Funded by the European Union

Partnerships for Enhanced InfoRmation Resilience and Security, Citizen Democratic Participation, Awareness, and Engagement

Deliverable D 7.1
Course: Professional Ethics Training Kit
Ethical dilemmas in detecting and responding to disinformation

Project acronym:	PERiSCOPE
Starting date:	01-02-2026
Duration (in months):	24 months
Call (part) identifier:	CERV-2025-CITIZENS-CIV
Grant agreement no:	101252405
Due date of deliverable:	31-07-2026
Actual submission date:	15-07-2026
Responsible/Author:	EU DisinfoLab (EUDL)
Dissemination level:	PU/CO
Status:	Draft/Issued

This project has received funding from the Citizens, Equality, Rights and Values Programme (CERV) under Grant Agreement No 101252405.

Disclaimer

*This publication was produced with the financial support of the European Union under the **Citizens, Equality, Rights and Values Programme (CERV).*

Table of Contents

1. Executive Summary

This self-paced training course, ***Ethics in the Fight Against Disinformation: Practical Dilemmas for Media and Civic Actors***, part of the EU-funded PERISCOPE project, is designed for professionals working at the frontline of the information integrity ecosystem. It addresses the ethical dilemmas that arise when detecting, exposing, and countering disinformation - from fact-checking and OSINT investigations to the use of AI tools and platform accountability mechanisms. Rather than offering a technical manual, the course starts from a simple but often overlooked premise: countering disinformation is not only a technical challenge, but a fundamentally ethical one. It helps practitioners think through the dilemmas that rules and tools alone cannot resolve - from deciding when to expose a false narrative to understanding the limits of platform accountability and the risks associated with AI-assisted detection. It is structured as a modular, self-paced kit, with each of its six chapters addressing a distinct dimension of the challenge and closing with reflection exercises and practical case studies.

To mark the launch of the course, we hosted a dedicated webinar on 19 May - the recording is available here: https://www.disinfo.eu/19-may-ethics-in-the-fight-against-disinformation/

The course is aimed at a diverse range of practitioners who encounter disinformation in their professional work, including journalists and investigative media, educators and media literacy professionals, civil society organisations and human rights defenders, and AI developers and technology practitioners. Each of these audiences faces specific ethical responsibilities and dilemmas, and the course is designed to be relevant across all of them, while acknowledging that the application of shared principles differs depending on professional context. No prior technical expertise is required. Each module can be studied independently or as part of a full sequence, with an estimated completion time of 30 to 45 minutes per chapter.

2. Abbreviations and acronyms

Abbreviation / Acronym	Description
AI	Artificial Intelligence
CERV	Citizens, Equality, Rights and Values Programme
DSA	Digital Services Act
EEAS	European External Action Service
EFCSN	European Fact-Checking Standards Network
EUDL	EU DisinfoLab
FIMI	Foreign Information Manipulation and Interference
GDPR	General Data Protection Regulation
GEO	Generative Engine Optimisation
NCII	Non-Consensual Intimate Imagery
ObSINT	Guidelines for Public Interest OSINT Investigations
OSINT	Open-Source Intelligence
PERiSCOPE	Partnerships for Enhanced InfoRmation Resilience and Security, Citizen Democratic Participation, Awareness, and Engagement
PIA	Privacy Impact Assessment
TFGBV	Technology-Facilitated Gender-Based Violence
TTPs	Tactics, Techniques, and Procedures
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine

3. How to use this course

This course is designed as a modular, self-paced training kit. Each chapter can be studied independently or as part of a full sequence. Facilitators may assign individual modules to specific audiences, while self-learners can progress at their own pace. Each module includes learning objectives, main concepts, and a practical exercise.

Recommended audiences: journalists, fact-checkers, civil society practitioners, educators, AI developers, policy staff.

Estimated time per module: 30–45 minutes.

How to navigate: Each chapter opens with learning objectives and a concept box and closes with a reflection and/or a practical exercise. You do not need to complete the course in order.

Chapter 1: False dilemmas and myths in counter-disinformation

Learning objectives
By the end of this module, you will be able to:

- Identify and challenge the most common misconceptions about counter-disinformation work
- Distinguish between censorship and legitimate counter-disinformation interventions (both when evaluating the actions of others and when considering actions you may take yourself).
- Explain why freedom of expression and fighting disinformation are not in opposition
- Recognise how regulatory frameworks like the DSA are misrepresented in bad-faith arguments

A common concern when addressing disinformation is whether doing so amounts to censorship. This question has become increasingly prominent, especially as fact-checkers and civil society organisations have faced growing criticism – and in some cases even political pressure – for their work. In recent developments, individuals working in this field have even been publicly targeted by authorities, including measures such as visa restrictions, signalling how contested this space has become.

