

## Module 9

### AI manipulation in social media

*"Arguing that you don't care about the **right to privacy** because you have nothing to hide is no different than saying you don't care about **free speech** because you have nothing to say." - Edward Snowden*



# About the module

This module consists of two parts and both deal with a different perspectives of **manipulation in social media** . The two parts can be used separately, but can also be used in conjunction with one another.

The first part of the module deals with the handling of **private data on social media platforms**. The students deal with their own behavior and reflect on the disclosure of their data. In a further step, the extent to which this data is used by companies to generate customized advertising is discussed. Finally, this part of the module deals with the difficulty of deleting data on the Internet and discusses the possibilities.

The second part deals with the **phenomenon of "deepfake"** as an excerpt on the topic of spreading disinformation on the Internet. The aim is to clarify what deepfakes are, how they are created and what opportunities and risks they entail. There is also the possibility to create deepfakes themselves with apps such as Wombo and to discuss the results in the plenum.

## Objectives

The students can ...

- ... Analyze and evaluate the social media channels they use
- ... Knowing, reflecting on and considering risks and dangers in digital environments
- ... Develop and apply strategies for protection
- ... develop an understanding of the value of data and understand that they pay for supposedly free offers on the Internet with their data
- ... analyze their own data trails
- ... explain the intentions and strategies behind deepfakes
- ... assess the dangers of fake news in the form of deepfakes for democracy and society
- ... Recognizing and describing deepfakes as a special form of **AI**-based disinformation

## Agenda – Part 1: Manipulation by algorithms

<b>Time</b>	<b>Content</b>
15 mins	Discuss opening questions in groups and in plenary
15 mins	Thought Experiment + Video
5 mins	Social Media Data Quiz
5 mins	Theory - Algorithms and the GDPR
40 mins	Research Task - Show me your account and I'll tell you who you are
15 mins	Theory - The Internet Never Forgets
15 mins	Brainstorming - Can images be permanently deleted from the web?

## Agenda – Part 2: Deep Fakes

<b>Time</b>	<b>Content</b>
15 mins	Quiz - Deepfake or Real?
10 mins	Brainstorming - What are deepfakes?
20 mins	Research - how deepfakes work
10 mins	How to unmask manipulative deepfakes?
20 mins	Task - create deepfakes and reflection

# Introduction

## Manipulation through data on social media

This module begins with a well-known quote from former US secret service agent Edward Snowden, who quickly attracted a great deal of attention by publishing surveillance strategies for the US and Great Britain. This quote sums up the core of the learning objectives of this module: **the focus is on the right to privacy in the digital space and how carelessly we deal with our private data in this public space**. All of our activities are meticulously documented on the Internet and result in a detailed profile of our interests, our consumer behavior and the contacts we maintain.

As a start to the teaching module, students can discuss a questionnaire in groups and record the results to present in the plenary (e.g. through a poster or through a padlet page).

The following questions are suitable as input:

- What apps and websites do you use regularly?
- Do your social media providers (e.g. Instagram) know your real name?
- On your social media profiles, can everyone see your pictures or only your contacts?
- Why are you active on social media platforms?
- Have you ever read a privacy policy? Are you interested in what data is collected from you?
- Have you ever googled yourself? If so, did anything surprise you? Why?

Since these are personal questions, students should be able to choose **how much they want to reveal about themselves** in these group discussions.

Often **collecting data is not obvious**, which means that we are not at all or only less aware of the consequences of our behavior in the digital space.

As an example, the following scenario can be described by the teacher: The students should imagine that all the people they meet on the street know immediately where they were just before, where they bought their clothes, what they just ate or what they last googled.

- How would the students feel?
- How aware are they of what information they are posting online and who has access to it?

As a visual example, the **video** by Tomatolix can also be shown during this lesson. On the one hand, this video shows ways of finding out for yourself what data Internet platforms store, on the other hand, the reporter shown tries to find a person in real life only with the help of social media posts (and manages to do so within a very short time). The corresponding sequence begins in the video at minute 7:00 and lasts about 8 minutes. Here, too, the students can then be asked how they feel about it and whether they think it would also be possible to simply track them down with their profiles. Among other things, it is interesting to address the "Snap Map" feature of Snapchat and the associated opportunities and risks, since the live location of the user is displayed here. On the one hand, appropriate filters can be suggested (e.g. a Vienna filter if you are in Vienna) and the location information can be interesting for friends, on the other hand, this function offers a breeding ground for stalking.

