EFCSN

How to spot and understand AI-generated or digitally altered content

Accessible and easy-to-use AI models can help people learn and create content, but they can also **amplify the risks dis- and misinformation pose** for open societies and democratic discourse. It is important to prevent our shared information spaces from being cluttered with AI-generated and digitally altered dis- and misinformation.

Part of the solution is new technology like content provenance initiatives and detection software. **But technological solutions are far from perfect and we need the work of independent fact-checkers to provide society with a shared set of verified facts.**

Here is a quick overview what professional, independent fact-checkers do to identify and debunk AI-generated mis- and disinformation and what you can learn from them.

AI-generated content is on the rise

Today, AI-generated mis- and disinformation makes up a small percentage of all claims investigated by professional, independent fact-checkers. Digitally altered content is more prevalent in fact-checkers' work.

But: in an internal survey of EFCSN members, **most fact-checkers agreed that AI-generated and digitally altered content will only increase** in relevance in the future. And recent examples in the context of the European Elections support this prognosis.

UNDERSTAND: *Digitally Altered* refers to any form of content that has been altered significantly to manipulate or change the message it originally conveyed, including edits by AI tools. This does not include edits for clarity or quality.

AI-generated refers to any form of content that has been created by an artificial intelligence system.

Information brought to you by **AI@EUElections,** funded by Meta & managed by the **European Fact-Checking Standards Network** in association with its members and external experts.

Technology moves quickly, but we can't rely on it alone.

AI experts and professional fact-checkers agree: **AI detection tools alone are not enough to detect or debunk AI-generated or digitally altered content.**

Before using a detector, experts recommend developing a familiarity with AI content generators and detectors. With an understanding how models are trained, and a little bit of statistics, fact-checkers can start to recognize the strengths and weaknesses of a tool, and its likelihood of success. **Even still, tools can be a helpful starting point**.

Content provenance initiatives such as the C2PA Specifications can help to certify the source and history of media content, but watermarking and verification is not unimpeachable.

How AI disinformation affects people

> *"Each time you are reacting, if I may say, with your guts, it bypasses your reflection."*
> - Christine Dugoin*

PSYCHOLOGY: Influence operations are often designed to take advantage of psychological biases.

Understanding your own, and your audience's, can help to counter disinformation.

GOALS: Why might a bad actor be relying on AI to create or spread disinformation? What is their intended impact in the real world?

- ➔ Expand their reach into another country or community?
- ➔ Avoid detection or overwhelm fact checkers by generating many variants of similar claims?
- ➔ Influence thoughts or beliefs by establishing credibility via networks of inauthentic accounts?

* *Christine Dugoin is a researcher in informational influence at La Sorbonne.*

Debunking requires a multi-faceted approach and nuanced understanding

So if detection tools don't work, what does? It's important to understand the context of a claim as much as its content. Professional fact-checkers are experts in the necessary investigatory skills. Here are some tips.

> *"Detection tools will never work for 100% - I don't expect they ever will."*
> - Henk van Ess**

CONSIDER THE SOURCE: Can you confirm their identity? What do they talk about and share? Who is interacting with their content? What effect might this content have on readers?

ESTABLISH CREDIBILITY: Independently verify the information with credible sources such as experts with hands-on experience in the field. Does what is depicted make sense based on your knowledge?

Use **FORENSIC MEDIA** techniques to supplement traditional investigative reporting and documentary research. Some techniques include: data scraping, geolocation, biometric recognition, pattern analysis and more.

LEARN & ADAPT: Creators of AI-generated mis- and disinformation are constantly adapting. Adjust your approach to the changing landscape.

SHARE YOUR WORK

Along with a debunked claim, experts recommend providing a transparent analysis and links to sources. This can help readers to follow along with an investigation and understand a nuanced narrative. In some cases, the investigation is more important than if the content is AI-written.

*** Henk van Ess is an expert in OSINT and fact-checking techniques.*

EFCSN

Quick Reference Guide: Dos, Don’ts and Clues

The following are clues that might indicate a piece of content is AI-generated or digitally altered. Together with the other tips mentioned in this guide (context, investigatory techniques and detection tools), they can help you to understand the truth behind what you see.

Text

- ➔ Often (but not always) has **better grammar** than a human.
- ➔ Likely to use **overly formal or structured language**, especially for a social media context.
- ➔ **Excessive adverbs or adjectives**.
- ➔ Lack of human emotion, humor, sarcasm, and idiomatic expressions.
- ➔ May be **lacking in specific details** (names, dates, locations) or original ideas.
- ➔ Most important: are facts stated in the text correct?

Video

- ➔ Don’t use an AI-generated image detector on stills from a video.
- ➔ Look at **facial expressions and movement**, such as blinking, and whether the movement of the mouth matches the audio.
- ➔ Can be marked by **sharp transitions or cuts**.

Audio

- ➔ **Compare suspicious audio to an authentic sample** using tools that can detect differences in speech and breath patterns, intonation…
- ➔ When using detectors, avoid low quality audio samples with static or background noise.
- ➔ May be marked by **unnatural or mechanical speech patterns**, lack of pauses or natural breath.

Images

- ➔ Look for areas with **unnatural details**: perfect skin, blurred backgrounds, unnatural beauty or light, and oddities such as additional fingers.
- ➔ **Look for a watermark** from common image generators.
- ➔ Pay attention to the details of what is depicted: is it logical? Is it appropriate?
- ➔ When using detectors, **opt for a high-resolution or early upload version of an image over one that has been shared and re-shared**.