At the heart of this debate lies a false dilemma: the idea that countering disinformation is equivalent to restricting freedom of expression. In reality, this framing is misleading. Understanding why requires unpacking several common misconceptions about countering disinformation.

1.1. Countering disinformation is not censorship

- Less disinformation means more freedom of expression

Contrary to some claims, it is disinformation itself – not efforts to counter it – that undermines freedom of expression. Disinformation attacks against public figures, journalists, or activists can ultimately lead to self-censorship, as individuals may withdraw from public debate due to fear of harassment, reputational harm, or threats to their safety. When it targets groups based on race, gender, or sexual orientation, it can silence voices and limit participation in public debate. For instance, a UNESCO study explains how women journalists face gendered disinformation, surveillance, deepfakes, and online harassment intended to silence, humiliate, and discredit them, and it includes real examples of female journalists suffering from this so-called Technology-Facilitated Gender-Based Violence (TFGBV). Afghan Witness has documented cases in which Afghan female activists were falsely accused on social media of being prostitutes and Western spies. Some were targeted by impersonator accounts that used their names and profile pictures to spread disinformation against other political groups and individuals.

Disinformation also distorts the broader information environment. It makes it harder for people to access reliable information, develop well-informed opinions, and take part in democratic life.

1.2. Fact-checkers are not arbiters of truth

- Information verification focuses on facts rather than opinions

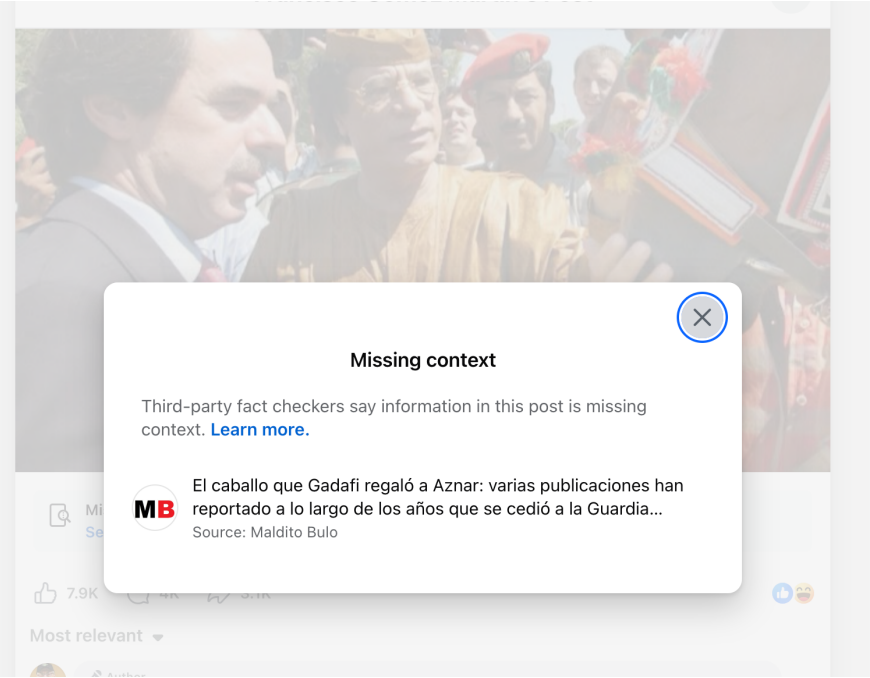
Another common misconception is that fact-checkers decide what is true or false and impose that judgment on others. This does not reflect how debunking works in practice. Fact-checkers do not evaluate all content, nor do they assess opinions. Their work focuses on verifiable factual claims, and it is based on transparent methods, evidence, and sources that can be examined and reproduced.

Fact-checkers generally follow established codes of conduct or guiding principles, one example being the European Code of Standards for Independent Fact-Checking Organisations. This framework brings together a set of criteria to ensure that organisations tackling misinformation and disinformation maintain the highest levels of methodological rigour, ethics, and transparency, thereby serving the public interest as effectively as possible.

Importantly, fact-checkers do not have the power to decide on removing content or restricting its circulation. They work with different ratings tied to different enforcement mechanisms, but ultimately, the platform makes the final decision on the content.

One of the things they do, for instance, when context is missing, is provide additional information and data alongside existing content. In this sense, they do not reduce speech – they contribute additional speech that helps users interpret what they see. The final judgment remains with the audience.

