MODULE 3.
IDENTIFYING AI-GENERATED CONTENT

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the National Agency. Neither the European Union nor the granting authority can be held responsible for them.

Co-funded by the Erasmus+ Programme of the European Union

Module 3: Identifying AI-generated content

The rise of AI-generated content has made it tougher to tackle false information and disinformation. AI can create highly realistic fake videos, images, and text at a rapid pace, making it hard for fact-checkers to keep up. This means that false information can spread quickly and appear very convincing, making it challenging to tell what's true and what's not. As a result, misleading content can easily go viral, damaging public trust and creating confusion.

At the current state of AI, it is still possible to identify AI-generated content, either by paying attention to certain flaws or by making use of helpful tools. This module will explore both the challenges of AI as well as the options users have to tackle these challenges.

3.1 Activity 1: Input – Introduction and AI-generated Text

Type: Input

Duration: 10 minutes

Learning outcomes: Participants will have an improved understanding of generative AI, its challenges for the spread of disinformation and challenges of detecting AI generated text.

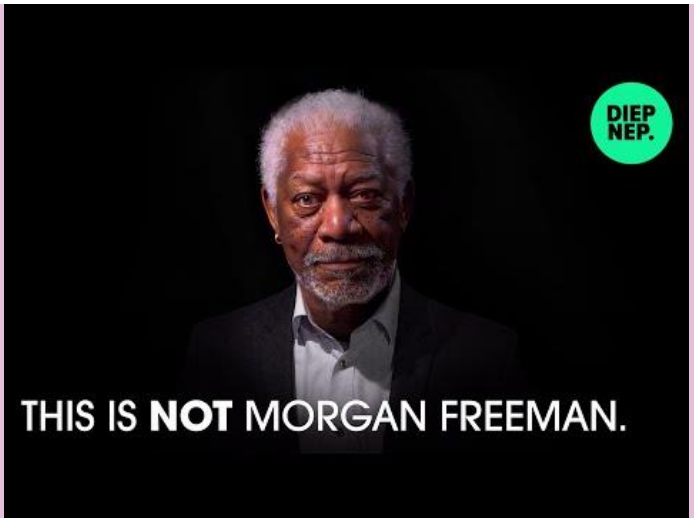
Materials/Resources needed: Notebooks and pens for participants. Input script for facilitator. Large screen or projector and a laptop connected to it for facilitator.

Optional: Facilitator can prepare a PPT adjusted to the group's age and background knowledge.

The facilitator uses the following information for their input.

> **Introduction:**
>
> AI can generate deceptively realistic images of events that never occurred, artificially generated voices, and sophisticated texts without human authors. It's getting harder to tell what's real and what's fake. This highlights a general truth about technology: it can be both a blessing and a curse, depending on how it is used and for what purpose.
>
> While much of the AI-generated content online today is the result of users exploring their creativity, AI can also be also employed to deceive and spread disinformation.
>
> Generative AI is a type of artificial intelligence that can create new content, like texts, pictures, music, or videos, by learning from existing data. It uses smart computer programs called neural networks to create things that look or sound like they were made by humans. Examples include chatbots like ChatGPT that write texts, AI tools that make art, and programs that create fake videos.

- *Show video:* https://www.youtube.com/watch?v=oxXpB9pSETo



AI-generated text:

AI-generated text is created using a type of artificial intelligence called language models. These models are trained on large amounts of text data from books, articles, and websites. By learning patterns and structures in the data, the AI can generate new text that sounds like it was written by a human. When given a prompt or a topic, the AI predicts the next words and sentences based on what it has learned.

AI-generated text poses a challenge with regard to disinformation because it can create realistic and convincing articles or social media posts. This makes it easier to create deliberately misleading content. The content can be produced in large volumes and tailored to target specific audiences.

It is also important to remember that AI generated content does not necessarily have to be true. One reason for this is that the language models behind the chatbots themselves have no sense of reality and no idea of what is true and what is false. We must be aware that an AI - no matter how authentic its responses may sound - has no knowledge of the world around us knowledge and no human perception. So while an AI can calculate which word is most likely to follow another word, the AI itself cannot grasp the meaning of the resulting text. How likely it is that the text is factually accurate depends on how often a sequence of words appeared in the training data.

