.e. in this case, why they classified a given video input as deepfake). If we cannot comprehend why a system made a specific decision, while it is still making some errors, it is hard to pinpoint when and why errors were made. This plays into the issue of bias in AI. The quality and robustness of neural networks' outputs is also dependent on the training data. Unbalanced training data has often lead to biased decisions – from misclassifying an African American couple as "gorillas" by a neural network of Google to another AI suggesting longer sentences for dark skinned people than for light skinned ones [144], [145]. In the context of neural networks for deepfake detection, it might analogously be the case that some detection methods are performing poorly for some inputs – e.g. for videos featuring dark skinned women because they were underrepresented in the model's training data. Furthermore, for different uses and users different kinds of explanations are relevant [146]. Since these concerns first came up, the field of explainable AI has been fruitful and much progress has been made [147]–[152]. However, it is still the case, that humans have a hard time to get an intuitive grasp of the decisions of neural networks – especially for users who are non-experts.

Despite these general challenges, detection methods have one major advantage – there are contemporary versions, that can deal relatively well with contemporary deepfake creation methods. Thus, let us have a closer look at the useful detection might be for SM and TJ.

6.2.1.2 Detection in SM

The incorporation of detection methods into SM platforms could be relatively feasible. However, detection methods are not 100% error-free. Thus, an important open question for such implementations is what should best be done with the gained information.

How to Communicate the Result to the User?
Let us consider an example. If the detection model outputs a probability of 98% that a given video is a deepfake, should it simply be labelled to be a deepfake? Should the 98% be displayed for the user? Should it be less preferred by the algorithm, that determines whether the post is displayed for other users? Or should the video be taken down completely? And how does the situation change if the probability is at 70% for instance? Where would be the threshold for a binary labelling[46] or the decision to take down content?

Twitter recently held a poll among its users to gauge their preferences on such questions [153] and is continuing to gather more feedback [154]. Next to users reported preferences, it it also relevant how people tend to react to either way of labelling and content moderation. Some preliminary work indicates that some forms of labelling may indeed help reducing the effect of false information online [155], [156]. For instance, there may be unintended effects like for labels about the political stance of an article ("Democrat" and "Republican"). These labels are intended to enable a more diverse online exposure and thus mitigate the effects of filter bubbles and echo chambers.[47]. Apparently, such stance labels can make articles come across more trustworthy and less extreme, furthermore, people may use labels rather for even stronger selective exposure of their held views [160]. There is still a lot of research required to settle these questions. We face a classical exploration-exploitation trade-off [161] – For how long should we continue exploring the effects of labels before exploiting the gathered knowledge to implement labels? At the same time, potential negative side effects of labels are a point in case to consider removing deepfake content instead of adding labels. In the first days of 2020, Facebook and Reddit have both announced to ban malicious deepfakes [162], [163]. Naturally, an option that is less in tension with the right to free speech

[45] One might suggest to limit within these tools, how many images or videos a specific user or IP address can test per time unit. This could be a valuable approach, but might be circumvented by using a multitude of bots.

[46] I.e. either being labelled as "real" or "fake"

[47] Echo chambers roughly being caused by the user choosing an exposure to news, that they already agrees with and filter bubbles being the SM algorithm doing their part in this information segregation – selectively presenting content, that is compatible with the users views [157]–[159]

would be to not recommend classified deepfakes. This would substantially throttle the speed at which misleading deepfakes could be expected to spread.

Privacy versus Safety?
In some regions, the prevailing form of SM for news are services like WhatsApp. Such messaging services may tend to be regarded as more private than posting on ones Facebook timeline or uploading a video on YouTube for instance. Consequently, users might be in opposition to their messages being scanned – also when it is in the name of protecting them and the community from disinformation.[48]

In China, such pervasive censoring is already happening, for example in the social media app called WeChat. WeChat is a "mega app" that combines a staggering multitude of possibilities. By analogy, one might say that it incorporates the functionalities of what in the West is thought of as social media (from posting more publicly to private messaging). Furthermore, it enables all kinds of payments and bookings within China, but these functionalities are naturally not in the focus of this thesis. What we are interested in, is the (real-time) censoring of private messaging features. It is hard to find reliable information on how this is perceived within China. There are individuals who leave China in order to communicate freely and they will still be censored within WeChat if they originally signed up with a Chinese phone number [164]. Evidently, these individuals do not appreciate this, but we could not find data on how the distribution of opinions is within China.

Even though, there are two different versions of the app – WeChat for the West and 微信 (Wēixìn) for China – also WeChat users are being censored [164], [165]. This might be the only way that people in the West are familiar with having private messages censored is by using Chinese apps like WeChat.[49] If it turns out, that users oppose such a scanning of their messages for detecting deepfakes, this could be another hurdle for relying on detection as a solution within SM.

6.2.1.3 Detection in TJ

Detection applications for TJ face many of the same challenges. Bias and lack of explainability are an issue, as well as the brittleness or short-lived effectiveness of detection, that comes with the cat and mouse dynamic between detection and creation. However, it seems like TJ is in a better position to face these challenges. Comparing the number of authors within SM and TJ – TJ's number is only a small fraction of the number of SM content creators. This reduces the needed costs of educating authors about the limited explainability and the potential for bias. Furthermore, TJ authors have large direct incentives of only publishing truthful content. Consequently, even commercial detection tools for journalists may be a viable option. Commercial tools will likely do their best to be up to date and on top the cat and mouse game, in order to be competitive. In contrast, the sharing incentives for SM users seem often to be more about group identity signalling than about sharing correct information [167][50] and the vast majority of citizen seems to prefer spending limited resources on entertainment subscriptions rather than on news-related ones.[51]

To conclude, the issue of the cat and mouse dynamic is a major drawback for detection methods. Because SM users tend to have lower incentives and possibilities to invest resources, the issue of lacking explainability is exacerbated within SM. Furthermore, there is still significant uncertainty around the exact effects of labeling content online. Lastly, SM potentially needs to face a tension between privacy and detection within more private SM environments (e.g. within WhatsApp, Telegram or Facebook messenger).

6.2.2 Effectiveness of Authentication

Let us now have a look at how applicable authentication might be within social media and traditional journalism.

6.2.2.1 Authentication in TJ

Authentication offers low-hanging fruits for TJ. As mentioned, there are existing authentication apps, with which TJ outlets can equip their staff (including free-lancers) with. Like that, their own raw material would be verified. Consequently, TJ would not need to fear incidents like one mentioned above in figure 4 – manipulated photos from freelancers of reputable journals.

6.2.2.2 Authentication in SM

For SM, the situation currently looks a bit more challenging. We will look at three possible authentication uses for SM that we came up with, (1) SM users use authentication app independently, (2) future integration of an authentication service into the SM platform, and (3) authentication integrated in mobile phone operating systems (OS). Let us have a look at these options in turn.

(1) SM Users Use Authentication App Independently
Authentication apps for taking verified pictures and videos are already available. SM users can already make use of them to take and share verified material. Thus, could a well-planned marketing and promotion of these apps solve a major part of visual disinformation? It could for example be an option to tweak the SM algorithms, such that

[48] It is to note, that these challenges are ideas from our side. As we have not been able to find data from surveys or experiments on this, it is not certain that these challenges would need to be faced.

[49] Keep in mind that this is not the case for all Chinese apps, that have an international version. TikTok for example – the international version of the Chinese video sharing app 抖音 Dǒuyīn – is mostly uncensored outside of China. [166]

[50] or more generally [134]

[51] Over 13000 citizens (age < 45 years) of 14 countries were asked on what kind of subscription, they would settle for the next year if they could only have one. Results: 37% chose online video subscriptions, 15% chose online music subscriptions and only 7% chose online news subscriptions [75].

content with verified visual material is preferred over content with unverified material. If people know about that, the users themselves have a stronger incentive to use authentication apps for taking their visual material. However, there is one crucial issue left so solve. People do not want to upload raw pictures. The majority of uploaded visual content is edited in some form or another. It seems irresponsible to rely on a normative change of everyone preferring raw material by having something like a verified *#nofilter*. A combination with provenance possibilities might be able to help. In a nutshell, the capture of visual material could be verified via authentication and edits like changing the contrast or brightness could be tracked via provenance. If the capture is verified and all edits are innocuous, the post could receive the mentioned boost within the social media algorithms. However, this incentive tweaking is not applicable for private SM, such as WhatsApp or Facebook messenger.

(2) Future Integration of Authentication Service Into the SM Platform
A future integration of an authentication service into the SM platform has the advantage of making the verified capture the (unavoidable) default. Consequently, the success of this approach does not depend on the behavioral change of billions of users. Because people will still want to edit their photos and videos, also here a combination with provenance methods seems necessary.