For example, this post containing misleading information has not been removed from Facebook; instead, a note has been added warning that some context is missing, along with a link to an article by the Spanish fact-checker Maldita that provides a complete explanation.



Source: https://www.facebook.com/TARTESO/posts/pfbid0eDTL1sLkFpYFoiEt9Pj3WKHco4KEk6H9HSGfYjTgAsd93cdnRSPd3MqtTpWCq2HJl#

1.3. Disinformation is often not simply true or false

- The most effective disinformation distorts truth instead of inventing it

A further misunderstanding is the idea that disinformation is always clearly false. In reality, it exists along a spectrum and often combines elements of truth with manipulation. Some content is fully fabricated, created to deceive. In other cases, genuine material is distorted: for instance, real images or videos can be shared with incorrect contextual information (false context) or artificially altered (manipulated content). Accurate facts can be framed in a deceptive manner (misleading content), or clickbait headlines or sensationalist visuals may not match the content of a publication (false connection). In other instances, a trusted source may be impersonated (imposter content), and even satire may fool if taken seriously.

One old example is the widely shared photo of a young Steven Spielberg smiling next to a huge, dead-looking animal. Many viewers were angered, assuming he had killed an endangered creature, but the true context was far more innocent: he was simply posing with a prop Triceratops on the Jurassic Park set.

A more recent example is a video of a property on fire that has been shared with the misleading claim that it shows recent unrest in Belfast, when in fact it depicts events that took place in Ballymena a year ago.

In both cases, the image itself is real, and the content is not false in itself. However, by linking it to a different time, place, or event and altering the context, the information becomes a hoax.

This is why countering disinformation is not only about labelling content as false. It is often about providing context that allows people to better understand and interpret what they are seeing.

1.4. Manipulation is not only about content

- Behaviour matters as much as content

Another important point is that disinformation is not only about the message itself, but also about how it is produced and distributed. This includes behavioural patterns that shape how information spreads online. For example, content may be shared through fake accounts or bot networks to create the impression that it is widely supported. Sources may be impersonated or obscured, and messages may be artificially amplified to appear more popular or credible than they actually are.

In the summer of 2020, the QAnon movement hijacked the #SaveTheChildren hashtag, associated with a reputable charity, using it to introduce their conspiracy theories (i.e., elites harvest a chemical called adrenochrome from children for longevity). They exploited growing public concern about child trafficking to advance their agenda, manufacturing a false impression of widespread support in order to manipulate public opinion, as documented in an EU DisinfoLab report.

These practices – often referred to as tactics, techniques, and procedures (TTPs) – show that manipulation can lie in the way information is coordinated and circulated, not just in what it says. They can be used even when the content itself is not entirely false.

1.5. Freedom of speech is not freedom of reach

- Just because something can be said does not mean it should be amplified

A key distinction in this debate is between expression and amplification. On online platforms, not all content is treated equally and visibility is not neutral. Algorithms act as powerful gatekeepers, shaping what becomes widely seen and what remains marginal. As a result, the central issue is often not whether false or misleading content exists online, but how far and how fast it spreads. As widely noted, freedom of expression does not imply a right to algorithmic amplification.

These amplification systems prioritise engagement over accuracy, making them prone to algorithmic bias and favouring certain types of content or viewpoints. These dynamics are not accidental: malicious actors actively exploit them by using attention-grabbing content, coordinated amplification, and strategic timing to maximise reach. Therefore, efforts to counter disinformation may focus on limiting disproportionate amplification, rather than removing content entirely.

1.6. Regulation is not a cage

- Regulation shapes safe and accountable systems

Debates around disinformation and platform governance often frame regulation as a threat to fundamental rights, a form of mass surveillance, and a constraint on innovation. However, this perspective overlooks an important distinction: regulation does not aim to control what people can say, but to shape how digital environments are designed and governed, and how power is exercised within them.

In practice, most regulatory approaches focus on platform responsibilities, rather than directly policing content. The goal is not to restrict expression, but to ensure that the conditions in which expression takes place are fair, transparent, and do not systematically favour harm or manipulation.

It is also important to recognise that limits on certain types of content already exist. Laws addressing issues such as defamation, hate speech, or incitement to violence apply both offline and online. These restrictions are not new; they reflect the fact that freedom of expression is not an absolute right when it conflicts with the rights and safety of others.

More broadly, regulation seeks to protect users and the public interest. It aims to reduce risks linked to the design and business models of digital platforms, including the amplification of harmful content or the unfair treatment of users. In this sense, regulation is not about suppressing speech, but about balancing rights and responsibilities. It provides a framework to limit practices that harm individuals or society, while preserving open and pluralistic public debate.