Recognizing AI generated text:

It becomes increasingly hard to identify AI-generated text, especially for the average user. However, some online tools can help identify AI content.

For example:

GPTZero
https://gptzero.me/

AI Detector by Content at Scale
https://contentatscale.ai/ai-content-detector/

Show one of the AI detectors and briefly explain how to use it:

Simply copy the text into the text box or copy a URL into the URL field.

You will then receive an estimation of how likely this content is human-generated.

However, it's important to know that this estimation is not fully reliable! It is only a prediction and not 100% accurate. It can help to give you an idea if a text was created by AI, but you should never blindly trust it. These tools also works better in English than in other languages.

Participant Instructions: Listen to the facilitator's input, take notes and ask questions if anything is unclear.

Possible alterations: Depending on the participants' age, knowledge and background, the input can be adjusted. In general, as a lecture-type activity, the facilitator can implement it in a way that best suits the group. However, all necessary information should be transmitted in order for the following activities to be feasible.

Debrief: Ask participants if they have any questions.
Background reading/External resources:

https://www.digital-mobil-im-alter.de/wissen/kuenstliche-intelligenz/ki-generierte-inhalte (in German)
https://www.makeuseof.com/accurate-ai-text-detectors/
https://research.ibm.com/blog/what-is-generative-AI
-

3.2 Activity 2: Test AI text detectors.

Type: Group Activity + Discussion

Duration: 10 minutes

Learning outcomes: Participants will know how to use AI-text detector tools and will be more aware of their limitations

Materials/Resources needed:

One device per group with a ChatGPT account (or other AI text generator)

Laptop and large screen / projector for facilitator

Instructions for Facilitator:

Divide the participants into groups of 2-3 people. Each group should have one device with a ChatGPT account.

After 6-7 minutes, ask participants to come to an end and discuss their findings with the debrief questions.

Participant Instructions:

- Log into ChatGPT. Create an article on a topic of your choice. You can deliberately choose to create a false article, for example on Chemtrails
- Copy the text into one of the AI Detector tools. How accurately does it recognize identify the AI-generated text?
- Try to adjust your original text with prompts like "Rephrase in a way that the text sounds less like an AI-generated texts" or try different languages with prompts like "Translate this text into English"
- Copy the adjusted text into the AI Detector. How has the rating changed?

Possible alterations:

If participants don't have ChatGPT accounts or don't want to set one up, this activity can also be done together with the full group. Participants tell the facilitator, who's sharing their screen, which prompts to write. Facilitator then copies the text into an AI text detector and discusses the outcome with the group.

Debrief: What did you notice? How reliable were the Detectors in your test? What does this show you about the challenges of recognizing AI generated content?

Background reading / External resources:

https://www.makeuseof.com/accurate-ai-text-detectors/

3.3 Activity 3: Input and Quiz – AI generated Images

Type: Input and Quiz

Duration: 10 minutes

Learning outcomes: Participants will have an improved understanding of AI generated images and how to spot them

Materials/Resources needed:

- Notebooks and pens for participants.
- Input script for facilitator.
- Large screen or projector and a laptop connected to it for facilitator.
- Annex 3: Images

Either show the posts on a large screen (projector/TV) via a laptop and PDF/PPT file.
Or Preferably: Mentimeter Quiz (PC + Projector/TV, one smartphone / tablet per participant)

1. Set up a Multiple Choice poll on www.mentimeter.com (or comparable platform).
2. Set up five slides and add the five example images from Annex 4, all with the question "Do you believe this image is real or AI-generated?"
3. Add one image image per slide (it might not be displayed ideally in the slide due to the format, but participants will see it correctly on their devices). Add the possible answers "Real" and AI generated" and tick the box for the correct option
4. Disable "More points for fast correct answers"

Instruction for Facilitator:

Give a short introduction to the topic of AI-generated images and show the checklist on a large screen

Recognizing AI-Generated Images and Videos

AI-generated content like images and videos is created using advanced neural networks trained on vast datasets. These networks learn to generate new content by recognizing patterns in the data they are fed. Given a description or an idea, they can produce highly realistic images or videos. However, despite their increasing sophistication, there are still ways to identify AI-generated content by looking for specific clues.