(3) Authentication Integrated in Mobile Phone Operating Systems
If authentication (and provenance) would be integrated in the operating systems of the mobile phones, by default only verified images and videos would be taken. Thus, there is no additional burden created for the user. Such a system could not only benefit the information ecosystem on SM, but also be useful to keep alive the value of photos and videos for court use. Furthermore, if only the operating system of Android mobile phones and iPhones would tackle this undertaking together, the vast majority of mobile cameras would be covered. The captured visual material could then also be shared on SM, where only a last check would need to take place. An additional hurdle for this direction is that, in contrast to SM platforms, mobile phone producers are currently not seen as, and do not seem to feel, responsible for resolving issues around the misuse of deepfakes.

The main disadvantage of using authentication for SM is that a complex system (combining authentication with provenance) would need to be implemented and integrated by multiple parties – be it the users themselves, the SM platforms or the mobile phone producers. Even if everyone would be on board, carrying out such a plan would likely take a long time. It could be interesting to explore what role legal regulations could play to make this path more feasible.

6.2.3 Effectiveness of Provenance

Let us now have a look at how applicable provenance might be within social media and traditional journalism.

6.2.3.1 Provenance in TJ

It is long established, that it is a responsibility of TJ to trace back the path and origins of the used evidence. Contemporary provenance tools are not much more than an updated instrument in journalists' toolboxes. Furthermore, due to journalists' expertise and exercise to judge the roots of a given piece of evidence, provenance could also be used without integration with authentication. The main difficulty might be to communicate the uncertainty and low explainability that comes with opaque neural network tools. Thus, journalists might need some training in how to use and make sense of the new tools, but this should be only a relatively small obstacle.

6.2.3.2 Provenance in SM

The applications of provenance for SM are clearly more troublesome. The most realistic application, that we found is to integrate provenance methods into SM platforms – for the provenance of edits before sharing it for the first time online and/or for tracking how the video was shared online. The former is similar to the already described in above in the SM applications of authentication. The latter would likely use similarity search as well as timestamps and geo-tags. Furthermore, the online tracking would result in a provenance graph (see figure 18). For the average SM user, this graph by itself would likely be of little use – the interpretation is not straight-forward. However, it could be possible to either accompany or replace the provenance graph by an explanation. A possible explanation might be, "This video was originally published in a different context. It might have been manipulated to communicate a misleading message. Check out this link to see its original context." with a link to the previous upload. This approach could be useful for face replacement deepfakes and lip-syncing deepfakes. Regarding the provenance before the first upload, however, this approach is likely to fail for media that are entirely fabricated – like current face generation and voice synthesis. In a few years, it might be possible and not too much effort to create entirely fabricated deepfakes. Such deepfakes would not directly re-use the the background and body from a conventional video. Rather, the full video would be created from scratch. Similar to the current possibilities for photos in the realm of face-generation. The online-tracking provenance could still be useful to discover invalid re-contextualizations – a simple form of false information. The approach could therefore address a shared root cause of disinformation. However, it would not actively targeting deepfakes. This would render it rather ineffective regarding more advanced forms of false information such as deepfakes.

6.2.3.3 Provenance Overall

In conclusion, provenance is a welcomed tool for TJ and is not expected to lead to many complications here. For SM and for TJ, the provenance before the first upload will inevitably lose its power when videos can be entirely generated. At the same time, we do not foresee the more general online-tracking provenance to lose its effectiveness.

For SM, a provenance graph and/or an explanation for the user needs to be provided – this is likely more difficult than a label for detection because only relevant roots should be presented. For the time being, provenance by itself could be a useful deepfake counter measure for SM. However, if a high-quality and intuitive SM integration would require too much time, complete fabrication of deepfakes might already be possible. This would render provenance solely useful for re-contextualizations of videos. The previously mentioned combination with authentication seems more promising (see **??**).

One frequently voiced disadvantage is that provenance and authentication could be a trade-off with anonymity [83]. If a given video could be traced back to a specific camera, this could risk the anonymity of the person who provided the video. However, even if a video is traced back to one unique camera, there is no accessible registration of who bought which camera. Thus, it does not imply who owns the camera and who captured the video.

6.2.4 Effectiveness of Anonymization

Adding perturbation filters onto your images achieves partial anonymization and could thus prevent your pictures from easily being traced back to you – at least by current facial recognition systems.[52] Unfortunately, there are at least two flaws, which render the effectiveness of this approach limited. The first flaw is that attackers can still find photos of you around when your name is mentioned (think profile pictures, Instagram accounts, other social media presences or articles about you). The second flaw is that researchers, companies and 'hobbyists' the like are aiming to arrive at creation techniques, which only require one or very few pictures to deliver a realistic result. As a consequence, progress in this direction is being made (see for example the Zao app, in which one (profile) photo of a user is enough to create 30 seconds of fairly realistic deepfake video. You can watch it in [169].). If only one or a handful of images are needed from you, attackers will not have too high costs of gathering manually.[53]

In conclusion, the effort of manually searching for many photos of one individual can be a hurdle, that creators might want to automate by using facial recognition systems to scrape the internet for photos of that person. If partial anonymization is applied to most online photos of the target, the creators are forced to manually search and download the required pictures. This may be somewhat of a protection. However, if deepfakes can be created with only very few images of the target, partial anonymization loses its effectiveness of preventing harmful deepfakes. Complete anonymization can be used for beneficial deepfake applications and it might mitigate the Streisand effect for deepfake victims. However, we could not find uses of complete anonymization that would prevent malicious uses of deepfake technology. These points hold for both, SM and TJ.

[52] One partial anonymization tool is already available and offered by equalAIs [168].

[53] It is unclear to us whether an adversarial attack filter on your pictures could impede creating deepfakes.

6.3 Effectiveness of Educational Solutions

Educational solutions aim to empower people. The goal is to equip individuals with the tools to independently judge the veracity of videos. Let us now have a look at how this approach applies for TJ and SM.

6.3.1 TJ Perspective

Traditional Journalism benefits once more from their limited amount of content creators and the high resources and incentives to only publish true content. These characteristics make it feasible to provide workshops and dedicated resources for a bigger percentage of authors than it would be possible for SM. Furthermore, the situation of the incentive make it even possible that these services can have commercial value.

6.3.2 SM Perspective

The complementary attributes of SM – vast amounts of content creator, lower resources and incentives for sharing verified content – may shatter the potential of educational approaches for SM. However, before declaring this subfield as neglectable, let us have a look at the requirements, that different possibilities would have. We distinguish between solution ideas, whose target group is citizens in their leisure, and solution ideas, that aim to use the resources of non-leisure time to increase media literacy levels.

6.3.2.1 Within Leisure

In order to motivate people to educate themselves in their precious leisure time, the experience likely needs to be entertaining and available for free. Furthermore, it should probably be the default to encounter the material, rather than requiring effort and initiative of each individual to unearth it.

Entertaining

In their leisure, people prefer to relax and thus can be expected focus more on low-effort, entertaining content than on education. Especially if the education is around something that does not pay off directly (like increased likelihood to get promoted), but rather is something that promises a gain in judgement without direct feedback of whether one is actually progressing. There might be some people, who place a lot of value on their media literacy and consequently are open to spend some of their leisure with more conventional education on the topic. However,

also given the current situation around mis- and disinformation[54], it seems naive to rely on the hope that most people will invest their leisure to do something that feels like studying. Consequently, the education solutions that people should engage with in their leisure should aim for packaging the relevant information in an entertaining fashion, that is low-effort in its consumption.

Free
If most people should be able to access such materials, they should not come with a financial burden, but be easily available (i.e. low-effort to find) and for free. This is already the case for entertainment material, for which people have a higher (innate) incentive to access them. Consequently, this should even more be the case for educational material.

Long- versus Short-Lived
Whether the focus of within-leisure education needs to strictly be long-lived advice or whether also short-lived advice may be useful, depends on the kind of within-leisure education. If individuals will likely not revisit the educational material frequently, a strict focus on long-lived advice seems appropriate.