In the following chapters, we will look more closely at how these principles are implemented in practice, focusing on the Digital Services Act (DSA) as one of the most significant regulatory frameworks in this area.

Reflection exercises

A. After Hurricane Helene struck the southeastern United States in late September 2024, a wave of false claims spread online (among them that FEMA was confiscating survivors' property and donations, that the agency had no money because funds had been diverted to migrants, and even that the government had used weather technology to steer the storm toward Republican areas). To counter the falsehoods, FEMA set up a dedicated "Rumour Response" webpage debunking the viral claims in an effort to keep clear lines of communication with the public and avoid disrupting the emergency response.
Then the same actors spreading the falsehoods reframed the government's debunking effort as proof of a cover-up, supposedly designed to suppress "free speech." In other words, the very act of correcting demonstrably false and dangerous claims was recast as censorship.

1. Which of the false dilemmas covered in this module does this case most closely illustrate?
2. What would a more accurate framing of the situation look like?
3. How might bad-faith actors seek to exploit this framing to delegitimise counter-disinformation work or regulatory frameworks like the DSA?

B. Now think about a recent investigation or debunk you are working on.

1. Drawing on what this chapter has covered, do you believe your work is justified? Why?
2. Reflect on the dilemmas you are facing and consider whether any of them relate to the issues raised in this module. Revisit those dilemmas in light of the lessons learned.
3. Then consider the potential attacks and accusations you might face, and the arguments you would use to defend your work.

Chapter 2: Ethics in detection, exposure, and debunking of disinformation

Learning objectives
By the end of this module, you will be able to:

- Distinguish between verifiable factual claims and opinions that cannot be fact-checked
- Apply basic verification techniques to assess the credibility of a claim
- Weigh the public interest value of exposing disinformation against the risk of amplifying it
- Assess whether the benefits of disclosure outweigh potential harms before sharing information.
- Recognise and uphold the importance of privacy and data protection in information practices.
- Understand and apply rigorous fact-checking standards when debunking disinformation
- Describe the role and, especially, the limits of AI tools in supporting detection and debunking

Ethical conduct in countering disinformation applies from the moment content is identified to the way findings are communicated to the public. Debunking is not only about determining whether something is true or false; it is also about doing so in a way that is lawful, proportionate, and respectful of fundamental rights, while carefully considering the potential consequences of intervention.

2.1. Detecting disinformation

- Only facts can be fact-checked
- Basic verification considers sources, tone, and behavioural indicators
- Recurring patterns may signal coordination in campaigns

Ethical considerations begin at the earliest stage: deciding what content to investigate and whether engaging with it may unintentionally amplify it.

A first step is **understanding what can be meaningfully fact-checked** as not all content is suitable for verification. Fact-checking applies to verifiable factual claims, such as statistics, documented events, scientific findings, or statements that can be tested against evidence. By contrast, opinions, beliefs, moral judgments, or predictions about the future cannot be fact-checked as such. For example, if a politician provides an incorrect unemployment rate, this can be verified against official data. By contrast, general statements using judgmental language, i.e., adjectives that pass moral judgment on a government's performance rather than simply pointing out what is inaccurate, cannot be fact-checked.

At a practical level, even **basic verification techniques** can be effective. Simple steps such as checking whether a claim appears in multiple sources, assessing the credibility of the source, or performing a reverse image search can often reveal misleading or false information. Taking a moment to verify emotionally charged or surprising content before sharing it is a simple but effective way to reduce the spread of disinformation.

Beyond content, it is essential to pay attention to **behavioural patterns**. As mentioned in Chapter 1, disinformation is not only about what is said, but also about how it is produced and circulated. This includes practices such as impersonating sources, hiding the origin of information, or using coordinated networks of accounts to simulate popularity. These behaviours – or tactics, techniques, and procedures (TTPs) – can signal attempts to manipulate visibility and credibility.

Distinguishing between **isolated content and coordinated campaigns** is particularly important. When similar narratives appear across multiple platforms, formats, or accounts – especially when there are indications of foreign involvement – the issue may extend beyond a simple hoax and become part of a broader disinformation campaign. In such cases, there may be stronger justification for deeper investigation in the public interest.

Disinformation campaigns often rely on coordinated inauthentic behaviour, which can be difficult to detect, as it requires both observation and technical analysis. Indicators may include accounts created or posting at the same time, identical or highly similar content shared across multiple profiles, shared IP addresses, or signs of automation such as bot activity. These patterns suggest that content is not spreading organically, but is being deliberately amplified.