After the students have talked about their usage behavior on social media platforms, a short **quiz** can be played in the plenum. The students should guess which data platforms such as Instagram, Snapchat and TikTok collect. The quiz can either be completed using the worksheet or the quiz can be transferred to a platform such as kahoot or the students' assumptions can be discussed directly in the plenary session.

Further you can ask the question why they are collecting all this data? What benefit do you get from it? Social media platforms **use our data to offer it to companies** so that they can place targeted advertising. Because Instagram or TikTok know exactly that you are female and between 15 and 25 years old and that you are interested in sports and sustainability, you will then be shown advertisements such as bamboo yoga mats or organic sports nutrition without additives.

So social media platforms do not primarily want to connect people with each other, but only see it as a means to the end. Rather, it is about using personal data so that other companies can **place efficient advertising. So even if these websites are free at first glance, we pay passively.** We pay for our time and attention, which the websites use to collect data about us and resell it for advertising by companies.

The students should now reflect briefly, but also check their apps to see what data they are publicly disclosing (date of birth, place of residence, gender, relationship status...) and how comfortable they feel about this information being passed on and shared are also public. As a result, it can be noted that some things can be improved by adjusting the privacy settings. The students can independently research which data can and want to be set to private. In the plenary session, you can discuss what options you have to protect your personal data.

## The GDPR and what it means

What data protection basically means is defined differently from country to country. This can lead to problems, since there are no national borders on the Internet and personal **data often travels through several countries** before it reaches its destination. For this reason, there are international agreements such as the "EU-US Privacy Shield". This agreement regulates data exchange between the USA and Europe. Within Europe, the General Data Protection Regulation ("GDPR" for short) has been in effect since May 25, 2018, which regulates the processing of personal data by authorities and companies. The GDPR obliges companies to provide users with information about their stored data. If this is not implemented, high fines can be imposed.


Further information (in german) and exciting explanatory videos on the GDPR can be found on the website: [deinedatendeinerechte.de](https://deinedatendeinerechte.de)

## Algorithms know what you need

These seemingly well-tailored ads use algorithms that are fine-tuned **based on user browsing history and metrics**. Advertisements can be created for specific target groups and thus match the interests and previously viewed content of the users. However, these recommendation algorithms can create a completely **insightful world view** among users and even lead to **radicalization of users**.

It is therefore important that we as users recognize and distrust distortions caused by one-sided offers. Another important step is making **algorithms transparent, ethical guidelines and an analysis of previously used algorithms**. For example, an NGO pushing these demands would be AlgorithmWatch.

## Material

-  Social Media – Introduction.pdf

## References

1. (GER) Video: Das weiß das Internet über dich! – Selbstexperiment
2. (GER) Deine Daten Deine Rechte

# **Show me your account and I'll tell you who you are**

In the first steps, the students have already reflected on which data they disclose publicly and how comfortable they feel with it. Now the recommendations, suggestions and advertisements of the profiles are to be analyzed in small groups and only based on this a characterization of the user is to be created. It is important that nobody should be forced to do anything and that only those profiles are analyzed that feel comfortable with it.

In a further step, the students can now independently request and analyze the data stored in their social media accounts. Instructions on how to request this data can be found in the appendix for the more frequently used platforms Instagram, YouTube and TikTok. Basically, you can also find detailed instructions on the Internet and in the FAQs of the platforms. Requesting the data can take anywhere from a few hours to a few days. It would therefore be advisable to request the data the hour before.

## **What does Instagram, TikTok and Co. know about me? Analysis of your own data in small groups (2-4 students)**

Due to the GDPR, you have the option of requesting all of your stored data via Instagram, Facebook and Co. But do you know what data these sites actually have?

### **Part 1 – Analyze your profiles' advertisements, recommendations, etc.**

1. In your small group, select one or more platforms that you are logged on to. You can analyze one profile together or several profiles in a row. But always talk to each other in the group and treat each other with respect! Nobody has to show their profile if your classmates don't want it!
2. In the following you try to characterize the owners of the profiles based on the content that is presented to you. Pay less attention to the things that the person has liked themselves and more to those that are suggested to them.



3. Click through the app's feed to do so. What ads do you see? Which videos or profiles are suggested to you?
4. Note any qualities, interests, hobbies, etc. that the profile owners might have.
5. How closely do the characterizations resemble the actual profile owners? What surprised the profile owners? Think about why you think some ads or suggestions don't fit the person, but are suggested anyway.