Example
To have a situation in mind, in which this decision might not be so clear, let us consider the following example. Let us imagine a Facebook feed with a post that has a label. The label marks it as potentially containing disinformation. A promising idea could be to embed the educational material precisely where people tend to encounter potential mis- and disinformation. If that label comes with the educational information, the user would not need to actively try to find relevant information. The educational information could for example come in the form of a link or by being displayed when one is hovering over the label. Like that, all identified relevant aspects would be combined.[55]

For this example, the users would steadily and repeatedly encounter the educational material. If users would periodically engage with the material, it may be the case that also short-lived advice might have a net benefit. However, because repeated engagement can not be guaranteed, a focus on longer-lived advice seems to be a more responsible starting point. Another crucial open question is whether the label itself might nudge people away from educating themselves. It may be the case that people think that the fact-checking task has already been done by an algorithm and consequently there is little use for them to do anything on top of that. If that would be the case, we would want to find ways to avoid such a counter-productive nudge.

6.3.2.2 Within Education System

Surprisingly, the restrictions that we encountered for leisure time education also seem to hold for education efforts within the educational system.

Entertaining and Free
If one aims equip students for their online activities, it can make sense to relax the restrictions that SM is generally facing by utilizing some of the time and attention resources that fall into their hours at school. However, in order to spark engagement with and their interest for the topic, it could still be a good choice to aim for entertaining content. Furthermore, teachers are often already hard pressed on time and budget[172]. In order to make it more feasible for teachers to shift the needed resources towards increasing their pupils' media literacy, it seems wise to make the material available for free. A current example of this approach is the before mentioned *Interland* game by Google.

Long- versus Short-Lived
Because such special projects at school are likely to reach each student not more than once, it seems crucial for interventions, to focus on long-lived advice and tools. It may be useful to also include examples of short-lived advice, that worked in the past, while emphasizing strongly that such measures are loosing their effectiveness quickly and thus should not be relied on (which itself can be seen again as a long-lived advice).

In general, we should keep in mind the viscerality of deepfakes. The best approaches might include ways to circumvent the situation of viewing deepfakes altogether. The US House Intelligence Committee Chairman Adam Schiff said "[Psychologists] will tell you that even if you're later persuaded that the video you have watched is a forgery, you will never completely shed the lingering negative impression of the person" [89]. Furthermore he added that your "brain will tell you, 'I shouldn't hold it against Joe Biden or Donald Trump or Bernie Sanders or Elizabeth Warren because that video I saw that went viral, I now know to have been a fake. But I cannot shake the feeling that that person is,' you know, fill in the blank. So part of the damage is done once you see it or you hear it" [89]. Even if we can identify and authenticate deepfakes and inform the people, there might be such effects that will haunt the viewers' minds.

7 Discussion

Let us now summarize what we have found and what our findings imply for the future decisions around counter measures for deepfakes.

[54] Mis- and Disinformation differ by whether intend to harm is involved. For disinformation, intend to harm is involved while misinformation includes honest mistakes. [170], [171]

[55] Naturally, we cannot guarantee the effectiveness of such an implementation because the effect of such information, as well as the user experience would need to be tested. Furthermore, it is currently unclear whether only one tip should be seen as a preview, for example in the style of "Did you know... you can(/not) spot fake material by... [one option]". This could take a similar form as the overview of the link content, that Wikipedia uses when hovering over a the link to another Wikipedia page.

7.1 Summary

Within this project, we have looked into whether current (technical) counter measures for deepfakes are generally less effective in the SM environment than in the TJ environment. To that end, we have glanced through earlier forgery methods to compare them with the possibilities and limitations of current deepfake categories. The exploration of various deepfake counter measures has enabled our analysis of the counter measure's applicability and effectiveness within the two media environments (social media and traditional journalism).

7.2 Limitations

A main limitation of our work is the assumption that the technical counter measures is the central subfield of approaches. We needed to make this assumption to not burst the scope of a Bachelor Thesis project, but it may very well be the case that more free and creative work can lead to more powerful counter measures. Due to the same reason, we had to leave intricate details within the counter measures underexplored. As a consequence, our action recommendations might turn out to be not feasible or we might have missed a better solution approach.

Another constraint is that there is very little research done around educational solutions. As a result, our recommendations in this field still include many open questions to be answered in future research.

7.3 Conclusion

From our analysis, it follows that the current technical (and educational) counter measures are indeed more effective in the context of traditional journalism. Let us now look at the gist of our takeaways from this project in terms of technical counter measures.

7.3.1 Detection as a Patch – a quick solution for now.

Detection has the major advantage of being more or less ready to be implemented and launched. However, because of the cat and mouse dynamic between creation and detection, this approach tends to be short-lived and brittle. Furthermore, detection introduces other complications as lacking explainability and bias. Consequently, we would recommend to not rely on detection as the main solution for the longer term.

7.3.2 Invest Now Into Implementations and Research Around Authentication and Provenance – For a Robust Longer-Term Solution.

Authentication seems promising thanks to its great robustness to developing capabilities behind deepfakes. The combination with provenance seems especially favourable. We arrived at three different actor categories that could be valuable for such solutions – SM users, SM platforms and providers of mobile phone operating systems. The most promising one of them might be the integration into major mobile phone operating systems. A fruitful integration could mean that the vast majority of captured photos and videos would be verified. Collaboration between OS providers and external apps for editing could enable a safe solution without unnecessarily restricting harmless editing.

Moreover, it is to mention that of the four main categories of technical counter measures (i.e. Detection, Authentication, Provenance and Anonymization) Anonymization seems to promise the least use to prevent malicious deepfakes.

7.4 Significance

Implementing reliable solutions as soon as possible is important for our modern societies. Whether our motivation stems from protecting individual lives form major turbulences from involuntary porn, whether we want to mitigate the consequences of a whole new layer of cybercrime, or whether we aim to protect the democratic process, we believe that it is important to keep in mind how such technology of developed nations is influencing countries like India, Myanmar and Sri Lanka. Here, many risk factors come together – mis- and disinformation is common [173], local languages have smaller numbers of speakers[56], and the a large part of the population relies mostly on social media to access news, while many people are starting to use the internet for the first time and do so via social media [175]. These aspects together form a volatile mix that can lead to countless deaths and enourmous suffering.

7.5 Future Research

An important point for future research is the realm around the details of authentication and provenance approaches. There are many more intricate details, which lie beyond the scope of this project. Many of them are important to consider if one aims to arrive at a market-ready implementation. One such aspect is which exact form of authentication should be used. Blockchain technology for instance, is a burden for climate change [176]. A major reason for aiming to arrive soon at deepfake counter measures is to serve developing countries. However, many of the least developed countries are expected to feel the negative consequences of climate change the most [177]. Thus, we want to be especially careful that our solution approaches are not exacerbating the tension and burden in these regions.

A crucial opportunity for future research is to find out, via which of these actor-based routes we should pursue to arrive early at an implementation, that is widely adopted. Such an implementations likely will require the integration of provenance within it in order to let users edit their

[56] this makes it harder to moderate the content because enough people who speak these languages need to be employed to take care of "cleaning" the pool online content and checking reported content [174]

photos and videos. We believe, that if such a solution could be launched within the next years, our societies would have gained a robust longer-term solution.

References

[1] H. Farid, "Digital forensics in a post-truth age", *Forensic science international*, vol. 289, pp. 268–269, 2018.

[2] H. K. Hall, "Deepfake videos: When seeing isn't believing", *Cath. UJL & Tech*, vol. 27, p. 51, 2018.

[3] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics", *Foreign Aff.*, vol. 98, p. 147, 2019.

[4] J. Fletcher, "Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance", *Theatre Journal*, vol. 70, no. 4, pp. 455–471, 2018.

[5] H. Farid, "Image forgery detection–a survey", 2009.

[6] *Lexico "social media"*, [accessed 11 Dec 2019], 2019. [Online]. Available: https://www.lexico.com/en/definition/social_media.

[7] *Traditional vs. nontraditional journalism "social media"*, [accessed 11 Dec 2019], 2019. [Online]. Available: https://msuzukij2618.wordpress.com/traditional-vs-nontraditional-journalism/.

[8] "Millennials and political news", *Pew research center*, vol. 1, 2015.

[9] J. Research, *News consumption in the uk*, [accessed 10 Sep 2019], 2019. [Online]. Available: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption (visited on 07/24/2019).

[10] L. Valla, *The Treatise of Lorenzo Valla on the Donation of Constantine: Text and*. Yale University Press, 1922.

[11] G. Clarke, *The photograph*. Oxford University Press, USA, 1997.

[12] A. C. Piccinnano, "Techniques for ditial image forensics and counter-forensics", PhD thesis, University of Siena, Feb. 2014.

[13] N. H. Department, *A very weird photo of ulysses s. grant*, [accessed 23 Dec 2019], 2015. [Online]. Available: https://www.npr.org/sections/npr-history-dept/2015/10/27/452089384/a-very-weird-photo-of-ulysses-s-grant?t=1577091526456.