Key Indicators to Identify AI-Generated Content

- **Anatomy and Limbs:** AI often struggles with accurately rendering human anatomy. In images, look for hands with too many fingers or oddly shaped limbs. In videos, watch for unnatural body movements, as deepfakes might produce mechanical or repetitive motions.
- **Text and Symbols:** AI-generated images might mishandle text, resulting in misspelled words, blurry letters, or strange symbols. Similarly, in videos, any on-screen text that appears off can be a giveaway.
- **Hair and Skin:** Human hair in AI-generated images may look like paint strokes rather than flowing naturally. In videos, the skin might appear too smooth or too wrinkled, and inconsistencies between the skin and other features like hair and eyes can be indicators of manipulation.
- **Symmetry and Consistency:** AI can struggle with symmetry. In both images and videos, one side of a face might look different from the other. Consistency issues are also a clue;

if you have multiple images or frames of the same subject, inconsistencies among them can suggest AI generation.

- **Textures and Patterns:** Rendering realistic textures, patterns, and fabrics is challenging for AI. Small discrepancies in how surfaces appear can indicate an image is AI-generated. In videos, look for unnatural textures on objects or backgrounds that seem out of place.
- **Geometry and Angles:** Check how objects fit together in the scene. In images, odd angles or elements that don't align naturally can be a sign of AI generation. In videos, watch for inconsistencies in spatial relationships and object interactions.
- **Facial Features and Movements:** Deepfakes often involve facial transformations. Look for unrealistic skin textures, mismatched age indicators between different facial features, and unnatural shadows or reflections. Pay attention to the blinking patterns, as deepfakes may blink too much or too little. Also, examine lip movements to see if they match the spoken words naturally.
- **Details and Refinements:** AI-generated content can sometimes miss finer details. In images, this might be evident in the way light interacts with surfaces, producing unrealistic highlights or shadows. In videos, pay attention to the reflections in lenses and small features like moles or facial hair, which may not appear as natural.

Now move on to the quiz:

When using a menti poll, share the code with the group. Show the "Present" mode on mentimeter.com on the large screen. Ask participants to select an answer.

Once everyone has voted, the results will show on the main screen.

If you're not using a menti poll, show the posts one by one on a large screen and let participants vote if they think an image is real or AI-generated, e.g. by raising their hands.

After each image, briefly discuss what were the indicators for the AI generated images.

Here's the solution to the quiz and some points to note:

1: German Chancellor Olaf Scholz : Real

- German Chancellor Olaf Scholz had an accident while running in a forest in September 2023. He fell and had injuries in his face. To cover his injured eye, he wore an eye-patch and posted this photo publicly on his X page
- Source: https://x.com/Bundeskanzler/status/1698627434508063228
- Photo credit: Steffen Kugler/Bundesregierung

2: Donald Trump: AI-generated

- This image was created by supporters of Donald Trump
- Fact check here: https://www.bbc.com/news/world-us-canada-68440150
- Indicators it's fake:
  - People are missing fingers
  - The space between two fingers of Trump's left hand looks unrealistic
  - The text on the hat on the left is blurry and unreadable
  - Texture of the skin: too shiny

3: Katy Perry: AI-generated

- This image circulated during the MET Gala 2024. However, Katy Perry was not even present
- Fact Check here: https://www.politifact.com/factchecks/2024/may/07/facebook-posts/ai-generated-images-of-katy-perry-at-2024-met-gala/
- Indicators it's fake:

- In the background, the faces, hands, arms and bodies of the photographers are not clearly separated from each other or from the surroundings, but flow into each other

4: Giorgia Meloni: Real

- This picture was taken on 9th February 2023 at the Elysée Palace
- Source: https://www.faz.net/aktuell/politik/ausland/darum-aergert-sich-giorgia-meloni-ueber-emmanuel-macron-18672739.html
- Photo Credit: EPA

5: Tom Cruise: Fake

- Fact-Check: https://www.snopes.com/fact-check/tom-cruise-stunt-doubles/
- Indicators it's fake:
  - Hand of alleged Tom Cruise double in the middle seems to miss finger nails
  - Texture of the skin is too smooth, shiny and wax-like

After the quiz give more advice and input on recognizing AI generated content

Advanced Detection Techniques

- **AI-Based Detection Software:** Various tools are being developed to help detect AI-generated content. These programs analyze inconsistencies in images and videos that might not be visible to the human eye. They can be particularly effective in identifying deepfakes by examining minute details and comparing them against known patterns of manipulation.