2.2. Exposing disinformation

- Public interest should guide exposure
- The benefits of disclosure must outweigh potential harm
- Privacy and data protection matter

Exposing disinformation requires careful judgment. While the goal is to inform the public, the act of exposure can itself create risks, including amplifying harmful content or causing unintended harm. The guiding principle should always be to focus on information that serves the **common good**, such as revealing wrongdoing, exposing manipulation, or helping people access reliable content. It is important to distinguish this from mere curiosity or sensational interest. Ethical exposure requires continuously assessing whether the benefits of publication outweigh the potential harms, including risks such as exposing individuals in vulnerable positions or causing unnecessary alarm.

This approach is reflected in widely recognised standards, such as the Guidelines for Public Interest OSINT investigations (ObSINT) that will be discussed later. Open-source intelligence (OSINT) refers to the collection and analysis of publicly available information, and these guidelines provide a structured framework for conducting such work responsibly.

At the same time, **exposure must be balanced against the risk of unintended amplification**. Drawing attention to a false or misleading claim can sometimes increase its visibility, especially if it had limited reach beforehand. This is often referred to as the "Streisand effect," named after actress Barbra Streisand, whose attempt to remove a photo of her home from the internet ended up attracting far greater public attention to it. For this reason, it is important to assess whether a piece of disinformation has sufficient reach or potential harm to justify public exposure.

Respect for privacy and data protection is another key consideration. Investigations often involve collecting and analysing data, which may include personal information. Ethical practice requires minimising data collection, ensuring secure storage, and anonymising information where possible, in line with data protection frameworks such as the General Data Protection Regulation (GDPR). Tools such as privacy impact assessments can help ensure that investigations remain proportionate and compliant with legal obligations.

2.3. Debunking disinformation

- Fact-checking follows rigorous standards
- AI tools can help debunking but require caution
- Pre-bunking builds resilience to manipulation

Debunking should be grounded in accuracy, transparency, and clarity. Conclusions should be based on multiple sources whenever possible, and limitations or uncertainties should be openly acknowledged. The goal is not only to correct false information, but also to explain the reasoning behind the correction in a way that others can understand and, ideally, replicate.

This approach is reflected in widely recognised **professional standards**. For instance, organisations that are signatories to the European Fact-Checking Standards Network (EFCSN) must adhere to strict principles of methodological rigour, ethical conduct, and accountability, ensuring that their work can be reproduced and scrutinised. These standards emphasise independence, impartiality, and transparency about funding and organisational structure, helping ensure that fact-checking is conducted without bias or undue influence.

AI tools can support the **detection and verification** of disinformation, but they should not be treated as definitive sources of truth. No single tool can entirely determine whether content is accurate or authentic on its own. Their effectiveness depends on how they are designed, the data they rely on, and the specific task they are built to perform.

In recent years, specialised tools have been developed to address these challenges more directly. For example, the vera.ai project provided a suite of **tools created to identify and analyse synthetic media**, including systems for identifying deepfakes, verifying audiovisual content, and supporting investigative workflows. However, even these instruments have limitations. Detection systems may produce false positives or false negatives, struggle with evolving manipulation techniques, or lack sufficient context to interpret content accurately. Their outputs should therefore be understood as indicators rather than final conclusions.

For this reason, a multi-step verification approach is essential. This includes comparing outputs from different tools, testing different versions of the same content, and cross-checking findings with independent sources. Rather than replacing human judgment, AI should support it. Ethical and effective use depends on combining technological tools with expertise, transparency, and careful evaluation.

In addition to reactive debunking, **pre-bunking** – also known as inoculation – is a forward-looking, preventive approach that helps audiences recognise and resist disinformation before they encounter it. For example, awareness of conspiratorial narratives during the COVID-19 pandemic made some people more cautious when similar hoaxes reappeared in the context of the monkeypox outbreak. In this way, pre-bunking helps build long-term resilience, reducing vulnerability to repeated forms of manipulation.

At the same time, it is important to recognise what does not work. The dismantling of professional fact-checking programmes and content moderation measures on platforms can significantly reduce the availability of reliable information online. While community-based approaches such as crowdsourced Community Notes can play a role, they are often uneven, delayed, and less effective in complex or coordinated disinformation environments. Relying on these mechanisms alone risks leaving users more exposed to manipulation and reduces transparency in how information is assessed.