[14] O. Yegorov, *How stalin's propaganda machine made people vanish from pictures*, [accessed 18 Sep 2019], 2018. [Online]. Available: https://www.rbth.com/history/329317-stalin-propaganda-photos (visited on 10/15/2018).

[15] E. Blakemore, *How photos became a weapon in stalin's great purge*, [accessed 16 Sep 2019], 2018. [Online]. Available: https://www.history.com/news/josef-stalin-great-purge-photo-retouching (visited on 04/20/2018).

[16] R. G. Nulph, *Edit suite: Once upon a time: The history of videotape editing*, [accessed 19 Sep 2019], 1997. [Online]. Available: https://www.videomaker.com/article/2896-edit-suite-once-upon-a-time-the-history-of-videotape-editing (visited on 07/01/1997).

[17] S. Pagin, *The evolution of photoshop: 25 years in the making*, [accessed 30 Sep 2019]. [Online]. Available: https://www.fastprint.co.uk/blog/the-evolution-of-photoshop-25-years-in-the-making.html.

[18] E. P̃. staff, *Update: Reuters fires photog over doctored pictures*, [accessed 9 Oct 2019], 2006. [Online]. Available: https://web.archive.org/web/20060908091324/http://editorandpublisher.com/eandp/news/article_display.jsp?vnu_content_id=1002950988&imw=Y.

[19] H. Farid, "Digital doctoring: How to tell the real from the fake", *Significance*, vol. 3, no. 4, pp. 162–166, 2006.

[20] PHLEARN, *How to use content aware fill in photoshop*, [accessed 11 Dec 2019], 2016. [Online]. Available: https://www.youtube.com/watch?v=Ge9jsJZ3lA0&feature=youtu.be.

[21] M. A. Kunz, *Cloak: Remove unwanted objects in video*, [accessed 28 Sep 2019], 2017. [Online]. Available: https://research.adobe.com/news/cloak-remove-unwanted-objects-in-video/.

[22] P. T. Channel, *How to use the face-aware liquify in photoshop*, [accessed 1 Oct 2019], 2016. [Online]. Available: https://www.youtube.com/watch?v=2zhgvNfJTnM.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *This person does not exist*, [accessed 2 Oct 2019], 2019. [Online]. Available: https://thispersondoesnotexist.com/.

[25] S. Cole, *We are truly fucked: Everyone is making ai-generated fake porn now*, [accessed 6 Oct 2019], 2018. [Online]. Available: https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

[26] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting photoshopped faces by scripting photoshop", *arXiv preprint arXiv:1906.05856*, 2019.

[27] Adobe, *Adobe photoshop pricing*, [accessed 5 Oct 2019], 2019. [Online]. Available: https://www.adobe.com/products/photoshop/pricing-info.html.

[28] T. @theInpaint, *Remove people from photo: The easy way*, [accessed 9 Oct 2019], 2019. [Online]. Available: https://www.theinpaint.com/inpaint-how-to-remove-unwanted-people-from-photo.html.

[29] B. Barron, *How to remove a person from a photo on your iphone*, [accessed 7 Oct 2019], 2019. [Online]. Available: https://enviragallery.com/remove-person-from-photo-iphone/.

[30] *Cambridge dictionary*, [accessed 3 Oct 2019], 2019. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/visceral.

[31] D. K. Citron, "Prepared written testimony and statement for the record", *U.S. House of Representatives Committee Repository*, 2019. [Online]. Available: https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=109620.

[32] D. Harwell, *Faked pelosi videos, slowed to make her appear drunk, spread across social media*, [accessed 11 Oct 2019], 2019. [Online]. Available: https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/.

[33] B. Feldman, *Why we love snapchat's face swaps*, [accessed 13 Nov 2019], 2016. [Online]. Available: https://slate.com/culture/2016/03/why-we-love-snapchat-s-face-swaps.html.

[34] *R/deepcage*, [accessed 21 Dec 2019]. [Online]. Available: https://www.reddit.com/r/deepcage/.

[35] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets", *arXiv preprint arXiv:1911.08139*, 2019.

[36] Wikipedia, *An example of deepfake technology: Actress amy adams in the original (left) is modified to have the face of actor nicolas cage (right)*, [accessed 23 Dec 2019], 2017. [Online]. Available: https://en.wikipedia.org/wiki/Deepfake#/media/File:Deepfake_example.gif.

[37] H. Kim, P. Carrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits", *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 163, 2018.

[38] R. University, *Deze filosoof bij ai pleit voor een robot 'rijbewijs'*, [accessed 25 Dec 2019], 2018. [Online]. Available: https://www.ru.nl/nieuws-agenda/vm/2018/januari/filosoof-ai-pleit-robot-rijbewijs/.

[39] ——, *Lonely at christmas? 'a robot is better than nothing.'* [accessed 25 Dec 2019], 2018. [Online]. Available: https://www.ru.nl/@1184856/lonely-christmas-robot-better-than-nothing/.

[40] T. B. Lee, *I created my own deepfake—it took two weeks and cost $552*, [accessed 23 Dec 2019], 2019. [Online]. Available: https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/.

[41] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[42] M. Zeymour, *Canny ai: Imagine world leaders singing*, [accessed 21 Dec 2019], 2019. [Online]. Available: https://www.fxguide.com/fxfeatured/canny-ai-imagine-world-leaders-singing/.

[43] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "Headon: Real-time reenactment of human portrait videos", *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 164, 2018.

[44] P. Esser, J. Haux, T. Milbich, and B. orn Ommer, "Towards learning a realistic rendering of human behavior", in *The European Conference on Computer Vision (ECCV) Workshops*, Sep. 2018.

[45] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis", in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5904–5913.

[46] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now", in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.

[47] C. for Data Ethics and Innovation, *Snapshot paper - deepfakes and audiovisual disinformation (independent report)*, [accessed 21 Dec 2019], 2019. [Online]. Available: https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation#about-this-cdei-snapshot-paper.

[48] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio", *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[49] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: One-shot anatomically consistent facial animation", *International Journal of Computer Vision*, pp. 1–16, 2019.

[50] S. Qian, K.-Y. Lin, W. Wu, Y. Liu, Q. Wang, F. Shen, C. Qian, and R. He, "Make a face: Towards arbitrary high fidelity face manipulation", in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 033–10 042.

[51] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation", *arXiv preprint arXiv:1802.07228*, 2018.

[52] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models", *arXiv preprint arXiv:1905.08233*, 2019.

[53] E. Zakharov, *Few-shot adversarial learning of realistic neural talking head models*, [accessed 23 Dec 2019], 2019. [Online]. Available: `https://www.youtube.com/watch?v=p1b5aiTrGzY`.

[54] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[55] L. AI, *Lyrebird ai*, [accessed 23 Dec 2019], 2017. [Online]. Available: `https://www.descript.com/lyrebird-ai?source=lyrebird`.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[57] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[58] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio", *arXiv preprint arXiv:1609.03499*, 2016.

[59] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, IEEE, vol. 1, 1996, pp. 373–376.

[60] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis", in *Advances in Neural Information Processing Systems*, 2019, pp. 14 881–14 892.

[61] Google, *The discriminator*, [accessed 21 Dec 2019]. [Online]. Available: `https://developers.google.com/machine-learning/gan/discriminator`.

[62] ——, *The generator*, [accessed 21 Dec 2019]. [Online]. Available: `https://developers.google.com/machine-learning/gan/generator`.

[63] A. M. Kaplan, "If you love something, let it go mobile: Mobile marketing and mobile social media 4x4", *Business horizons*, vol. 55, no. 2, pp. 129–139, 2012.

[64] Facebook, *Facebook places*, [accessed 19 Dec 2019], 2019. [Online]. Available: `https://www.facebook.com/places/`.

[65] Foursquare, *Foursquare*, [accessed 19 Dec 2019], 2019. [Online]. Available: `https://foursquare.com/`.

[66] Yelp, *Yelp*, [accessed 19 Dec 2019], 2019. [Online]. Available: `https://www.yelp.com/`.

[67] Twitter, *Twitter*, [accessed 19 Dec 2019], 2019. [Online]. Available: `http://twitter.com/`.

[68] Youtube, *Youtube*, [accessed 19 Dec 2019], 2019. [Online]. Available: `http://youtube.com/`.

[69] Wikipedia, *Wikipedia*, [accessed 19 Dec 2019], 2019. [Online]. Available: `https://www.wikipedia.org/`.