For images e.g.:
Hive Detection Tool:
https://hivemoderation.com/ai-generated-content-detection

For videos, e.g.:
Deepware Scanner or DeepFake-o-meter.

- **Human Analysis:** Despite the advancements in AI detection software, human intuition and analysis remain crucial.
- **Use common sense**: The same applies to all information on the internet. Pause, remain skeptical and critical and question what you see andhear. What motives could there be for putting something like this into the mouths of certain people? How realistic is the scenario?

Instructions for Participants:
Open menti.com on your handheld devices. Enter the code the facilitator gives you. Look at the image that shows on your device and choose whether you think it's real or AI-generated

Possible alterations: Facilitators can also use their own examples for the quiz, if they prefer to use different images.
Debrief: Did you find this activity challenging or easy? What is your main take-away from this activity?
Background reading:
https://www.theguardian.com/technology/2024/apr/08/how-to-tell-if-an-image-is-ai-generated

3.4 Activity 4: Prompt Battle – Who generates the most realistic AI image?

Type: Game / Contest Creativity

Duration: 15 minutes

Learning outcomes: Participants will have an improved understanding of how AI images are being created and how the instructions you give ('prompts') influence the result

Materials/Resources needed:

- Notebooks and pens for participants.
- Input script for facilitator.
- Large screen or projector and a laptop connected to it for facilitator.

Instruction for Facilitator:

Divide the group into two teams. Using the free Microsoft Copilot Designer Software (or paid software e.g. Dall-E or Midjourney), they will compete against each other in generating the most realistic AI image on a given topic.
They will play three rounds. Each round, they can choose or draw a topic from the following list or can invent their own topic. Then, each group gets 3-5 minutes to create a detailed prompt to get a realistic image. Be careful that the teams don't hear or see the prompt of the other team. After entering the prompt, the software will create 4 similar images, from which the teams choose their favorite. When time is up, both groups present the results. The whole group can look at the images and decide which is the winner (depending on which image looks more realistic). After three rounds, there is one winning team that was able to create the most realistic looking prompts.
List of topics:

- A scene of a protest action by climate activists
- a perfume advert with a model on the beach
- A politician in an election campaign, speaking in front of a large crowd (Note that many AI generators do not recreate real personalities, i.e. Donald Trump, to protect against disinformation)
- A devastating wildfire
- Paparazzi chasing celebrities

Tips on how to create a good prompt (can be presented in advance)

- Describe as precisely as possible what you want to see in the picture (subject in the foreground, background, details, what is arranged how...)
- Indicate the style of the picture: photorealistic, painting, black and white, etc.
- Context: Provide background information, data or context for the precise creation of content
- Keywords: Name important keywords or phrases to be included
- Examples: Provide examples of the desired style, structure or content
- Use lighting effects and moods to give your image a certain style, e.g. cinematic lighting, Christmas lighting, or descriptions such as gloomy, cheerful, bright, blurred, etc.
- Important: prompts currently work even better in English

Debrief: In a discussion afterwards, they can look at the different prompts used and try to spot errors in the images and signs for it to be AI-generated.

Like Get Bad News Game, this activity swaps perspectives. By putting yourself in the role of creating a deceptively real image with AI, you learn how quick and easy it is to create realistic-looking images with the right prompts these days. They practice recognizing errors in their own images that are indicative of the AI product. This is an entertaining and playful approach to sensitizing youngsters to take a closer look at images on the Internet.

Possible alterations: Facilitators can also use their own examples for deepfake videos, if they prefer to use different ones.

Possible alterations: Facilitators can also use paid software i.e. Midjourney or Dall-E. Instead of playing in two teams, there can also be only two people competing. In this case, each round, the competing people change. Next to the given topics, it if of course also possible to choose other topics or ideas for AI images.

Anti-RUMOUR
stiftung digitale chancen
CITIZENS IN POWER
challedu
Co-funded by the Erasmus+ Programme of the European Union