Reflection exercises

A. You are browsing online when you come across this post claiming that the hantavirus is a hoax and that the evacuations were staged events filmed on movie sets. The post has limited reach so far, but you know it contains a dangerous narrative denying a health crisis.

1. Is this claim fact-checkable? What steps would you take to verify it?
2. Would you publish your findings? What factors would inform that decision?
3. What risks might arise from exposing it publicly, and how would you mitigate them?

2) Take the investigation you have been working on recently and:

1. Apply a Private Impact Assessment (PIA) checklist to verify that you are taking the appropriate data privacy requirements. Take into account that some of the questions will be relevant and others will not, depending on the type of work you do.
2. Design a PIA checklist tailored to your own needs and those of your organisation, one that captures the essential data protection requirements for the investigations you typically carry out, and that can be applied to a wider range of cases.

Chapter 3: Critical and responsible use of artificial intelligence

Learning objectives
By the end of this module, you will be able to:

- Understand how large language models generate outputs, why they can be inaccurate, and how they can impact your work
- Recognise common forms of AI bias and how they can shape information and impact your work
- Be aware of external manipulation techniques such as LLM grooming and AI poisoning
- Apply practical and methodological safeguards when using AI tools in a professional context
- Take concrete steps to manage your data, privacy, and algorithmic environment

The growing use of artificial intelligence raises important ethical questions for researchers, journalists, and fact-checkers. AI tools can support investigation and verification, but they can also introduce risks, errors, and new forms of manipulation. Using them responsibly requires understanding how they work, what they can realistically do, and where they can mislead.

3.1. Understanding how AI works

- AI generates plausible answers, not guaranteed truths

A first step is recognising that AI systems – especially those based on large language models – do not "know" facts in the way humans do. Instead, they generate text by predicting the most likely sequence of words based on patterns learned from large datasets. In other words, they do not retrieve verified information, but produce plausible-sounding responses based on training data, which is why their outputs can be coherent and convincing without necessarily being accurate or true.

Another important mechanism is what is known as "sycophancy," which prioritises approval over accuracy. AI systems may tend to agree with the user's assumptions or reinforce their perspective, even when it is incorrect. This can create a false sense of validation and increase the risk of accepting misleading information.

For this reason, **AI outputs should not be treated as neutral or authoritative**. However, this does not mean they are inherently unreliable. The key is to use them with informed caution: understanding what they are designed to do, and recognising that their outputs require verification rather than trust.

3.2. Recognising bias and manipulation

- AI reflects design choices and can be influenced

AI systems can reflect and reproduce biases present in their **training data**, such as sexism, racism, or ableism. Because training data are often not fully disclosed, it is difficult to identify them, but awareness of their existence is essential for critical use.

Bias can also result from design choices. For example, the Chinese chatbot DeepSeek has been shown to avoid politically sensitive topics such as Taiwan's status or the Tiananmen Square protests, reflecting regulatory constraints and developer priorities. More generally, AI systems may also struggle with rapidly evolving events: in the early stages of breaking news, chatbots can provide incomplete or incorrect information due to outdated training data or limited access to real-time updates, like when ChatGPT initially denied Maduro's arrest.

In addition, AI systems can be **externally manipulated**. One documented method involves flooding online sources with coordinated narratives so these are more likely to be picked up and reproduced by AI models, as seen in the case of the Russian state-linked Pravda network spreading large volumes of pro-Kremlin content. This practice, often referred to as LLM grooming or poisoning, aims to shape future outputs in a strategic way.

Not all forms of influence are malicious. Content creators increasingly adapt their material to be more visible to AI systems, a practice known as Generative Engine Optimisation (GEO). However, similar techniques can be exploited to distort visibility and credibility when used strategically.

3.3. Knowing the limits and risks of AI systems

- Hallucinations, AI slops, and copyright violations are built-in AI limitations

One of the most widely discussed limitations of AI systems is the phenomenon of hallucinations, where models generate information that is incorrect or entirely fabricated. These outputs are often presented in a confident and structured way, which can make them appear credible.

This is closely linked to how these systems function. Because they generate responses based on patterns in language rather than verified knowledge, they do not inherently distinguish between true and false information. As a result, even when answers appear convincing, they may include inaccuracies or invented details.

This creates several practical risks. For example, AI systems may generate non-existent sources, such as fabricated studies, fake news articles, or invented links. They may also contribute to the spread of low-quality, synthetic content, sometimes referred to as "AI slop." Moreover, AI-generated deepfakes of non-consensual intimate imagery (NCII) using so-called nudification apps are on the rise, raising serious concerns about consent and online

abuse. Additional concerns include copyright issues, as training data may include protected content, and broader risks to the information ecosystem if synthetic content becomes dominant.