[70] M. Christensen, *What are the requirements necessary to become a journalist?*, [accessed 21 Dec 2019], 2019. [Online]. Available: `https://work.chron.com/requirements-necessary-become-journalist-12514.html`.

[71] BBC, *Guidelines*, [Accessed 9 Dec 2019]. [Online]. Available: `https://www.bbc.com/editorialguidelines/guidelines`.

[72] Kurzgesagt, *Kurzgesagt – in a nutshell*, [accessed 21 Dec 2019], 2019. [Online]. Available: `https://www.youtube.com/user/Kurzgesagt`.

[73] Veritasium, *Veritasium*, [accessed 21 Dec 2019], 2019. [Online]. Available: `https://www.youtube.com/user/1veritasium`.

[74] K. Quach, *Youtube thinkfluencer siraj raval admits he plagiarized boffins' neural qubit papers – as esa axes his workshop*, [accessed 21 Dec 2019], 2019. [Online]. Available: https://www.theregister.co.uk/2019/10/14/ravel_ai_youtube/.

[75] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. Nielsen, *Reuters institute digital news report 2019*. Reuters Institute for the Study of Journalism, 2019, vol. 2019.

[76] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on twitter?", *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 179–192, 2016.

[77] B. C. Andrew, "Media-generated shortcuts: Do newspaper headlines present another roadblock for low-information rationality?", *Harvard International Journal of Press/Politics*, vol. 12, no. 2, pp. 24–43, 2007.

[78] R. Dawkins, *The selfish gene*, 1976.

[79] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online", *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[80] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads", *PloS one*, vol. 11, no. 3, e0150989, 2016.

[81] N. Newman, "The rise of social media and its impact on mainstream journalism", 2009.

[82] B. Franklin and L. Canter, *Digital Journalism Studies: The Key Concepts*. Routledge, 2019.

[83] S. Gregory, *Deepfakes and synthetic media: Updated survey of solutions against malicious usages*, [Accessed 29 Nov 2019], 2019. [Online]. Available: https://blog.witness.org/2019/06/deepfakes-synthetic-media-updated-survey-solutions-malicious-usages/.

[84] D. Dr. Matt Turek, *Media forensics (medifor)*, [Accessed 30 Nov 2019], 2019. [Online]. Available: https://www.darpa.mil/program/media-forensics.

[85] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking", *arXiv preprint arXiv:1806.02877*, 2018.

[86] F. Marconi and T. Daldrup, *How the wall street journal is preparing its journalists to detect deepfakes*, [accessed 23 Nov 2019], 2018. [Online]. Available: https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/.

[87] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ode", in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[88] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[89] O. Beavers, *Experts are studying mannerisms of 2020 candidates to help offset threat of 'deepfake' videos*, [Accessed 30 Nov 2019], 2019. [Online]. Available: https://thehill.com/policy/cybersecurity/443018-experts-are-studying-mannerisms-of-2020-candidates-to-help-offset-threat.

[90] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 104–109.

[91] G. A. Daisy Stanton, *Advancing research on fake audio detection*, [Accessed 30 Nov 2019], 2019. [Online]. Available: https://blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/.

[92] G. N. Initiative, *Building a stronger future for journalism*, [accessed 19 Nov 2019], 2019. [Online]. Available: https://newsinitiative.withgoogle.com/.

[93] D. N. I. Fund, *Digital news innovation fund*, [accessed 19 Nov 2019], 2019. [Online]. Available: https://newsinitiative.withgoogle.com/dnifund/.

[94] ——, *Digger - deepfake detection*, [accessed 19 Nov 2019], 2019. [Online]. Available: https://tinyurl.com/wp2hje9.

[95] J. Lukás, J. Fridrich, and M. Goljan, "Digital" bullet scratches" for images", in *IEEE International Conference on Image Processing 2005*, IEEE, vol. 3, 2005, pp. III–65.

[96] P. Korus and N. Memon, "Neural imaging pipelines-the scourge or hope of forensics?", 2019.

[97] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?", in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 2019, pp. 506–511.

[98] A. E. Dirik and A. Karaküçük, "Forensic use of photo response non-uniformity of imaging sensors and a counter method", *Optics express*, vol. 22, no. 1, pp. 470–482, 2014.

[99] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva, "Spoc: Spoofing camera fingerprints", *arXiv preprint arXiv:1911.12069*, 2019.

[100] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces", *arXiv preprint arXiv:1803.09179*, 2018.

[101] F. A. Mike Schroepfer Chief Technology Officer, *Creating a data set and a challenge for deepfakes*, [Accessed 30 Nov 2019], 2019. [Online]. Available: https://ai.facebook.com/blog/deepfake-detection-challenge/.

[102] C. Dictionary, *Authentication*, [Accessed 30 Nov2019]. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/authentication.

[103] *Ewitness: Seeing can be believing again*, [Accessed 30 Nov 2019]. [Online]. Available: https://ewitness.commons.gc.cuny.edu/.

[104] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends", in *2017 IEEE International Congress on Big Data (BigData Congress)*, IEEE, 2017, pp. 557–564.

[105] J. Frankenfield, *Permissioned blockchains*, [Accessed 3 Dec 2019], 2019. [Online]. Available: https://www.investopedia.com/terms/p/permissioned-blockchains.asp.

[106] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security", 2018.

[107] D. Moreira, A. Bharati, J. Brogan, A. Pinto, M. Parowski, K. W. Bowyer, P. J. Flynn, A. Rocha, and W. J. Scheirer, "Image provenance analysis at scale", *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6109–6123, 2018.

[108] I. Archive, *Internet archive*, [accessed 7 Nov 2019], 2019. [Online]. Available: https://archive.org/.

[109] A. Bharati, D. Moreira, J. Brogan, P. Hale, K. Bowyer, P. Flynn, A. Rocha, and W. Scheirer, "Beyond pixels: Image provenance analysis leveraging metadata", in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2019, pp. 1692–1702.

[110] InVid, *Invid*, [accessed 5 Nov 2019], 2019. [Online]. Available: https://www.invid-project.eu/.

[111] ——, *Invid plugin*, [accessed 5 Nov 2019], 2019. [Online]. Available: https://www.invid-project.eu/tools-and-services/invid-verification-plugin/.

[112] I. Goodfellow, N. Papernot, S. Huang, R. Duan, P. Abbeel, and J. Clark, *Attacking machine learning with adversarial examples*, [Accessed 3 Dec 2019], 2017. [Online]. Available: https://openai.com/blog/adversarial-example-research/.

[113] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization", in *International Symposium on Visual Computing*, Springer, 2019, pp. 565–578.

[114] C. Silverman, *How to spot a deepfake like the barack obama–jordan peele video*, [accessed 5 Nov 2019], 2018. [Online]. Available: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed.

[115] T. W. Post, *Seeing isn't believing. the fact checker's guide to manipulated video*, [accessed 5 Nov 2019], 2019. [Online]. Available: https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/.

[116] P. on Computational Propaganda, *The comprop navigator*, [accessed 11 Jan 2020], 2019. [Online]. Available: https://navigator.oii.ox.ac.uk/.

[117] Google, *Helping kids be safe, confident explorers of the online world.* [accessed 23 Nov 2019], 2019. [Online]. Available: https://beinternetawesome.withgoogle.com/en_us.

[118] H. Baker, *Introducing the reuters guide to manipulated media, in association with the facebook journalism project*, [accessed 11 Jan 2020], 2019. [Online]. Available: https://www.reuters.com/article/rpb-hazeldeepfakesblog/introducing-the-reuters-guide-to-manipulated-media-in-association-with-the-facebook-journalism-project-idUSKBN1YY14C.

[119] W. M. Lab, *Deepfakes: Prepare now (perspectives from brazil)*, [accessed 23 Nov 2019], 2019. [Online]. Available: https://lab.witness.org/brazil-deepfakes-prepare-now/.

[120] K. Dodgson, *Dsi and unicri workshop on deep fakes and video manipulations*, [accessed 10 Dec 2019], 2019. [Online]. Available: https://datascienceinitiative.eu/projects/dsi-and-unicri-workshop-on-deep-fakes-and-video-manipulations/.

[121] UNICRI, *Addressing the challenge of deepfakes*, [accessed 17 Dec 2019], 2019. [Online]. Available: http://www.unicri.it/news/article/Workshop_Deepfakes.

[122] S. Wolpert, *Combating deepfakes: Leading scholars to discuss doctored content and how to fight it*, [accessed 17 Dec 2019], 2019. [Online]. Available: http://newsroom.ucla.edu/releases/deepfakes-workshop-how-to-combat-doctored-content.