## **Part 2 – Request your data**

### **Instructions for Instagram**

1. Go to your own profile.
2. Tap the gear icon to go to settings. From there, select Privacy & Security.
3. If you scroll down, the Data Download section will appear. Tap Request Download.
4. Enter the email address you use for your account.
5. Enter your Instagram password.
6. You should receive your report within the next 48 hours.
7. Download the information from the mail and unzip the folder.
8. If you click on the index.html file, you will be taken to the Instagram website where you can click through all of the information the company holds.

### **YouTube Instructions**

1. Open your profile in the top right corner and select "My data on YouTube".
2. On YouTube, you can view your results directly on the website.

### **TikTok Instructions**

1. In the TikTok app, tap Profile at the bottom.
2. Tap the Menu button at the top.
3. Tap Settings and privacy.
4. Tap Manage Account, then Download Your Data.
5. You should receive your report within the next 3 days.

## **Part 3 – Analyze your reports**



Check the reports.

What surprises you? How did you know that this app saves this information?

These self-experiments make it clear how much data is continuously produced. Up until a decade ago, this was mainly done through PCs. In the meantime, data is not only constantly being generated by smartphones and their apps, but also new technologies in the area of the Internet of Things (IoT) such as wearables (e.g. for measuring heart rate and movement profiles) or small environmental measuring stations (e.g. for measuring air quality) produce an incredible amount of data. In 2020, 1 GB of data will be produced per user per day! <sup>1</sup>

On Instagram, for example, more than 1000 photos are uploaded per second, that's about 100 million pictures per day! <sup>2</sup>

## Material

-  Social Media – Introduction.pdf
-  Social Media – Worksheet Research.pdf

## References

1. Wirtschaftsforum.de: data consumption
2. Instagram Statistics

# The internet never forgets!

This data doesn't just disappear from the internet either. In addition, much of this content is accessible unencrypted. To show that data on the web, whether encrypted or unencrypted, is forever out of your control, students can use the **Wayback Machine** can be taken on a little journey through time. The students can explore their own school website, club websites or blogs. Afterwards, there should be a **discussion about the advantages and disadvantages of such internet archives** and what you can try as a user if unpleasant content gets onto the internet.

If private content, such as naked pictures, finds its way onto the Internet, nobody simply has to accept it. Because even if a final deletion is extremely difficult, it is still worth taking action against it! The authors can be contacted on social media platforms. If they do not remove the content, they can be reported directly if they violate the guidelines (nudity, violence, etc.). In other cases, appropriate moderators and the site operators can be contacted. With other websites and blogs, too, you should first contact the owners of the site and ask them to delete the content. If the request is not complied with, a report can be made to the police. Among other things, this can refer to the "right to one's own picture" (§ 78 UrhG) or to "pornographic depiction of minors" in the case of nude pictures of under 18-year-olds.<sup>3</sup>

The following website is suitable for further information: [www.ombudsstelle.at](http://www.ombudsstelle.at)

## Material

-  Social Media – Introduction.pdf
-  Social Media – Worksheet.pdf

## References

1. [saferinternet.at](http://saferinternet.at)

## Part 2: Deep Fakes

Due to time constraints, this part of the module mainly deals with video deepfakes, although these only make up a small part of digital disinformation and fake news. Other exciting, in-depth aspects of manipulation in social media would be, for example, the function of social media trolls or social bots.

### What are deepfakes?

Deepfakes are **manipulated or artificially created sound or image media that appear real**. They show people who appear to be saying or doing something that they have never said or said before. Deepfakes are created using **artificial intelligence** such as machine learning and deep learning.

Thanks to new technological developments in the field of image processing and manipulation, deepfakes also appear more and more authentic. On the one hand, **algorithms have been developed** and improved in computer vision that automatically recognize and map facial structures (e.g. the position of eyebrows and nose), resulting in new technologies in face recognition. On the other hand, the triumph of the Internet – and in particular through platforms on which images and videos are shared – has created an incredibly **large data pool** with audiovisual data that can be used for this.

Two specific **AI** approaches are commonly found in deepfake programs: **Generative Adversarial Networks (GANs)** and **Autoencoder**. GANs are machine learning algorithms that can analyze a series of images and thereby create new images of comparable quality. Autoencoders, on the other hand, can extract information about facial structures from images and use that information to model a new facial expression.

Because these techniques can be used to realistically simulate facial expressions and types of movement of a person, it is now very difficult to tell whether you are looking at a deepfake or the original. However, not only the facial expressions of an existing face can be changed: Faces can be exchanged and created from scratch.

Convincing Computer Generated Imagery (CGI) technologies have been around in film and cinema for many years. For example, The Curious Case of Benjamin Button won the 2009

Oscars for Best Visual Effects. Brad Pitt, the film's protagonist, was reverse-aged using computer-based manipulations.