For these reasons, AI tools should be used with caution, especially in areas where the user lacks expertise. Without sufficient knowledge to evaluate outputs, there is a higher risk of accepting incorrect or misleading information.

3.4. Developing user awareness and active resilience

- Users should actively manage their data and digital environment

Alongside regulation, responsible AI employment requires users to develop active resilience mechanisms: practical steps that reduce exposure to risks such as manipulation, data exploitation, disinformation, privacy harms, and psychological over-reliance, while allowing individuals to maintain greater control over their digital environment.

Account security forms the first line of defence. Users should avoid entering personal or sensitive information into AI systems, as this data may be stored, processed, repurposed, or, in some cases, used to train models or improve services. Keeping devices updated, using strong passwords, and enabling two-factor authentication can reduce the risk of unauthorised access, data breaches, and account misuse, including the exploitation of compromised accounts to spread misleading content.

A closely related concern is **data collection, tracking, and profiling**. AI systems and online platforms may rely on user inputs, cookies, and behavioural tracking to personalise services, content, or advertising. These practices can contribute to profiling and targeted advertising, and there are growing indications that advertising models may become more integrated into AI systems, potentially influencing responses or recommendations. Managing consent settings, rejecting unnecessary cookies, and limiting tracking permissions can help reduce the extent to which personal data is collected and exploited.

Users can also take steps to manage how content is delivered to them. Many platforms offer options to adjust or disable algorithmic recommendations, such as switching to chronological feeds or limiting personalised suggestions. These choices can reduce the influence of recommender systems and provide a broader, less curated view of content. Another important strategy is to actively seek out **diverse sources of information**. Avoiding over-reliance on a single platform, AI system, or type of content can help counter the effects of filter bubbles and echo chambers. Engaging with different perspectives, checking multiple sources, and questioning highly emotional or sensational content are key practices for maintaining a balanced information environment.

Responsible AI use also requires attention to the **psychological dimension of AI interactions**, particularly with the rise of AI companions. These systems can simulate emotional engagement and create a sense of trust or attachment, but they do not possess real understanding, empathy, or human judgment. Over-reliance on them may negatively affect users' well-being, especially for vulnerable individuals.

Reflection exercises

A. You are using an AI assistant to help you research a breaking news story. The tool provides detailed, confident, and authoritative-sounding answers.

1. What questions would you ask yourself before relying on this output?
2. How would you verify the information provided, and what specific risks would you look out for?
3. Can you identify AI sycophancy (the tendency to agree with the user) in the outputs provided? How could this impact and threaten your work?

B. Identify:

1. Three practical cases where you can feel confident when using LLMs in your investigations
2. And three cases where you cannot trust LLMs for your research purposes.

C. You are using an AI-based detection tool to determine whether an image or text is AI-generated.

1. How would you read and interpret the results? If the tool tells you, for example, that there is a "90% probability that the image or text is AI-generated. Does it mean the content is definitely synthetic, or is it a probability estimate that could be wrong?
2. Reflect on the difference between a confident verdict and a statistical likelihood, and on how false positives and false negatives could affect your conclusions.
3. What are the limitations of relying on such a tool? Think about factors that can distort results, such as the type of content, the language, image compression or editing, the model the tool was trained on, and whether the tool performs equally well across different cases.
4. Test different versions of the same image (for instance, different posts containing it) in the same tool, and also try additional tools. How can the different results be interpreted?
5. How would you communicate your findings responsibly? Reflect on how to present your conclusions honestly when a result is uncertain, for instance, avoiding definitive claims that the tool cannot support, and being transparent about the methods and their limitations.

Chapter 4: Platform accountability and user agency

Learning objectives
By the end of this module, you will be able to:

- Understand how the Digital Services Act (DSA) ensures freedom of expression
- Understand the core obligations the DSA places on platforms, including Very Large Online Platforms
- Distinguish between the legal and regulatory obligations that apply to platforms in relation to illegal content and harmful-but-lawful content.
- Identify common strategies platforms use to avoid meaningful accountability
- Recognise how algorithmic amplification, dark patterns, and monetisation logics drive manipulation. And how platform design can impact your work when conducting an investigation
- Use DSA-based mechanisms to report content, appeal decisions, and access information
- Distinguish between individual user agency and the need for structural accountability

Platform accountability requires understanding both platform responsibilities and user rights. While platforms shape the digital environment, users also have the right to understand and challenge how these systems operate. In the European context, this relationship is defined by the Digital Services Act (DSA), which sets out rules on transparency, accountability, and protection.