[123] D. C. England, *Civil liability*, [accessed 19 Dec 2019], 2019. [Online]. Available: https://tinyurl.com/sglo2k7.

[124] *Advocacy groups obtain sixth circuit decision striking down ohio's political "false statement" law*, [Accessed 4 Dec 2019], 2016. [Online]. Available: https://www.jonesday.com/en/practices/experience/2016/02/advocacy-groups-obtain-sixth-circuit-decision-striking-down-ohios-political-false-statement-law.

[125] P. release from the Federal Ministry of Justice and C. Protection, *Gesetzentwurf der bundesregierung – entwurf eines gesetzes zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz – netzdg)*, [Accessed 4 Dec 2019], 2017. [Online]. Available: https://dip21.bundestag.de/dip21/btd/18/127/1812727.pdf.

[126] R. Q. Wan, *China prohibits 'deepfake' ai face swapping techniques*, [Accessed 4 Dec 2019], 2019. [Online]. Available: https://syncedreview.com/2019/04/24/china-prohibits-deepfake-ai-face-swapping-techniques/.

[127] *Ab-730 elections: Deceptive audio or visual media.* [accessed 17 Dec 2019], 2019. [Online]. Available: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730.

[128] C. Lecher, *California has banned political deepfakes during election season*, [accessed 17 Dec 2019], 2019. [Online]. Available: https://www.theverge.com/2019/10/7/20902884/california-deepfake-political-ban-election-2020.

[129] C. Karr, *Enforcing california's 'deepfake' ban could prove challenging*, [accessed 19 Dec 2019], 2019. [Online]. Available: https://www.theepochtimes.com/enforcing-californias-deepfake-ban-could-prove-challenging_3115476.html.

[130] K. Quach, *New york state is trying to ban 'deepfakes' and hollywood isn't happy*, [accessed 17 Dec 2019], 2018. [Online]. Available: https://www.theregister.co.uk/2018/06/12/new_york_state_is_trying_to_ban_deepfakes_and_hollywood_isnt_happy/.

[131] B. News, *Virginia bans 'deepfakes' and 'deepnudes' pornography*, [accessed 17 Dec 2019], 2019. [Online]. Available: https://www.bbc.com/news/technology-48839758.

[132] R. Cellan-Jones, *Fake news worries 'are growing' suggests bbc poll*, [Accessed 4 Dec 2019], 2017. [Online]. Available: https://www.bbc.com/news/technology-41319683.

[133] *Cambridge dictionary*, 2019. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/norm.

[134] K. Simler and R. Hanson, *The Elephant in the Brain: Hidden Motives in Everyday Life*. Oxford University Press, 2018, ISBN: 9780190495992. [Online]. Available: https://books.google.nl/books?id=mcM9DwAAQBAJ.

[135] F. of Life Institute, *Asilomar ai principles*, [accessed 23 Nov 2019], 2017. [Online]. Available: https://futureoflife.org/ai-principles/?cn-reloaded=1&cn-reloaded=1.

[136] Microsoft, *Microsoft ai principles*, [accessed 23 Nov 2019], 2019. [Online]. Available: https://www.microsoft.com/en-us/ai/our-approach-to-ai.

[137] Google, *Artificial intelligence at google: Our principles*, [accessed 23 Nov 2019], 2019. [Online]. Available: https://ai.google/principles/.

[138] IBM, *Everyday ethics for artificial intelligence*, [accessed 23 Nov 2019], 2019. [Online]. Available: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

[139] OECD, *What are the oecd principles on ai?*, [accessed 23 Nov 2019], 2019. [Online]. Available: https://www.oecd.org/going-digital/ai/principles/.

[140] P. release from the Federal Ministry of Justice and C. Protection, *Löschung von strafbaren hasskommentaren durch soziale netzwerke weiterhin nicht ausreichend*, [Accessed 4 Dec 2019], 2017. [Online]. Available: https://www.bmjv.de/SharedDocs/Pressemitteilungen/DE/2017/03142017_Monitoring_SozialeNetzwerke.html.

[141] E. C. P. release, *European commission and it companies announce code of conduct on illegal online hate speech*, [Accessed 4 Dec 2019], 2016. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937.

[142] J. Anderson and L. Rainie, "The future of truth and misinformation online", *Access: http://www. pewinternet. org/2017/10/19/the-future-of-truth-and-misinformation-online*, 2017.

[143] T. W. P. Drew Harwell, *Hany farid on race to detect 'deepfake' videos: 'we are outgunned'*, [Accessed 30 Nov 2019], 2019. [Online]. Available: https://www.ischool.berkeley.edu/news/2019/hany-farid-race-detect-deepfake-videos-we-are-outgunned.

[144] K. Crawford and R. Calo, "There is a blind spot in ai research", *Nature News*, vol. 538, no. 7625, p. 311, 2016.

[145] J. Angwin, J. Larson, S. Mattu, and L. (Kirchner, *Machine bias – there's software used across the country to predict future criminals. and it's biased against blacks.* [accessed 24 Dec 2019], 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[146] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges", in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 2018, pp. 19–36.

[147] H. Zeng, "Towards better understanding of deep learning with visualization", *The Hong Kong University of Science and Technology*, 2016.

[148] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks", *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[149] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization", *Distill*, 2017, https://distill.pub/2017/feature-visualization. DOI: 10.23915/distill.00007.

[150] C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, and S. Gumhold, "Visualizations of deep neural networks in computer vision: A survey", in *Transparent Data Mining for Big and Small Data*, Springer, 2017, pp. 123–144.

[151] P. Hall, W. Phan, and S. S. Ambati, *Ideas on interpreting machine learning*, 2017.

[152] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning", *arXiv preprint arXiv:1702.08608*, 2017.

[153] E. Culliford, *Twitter wants your feedback on its deepfake policy plans*, [Accessed 16 Dec 2019], 2019. [Online]. Available: https://www.reuters.com/article/us-twitter-deepfakes/twitter-wants-your-feedback-on-its-deepfake-policy-plans-idUSKBN1XL2C6.

[154] D. Harvey, *Help us shape our approach to synthetic and manipulated media*, [accessed 17 Dec 2019], 2019. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html.

[155] H. Seo, A. Xiong, and D. Lee, "Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation", 2019.

[156] P. Mena, "Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook", *Policy & Internet*, 2019.

[157] R. K. Garrett, "Echo chambers online?: Politically motivated selective exposure among internet news users", *Journal of Computer-Mediated Communication*, vol. 14, no. 2, pp. 265–285, 2009.

[158] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption", *Public opinion quarterly*, vol. 80, no. S1, pp. 298–320, 2016.

[159] E. Pariser, *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[160] M. Gao, Z. Xiao, K. Karahalios, and W.-T. Fu, "To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles", *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 55, 2018.

[161] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey", *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[162] M. Bickert, *Enforcing against manipulated media*, [accessed 11 Jan 2020], 2020. [Online]. Available: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

[163] R. u/LastBluejay, *Updates to our policy around impersonation*, [accessed 11 Jan 2020], 2020. [Online]. Available: https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/.

[164] Z. Schiffer, *Wechat keeps banning chinese americans for talking about hong kong*, [accessed 18 Dec 2019], 2019. [Online]. Available: https://www.theverge.com/2019/11/25/20976964/chinese-americans-censorship-wechat-hong-kong-elections-tiktok.

[165] E. Feng, *China intercepts wechat texts from u.s. and abroad, researchers say*, [accessed 18 Dec 2019], 2019. [Online]. Available: https://www.npr.org/2019/08/29/751116338/china-intercepts-wechat-texts-from-u-s-and-abroad-researcher-says?t=1576663883360.

[166] G. Verberg, *Are douyin and tiktok the same?*, [accessed 18 Dec 2019], 2019. [Online]. Available: https://www.whatsonweibo.com/are-douyin-and-tiktok-the-same/.

[167] M. P. Lynch, *Do we really understand 'fake news'?*, [accessed 24 Dec 2019], 2019. [Online]. Available: https://www.nytimes.com/2019/09/23/opinion/fake-news.html.

[168] equalAIs, *Equalais. empowering humans by suverting machines*, [accessed 3 Nov 2019], 2019. [Online]. Available: http://equalais.media.mit.edu/.

[169] H. H. Hi, *Zao deepfake - leonardo dicaprio*, [accessed 14 Nov 2019], 2019. [Online]. Available: https://www.youtube.com/watch?v=T89y4NYRsH0.

[170] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making", *Council of Europe Report*, vol. 27, 2017.