Manipulating the media and image processing are by no means new phenomena. Deepfakes are just a technological advancement of a much older phenomenon, so to speak. The emergence of social media platforms and the lively exchange and sharing of content (and thus also false content, e.g. fake news) has changed the media landscape significantly. In addition, apps like Snapchat, Instagram and TikTok already offer low-threshold filters within the applications that can be used to change faces and edit videos.

In addition, the **rise of visual media, particularly video**, as a means of communication is also significant. Visual media are considered to be a particularly efficient way of disseminating information. So far it was well known that misinformation in texts or photos is manipulated, but video was still considered by many to be hard evidence that was difficult to forge.

**False information** can be spread through deepfakes, and some users can no longer distinguish between truth and fiction. Many of these reports are deliberately created to cause some form of harm. The **spread of deepfakes creates uncertainty among Internet users**: What is the truth and what is a fact? In this case, which media can still be trusted and who is manipulating its content? Due to the sole existence of deepfakes, many users are no longer sure which content can still be trusted. <sup>4</sup>

Deepfake technologies can be used for a variety of purposes, with both positive and negative effects. Deepfakes can be helpful, for example, in the area of audiovisual media productions (e.g. if an actor is absent), human-machine interactions can run better, but they can also find a place in areas such as video conferences, satire and art projects or surgical facial reconstruction. However, there are also a number of negative aspects, such as blackmail, defamation, bullying, identity theft, damage to reputation, manipulation of news media, loss of trust in science, business and politics, manipulation of elections, damage to international relations and national security.

**Debunking deepfakes** can take a long time, which means that seemingly small videos can lead to big problems. In the course of the lesson, the students should independently find examples of who uses these technologies and who they can

harm. By playing through examples, you can get an idea of the dimensions that a manipulative deepfake can take on.

## A possible fictional example of the effects of deepfakes




An apparently real video was uploaded to Instagram and Twitter of a politician confessing on camera to evading millions of euros. This video not only damages the politician's reputation and inflicts psychological damage. For example, the politician or the party could be blackmailed by threatening further forged confessions. Voters are losing trust in the party and will not vote for it again in the next election. This distrust can go so far that the system is generally no longer trusted.

A video from funk.net (an offer from ARD and ZDF) offers further input on recognizing deepfakes: [www.funk.net](http://www.funk.net) – Do you recognize the fake?

Finally, in the plenum can be discussed how dangerous deepfakes can be prevented. Possible approaches would be, for example: do not post any videos of yourself on the web, avoid voice messages, do not allow unwanted ones to be recorded and insist on deleting the photos/videos, strict laws regarding Deepfakes, stricter controls (especially on social media platforms) to curb the spread.

From a legal perspective, there are no concrete measures or laws to date. However, strategies are already being developed, such as the deepfake action plan of the Austrian federal ministries. However, no concrete changes in the legal situation are required, but awareness-raising among the population and the use of software tools that are intended to recognize deepfakes and fact-checker platforms.<sup>5</sup>

## Material

-  Social Media – Deepfakes.pdf
-  Social Media – Worksheet Deepfakes.pdf
-  Social Media – Worksheet Research Deepfakes.pdf



## References

1. saferinternet.at
2. Deepfake Action Plan
3. How to detect deepfakes | Deepfakes explained (english)
4. Fake videos of real people – and how to spot them (english)
5. ganz konkret: Deepfakes gegen Fakten? | Zeit für Politik (german)
6. Deepfakes: Is This Video Even Real? | NYT Opinion (english)
7. How Dangerous are Deepfakes? | Explained (english)
8. Täuschung mit Deepfakes | Odysso – Wissen im SWR (german)

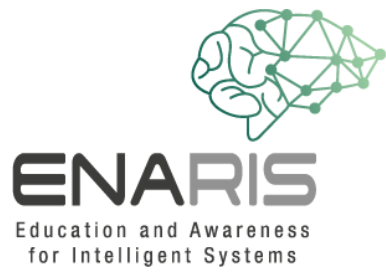
# We create our own deepfakes

In the last step, the students can try to **create convincing deepfakes** independently with the help of apps. A helpful app for this is the image manipulation app **Wombo**, which you can use to make selfies sing. Another possible app for this would be **Reface** (note: the paid Pro mode must be clicked away at the top of the X at the beginning)

## Material

-  Social Media - Deepfakes.pdf
-  Social Media - Worksheet Deepfakes.pdf





EUROPEAN UNION