[171] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, "The science of fake news", *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[172] N. Chokshi, *94 percent of u.s. teachers spend their own money on school supplies, survey finds*, [accessed 25 Dec 2019], 2018. [Online]. Available: https://www.nytimes.com/2018/05/16/us/teachers-school-supplies.html.

[173] C. Sanchez, *Misinformation is a threat to democracy in the developing world*, [accessed 25 Dec 2019], 2019. [Online]. Available: https://www.cfr.org/blog/misinformation-threat-democracy-developing-world.

[174] M. Fick and P. Dave, *Facebook's flood of languages leave it struggling to monitor content*, [accessed 25 Dec 2019], 2019. [Online]. Available: https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leave-it-struggling-to-monitor-content-idUSKCN1RZ0DW.

[175] K. Garimella and D. Eckles, "Image based misinformation on whatsapp",

[176] M. Orcutt, *Blockchains use massive amounts of energy—but there'sa plan to fix that*, 2017.

[177] R. Mendelsohn, A. Dinar, and L. Williams, "The distributional impact of climate change on rich and poor countries", *Environment and Development Economics*, vol. 11, no. 2, pp. 159–178, 2006.

[178] D. H. McKnight and N. L. Chervany, "What is trust? a conceptual analysis and an interdisciplinary model", *AMCIS 2000 Proceedings*, p. 382, 2000.

[179] E. M. Uslaner, *The Oxford handbook of social and political trust*. Oxford University Press, 2018.

[180] J. D. Lewis and A. Weigert, "Trust as a social reality", *Social forces*, vol. 63, no. 4, pp. 967–985, 1985.

[181] S. P. Shapiro, "The social control of impersonal trust", *American journal of Sociology*, vol. 93, no. 3, pp. 623–658, 1987.

[182] R. G. Taylor, "The role of trust in labor-management relations", *Organization Development Journal*, vol. 7, no. 2, pp. 85–89, 1989.

[183] J. Matthes, "The affective underpinnings of hostile media perceptions: Exploring the distinct effects of affective and cognitive involvement", *Communication Research*, vol. 40, no. 3, pp. 360–387, 2013.

[184] E. Theiss-Morse and D.-G. Barton, "Emotion, cognition, and political trust", in *Handbook on Political Trust*, Edward Elgar Publishing, 2017.

[185] M. Lieberman, *Reflective and reflexive judgment processes: A social cognitive neuroscience approach. w: Jp forgas, kr williams, w. von hippel (red.), social judgments: Implicit and explicit processes (s. 44–67)*, [accessed 10 Oct 2019], 2003.

[186] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships.", *Journal of personality and social psychology*, vol. 49, no. 1, p. 95, 1985.

[187] D. Johnson and K. Grayson, "Cognitive and affective trust in service relationships", *Journal of Business research*, vol. 58, no. 4, pp. 500–507, 2005.

[188] J. Morrow Jr, M. H. Hansen, and A. W. Pearson, "The cognitive and affective antecedents of general trust within cooperative organizations", *Journal of managerial issues*, pp. 48–64, 2004.

[189] D. Gefen, "Reflections on the dimensions of trust and trustworthiness among online consumers", *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 33, no. 3, pp. 38–53, 2002.

[190] D. H. McKnight, N. L. Chervany, and L. L. Cummings, *Trust formation in new organizational relationships*. Management Information Systems Research Center, Curtis L. Carlson School of ..., 1996.

[191] P. Bromiley and L. L. Cummings, *Transactions costs in organizations with trust*. Strategic Management Research Center, University of Minnesota Minneapolis, 1992.

[192] J. A. Heise, "Toward closing the confidence gap: An alternative approach to communication between public and government", *Public Administration Quarterly*, p. 209, 1985.

[193] B. R. Rawlins, "Measuring the relationship between organizational transparency and employee trust", p. 6, 2008.

[194] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security", 2018.

[195] A. Indoensia, *Afp fact check*, [accessed 7 Nov 2019], 2019. [Online]. Available: https://factcheck.afp.com/four-photos-have-been-doctored-and-one-shows-painting-indonesian-artist.

[196] A. F. C. Indonesia, *Four of the photos have been doctored and one shows a painting by an indonesian artist*, [Accessed 20 Nov 2019], 2019. [Online]. Available: https://factcheck.afp.com/four-photos-have-been-doctored-and-one-shows-painting-indonesian-artist.

[197] *Media literacy: A definition and more*, [accessed 12 Oct 2019]. [Online]. Available: https://www.medialit.org/media-literacy-definition-and-more.

[198] W. J. Potter, *Media literacy*. Sage Publications, 2018.

[199] Serelay, *Serelay*, [accessed 5 Oct 2019], 2019. [Online]. Available: https://www.serelay.com/.

Appendix

A The Trust

In order to estimate how trust might be influenced, it is critical to have a clear picture of which aspects of trust are in our centre of attention as well as how humans arrive at a certain amount of trust towards someone or something. Thus, the aim behind this section is to provide a focused overview of the current state of trust research and some of its relevant uncertainties. With that in mind, we explain the choices made for our trust space.

A.1 The Current State of Research on Trust

The aim behind this section is not to bore the reader with dry definitions. Rather, we aim to provide a rough overview of how variably this term is used within related research. From there, we explain the choices that were made for our trust space and what the various dimensions of trust refer to in the context of deepfakes.

A.1.1 Daily Life: Many Definitions

What exactly a person is referring to when they talk about trust is highly variable. However, that might be the case for many more abstract terms. A dictionary analysis to compare the terms *cooperation*, *confidence* and *predictable* with *trust* arrived at an average of $4,7$ definitions for the first three terms. At the same time, trust averaged to $17,0$ definitions [178].

A.1.2 Research: Many Definitions

This variability is not only the case in daily life, but also within the realm of research. Trust is part of many different kinds of relations – from one person to another, from a person to a group of people or to the workings of a thing like a government, a company, an NGO or even a religious organization. Generally, the terms truster and trustee are used – the truster being the person that potentially trusts someone or something and the trustee being the entity that potentially receives trust. A truster puts a certain amount of trust in the trustee regarding a specific behavior, in a given context and at a certain time [179]. The level of trust refers to the subjective probability of the truster that the trustee will act as expected by the truster.

However, depending on the context in which the term *trust* is used, people tend to associate different things with it. Furthermore, trust is crucial in relationships of bigger entities – like between countries or companies. Furthermore, each piece of academic literature aims to support a specific finding. Consequently, the authors are incentivized to choose a definition of trust, that seems to fit their research question best. These two reasons led to a staggering variety of usages of the term and researchers have called trust definitions numerous and confusing (e.g. [180]–[182].

A.1.3 Interdisciplinary Approaches

Fortunately, the work of McKnight and Chervany [178] provides a thorough conceptual analysis and an interdisciplinary approach to trust. However, their approach is limited by its focus on interpersonal trust relationships (i.e. from one individual to another). In order to also be able to use it for trust in a thing (like a government or a company), it needs to be adjusted.

A.1.4 Another Complication: Cognitive versus Affective

Another complication stems from how exactly we arrive at a specific amount of trust. This is investigated from a cognitive perspective as well as from an affective one. The cognitive approach focuses on conscious reasoning with all the gathered information that a person has available. From the affective stance, emotional cues and responses are used to arrive at a specific amount of trust – without strenuous, active reasoning. When comparing the amount of research done from each of the two perspectives, the cognitive approach is the well-trodden path. The affective approach as well as the interactions between the two perspectives is under-explored (e.g. [183] and [184]) – despite the high probability that they do influence each other (e.g. [185]). These two systems are likely in play for may different fields – from political trust to interpersonal, media, economic trust (see e.g. [184], [186]–[188]).

We will now see how the variety in literature and the information on the two processing routes influenced the choice of our trust space.

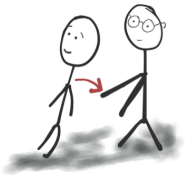
A.2 The Trust Space and its Dimensions

Our trust space adjusts and extends the interdisciplinary approach of McKnight and Chervany [178] in order to accommodate all relevant drivers for trust in the context of deepfakes. Their approach includes beliefs about the trustee – i.e. the person that may be trusted. These beliefs are called *Trusting Beliefs* and include competence, predictability, benevolence and integrity. In our space, beliefs about the context as well as media literacy of the truster are added.

These beliefs determine the trust amount of a truster in a trustee regarding a specific behaviour, in a given context at a certain time [179]. Within this thesis, the truster will always refer to a citizen, while the trustee will refer to an entity (i.e. a person or a thing, like a government or a company). The behavior may sometimes be rather abstract – an example behavior might be to *do well in the voted-for political position* or to *provide a satisfying product, which fulfills its intended function*. The context includes elements that are not about the truster or the trustee, but about the environment and its potential consequences for the result of the situation. The time of the situation in question influences the context, the truster's beliefs, the truster's media literacy and therefore also the truster's trust. For the sake of simplicity, we are only looking at static time-slices within this project. Thus, (before the conclusion) we will disregard how the mentioned aspects might develop over time.

Our handling of the tensions and uncertainties around the cognitive and the affective approach is the following. As there is few research on the interactions between the two, we will assume neither of them is clearly prevailing over the other. In other words, we assume that both are somewhat relevant. However, it is good to keep in mind, that results might need to be adjusted as soon as state-of-the-art gains more clarity on the interaction and relative importance of the two types of processing.

The following subsections provide short definitions and relevant examples for each of the dimensions. Figure 21 of the included dimensions. All of the dimensions have a direct influence on the truster's (trust-)decision. Most of the dimensions are beliefs that a truster has about a trustee. Furthermore, there are beliefs about the context and a relevant attribute about the truster.



A's Beliefs about B	Beliefs about Context	A's Attributes
- Competence - Integrity - Benevolence - Predictability - Transparency - Value Congruence	- Deception Capabilities (DC) 1. A's beliefs about DC 2. A's beliefs about B's awareness about DC	- Media Literacy

Figure 21: Dimensions overview

A.2.1 Competence

A truster has a certain belief about the competence of a trustee. This belief is determined by how confident the truster is that the trustee has the sufficient skills and performance ability for the task in question (e.g. [189] and [190]).

The most straight forward example to show that we humans usually care about competence when deciding whom to trust might just be the following. Imagine you need a high-risk operation and there are two surgeons available. One whom you believe to be competent and another one whom you believe to have lower competence. You would surely put your trust and thus your decision on the former.

A.2.2 Integrity

Definitions of integrity often focus on the honesty and truthfulness (e.g. [189] and [190]) and the fulfillment of promises ([191]). Accordingly, a truster believes a trustee to have integrity, if the truster believe the trustee to be honest, truthful and caring about fulfilling their promises.

If someone is honest and cares about whether they are fulfilling their promises, they are more likely to fulfill what they have promised. Consequently, all else being equal, it does make sense for you to rather put your trust regarding a specific hoped for behavior into a person who acts with strong integrity than in one that exhibits only meager integrity.

A.2.3 Benevolence

If a you believes that someone has an attitude of kindness and good will towards you [189] and if you believe that that person cares about your concerns ([189], [190]) and aims to act in your best interest [178], then you believe that that person is benevolent towards you.

It is probably intuitively clear to you, that benevolence is a crucial factor for trust. But let us still have a look at a quick example. Imagine you have two superiors in your current position of employment. Furthermore, you have a dilemma which combines professional choices and a personal problem or weakness. You think advice from any of your superiors would be helpful, but at the same time, opening up to a superior would make you vulnerable to them. Again, all else being equal, you would probably choose to trust the superior who you estimate to be more benevolent towards you.

A.2.4 Predictability

The meaning of predictability is close to the meaning of integrity. However, if a person is believed to be predictable, it only means that it is believed to be possible to foresee their actions – whether or not they are truthful or desired. Thus, it is a less value-laden term [178].

For a quick example that even predictability of negative details can be a relevant driver for trust, let us imagine you need to call a plumber. All else being equal, you would probably prefer to hire the plumber who predictably arrives five minutes late, to the one that arrives mostly on time but every now and then is an hour late.

A.2.5 Transparency

A truster believes that a trustee (another agent, party or thing) is transparent if the trustee seems to present relevant information openly to the truster. The information should be accurate, timely, balanced and unambiguous. Importantly, this includes information which might reflect negatively on the trustee. Here, information is considered relevant if it might have a bearing on the truster's decision about their relationship to the trustee (see [192] and [193]).

Let us assume you want to get honest feedback about the chocolate cookies, that you to baked to impress your partner's parents. You consider two friends, friend A and friend B. You remember friend A telling you that their colleague gave them a souvenir from their stay in Paris. Friend A did not like it but pretended they would (in order to not hurt their colleagues feelings). In contrast, friend B tends to tell you directly if they do not like something you did or chose. You would probably trust friend B more for the feedback on the cookies and rather ask them for their opinion.

A.2.6 Value Congruence

If a truster believes that theirs and the trustee's values overlap strongly then the truster believes that the two have a high value congruence. It makes sense, that value congruence is a driver for trust, because value congruence is a proxy for predictability and potentially even for benevolence.

For instance, let us assume that you are confident that you and your best friend generally have the same values. Then, you are probably rather trusting this friend to babysit your child because you expect them to make similar decisions as you would (e.g. what kind of explanation to give if the five year old would ask about where babies are from or whether Santa exists).

A.2.7 Deception Capabilities

The strength of deception capabilities and one's awareness of these capabilities can have a number of consequences. Let us inspect them separately from the perspective of the truster and the trustee.

Like in the above-mentioned example of the Austrian politician who considered corruption, a trustee can be tempted to deny a piece of video-evidence. This is what some refer to as *liar's divident* [194]. Whether this strategy works out

in the trustee's interest, depends on the quality of the video-evidence, as well as on how aware the trusters are of the state-of-the-art deception capabilities. Additionally, it is also the case that if the trusters know about the current deception capabilities (or if they believe them to be higher than they actually are), the trusters believe that the trustee is more likely to behave badly due to higher chances to get away with it. Thus, the trust of the trusters might decrease further.

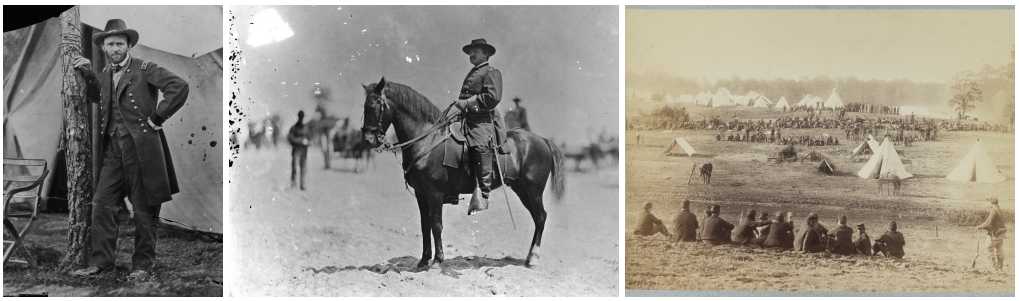
At the same time, if the truster believes that the state-of-the-art deception-capabilities are at least as high as they actually are, then the truster will be uncertain about the information they are finding online. Naturally, if the trusters disagree with the depicted message, they will be tempted to deny also real video-evidence. One might argue that we could simply adjust and ignore any video-evidence. However, we do not seem to be doing this currently for photos, despite the common knowledge that photo-realistic pictures can be created.[57]. This may be especially detrimental when combined with low media literacy. But even if we would nevertheless be able to ignore information from videos, we would expect the following. Via the cognitive route, our views should change little to not at all. However, the affective route might still update somewhat.

A.2.8 Media Literacy

In contrast to most of the above dimensions, this dimension is not about a belief of the truster about the trustee, but about a skill of the truster. Consequently, deepfakes cannot directly change this dimension. Furthermore, *media literacy* seems at first very close to *awareness of deception capabilities*. However, that is the case because the latter (being aware of state-of-the-art deception capabilities) is a necessary condition for the latter (media literacy). More specifically, media literacy is defined as a framework and set of tools of how to access, evaluate, create and interpret existing forms of media [197], [198].

Media literacy in the context of deepfakes is relevant for the critical evaluation of videos – e.g. knowing how to gather information on whether a given video might be manipulated in order to deliver a different message than intended. A person with high media literacy might be familiar with the different ways of how reality can be presented in a distorted way – from presenting parts out of context, to deceptive omission or splicing, up to doctoring and fabrication.[58] Furthermore, they might know about projects and tools like Serelay [199][59] or the European InVID [110] project to reveal manipulated videos and a browser plug-in [111] to detect fake videos.

B Components of the General Grant Image



(a) Head picture of Grant [13] (b) Horse and body picture of Alexander McDowell McCook [13] (c) Background picture of confederate prisoners [13]

Figure 22: Components of the composite image of Grant

[57] See for example the widely shared post [195] A real photo was edited into multiple pictures to create the false impression that the photo had attracted wide international attention [196]

[58] See for example the Washington Post's guide [115]

[59] With Serelay one can take pictures that include verification to ensure that pictures have not been tempered with.