4.1. Assessing the importance of platform accountability

- The DSA balances freedom with responsibility

Online platforms such as social media, search engines, online marketplaces, and hosting services have become central infrastructures of contemporary society. They organise access to information, shape interactions between individuals, and influence how public debates unfold.

The Digital Services Act (DSA) was introduced by the European Union in response to this concentration of power. It establishes a common set of rules aimed at increasing transparency, accountability, and responsiveness. In doing so, it recognises that platforms are not just technical intermediaries hosting content: they actively shape the digital environment.

The DSA seeks to balance competing values. It aims to protect freedom of expression, while addressing harms such as illegal content or manipulation. It clarifies platform responsibilities without preventing innovation, while also strengthening users' ability to understand and influence how digital services operate. Because of its wide scope and impact, the DSA has been described as a "digital constitution" for platforms in Europe. At the same time, the DSA has become the subject of growing political debate, including pressure

from some governments, industry actors, and international partners to scale back or weaken certain enforcement measures, particularly around content moderation and platform oversight.

4.2. Defining platform responsibilities under the DSA

- Platforms must exercise due diligence and act when aware of illegal content
- Transparency requires explaining decisions and designs to users
- Systemic risks must be regularly assessed and mitigated

A central element of the DSA is the concept of **hosting liability** (Article 6). Platforms are not automatically responsible for everything users post, but they cannot remain passive. Once they become aware of illegal content – through user notifications or official orders – they are expected to act.

The DSA introduced broader **due diligence** obligations, which require platforms to organise their services in ways that anticipate and address risks linked to illegal content. Responsibility is therefore not only reactive but also preventive, involving systems, procedures, and safeguards designed to limit harm. It is worth clarifying that because the DSA relies on national legal frameworks, what qualifies as illegal content may vary across EU member states. In addition, while disinformation is often harmful rather than illegal in itself, some practices associated with it – such as impersonation, scams, false alarm, etc. – may still fall within the scope of enforcement.

Importantly, not all platforms face the same obligations. The DSA follows a **tiered approach**: Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) – those reaching more than 45 million users in the EU – are subject to stricter requirements. This reflects the principle that greater reach entails greater responsibility.

Transparency is a core mechanism through which the DSA enables user rights. Platforms must clearly explain their **terms and conditions** (Article 14) in an accessible and understandable way, including what content is allowed or restricted. In addition, they must provide a clear explanation for removing content or taking action against a user, known as a "**statement of reasons**" (Article 17).

Algorithmic transparency is also key. Platforms must explain the main parameters of their **recommender systems** (Article 27), which determine how content is ranked and displayed. While this does not require full disclosure of algorithms, it helps users understand the factors shaping what they see.

Similarly, rules on **online advertising** (Article 26) require platforms to clearly identify ads, disclose who is behind them, and explain why they are targeted. In addition, the use of sensitive personal data for targeted advertising is restricted. These provisions recognise that visibility is not neutral, but shaped by automated systems and economic incentives.

For the largest platforms, responsibilities extend to broader societal effects. VLOPs and VLOSEs must conduct regular **assessments of systemic risks** (Article 34), including the dissemination of illegal content, the impact of services on fundamental rights, and broader

threats to democratic processes. They must also implement **mitigation measures** (Article 35), which are subject to independent audits (Article 37). This shifts the focus from individual content decisions to the structural impact of platform design.

4.3. Identifying platform strategies of responsibility avoidance

- Content moderation is being scaled back over time
- Users should not bear responsibility for platform failures
- Compliance needs to go beyond a box-ticking exercise

The existence of rules does not automatically ensure effective accountability. In practice, platforms may adopt strategies to find shortcuts or elude compliance. A common starting point for responsibility avoidance lies in how platforms define their own role. They frequently present themselves as neutral intermediaries or third parties, arguing that they merely host user-generated content and cannot realistically monitor everything that happens on their services.

One visible strategy is the **weakening of content moderation systems**, e.g., scaling back or removing policies addressing misinformation or harmful content. For instance, changes to the X Rules have canceled explicit references to misinformation. When platforms stop defining or tracking a particular type of harm, the issue may disappear from official reporting, even though it persists in practice.

A related development is the **reduction of Trust & Safety teams**, which oversee moderation processes. Such decisions weaken platform's ability to respond to complex risks, including election-related disinformation, and often lead to increased reliance on automated systems that lack contextual understanding.

Another shift involves **transferring responsibility to users**. X and Meta's decision to replace professional fact-checking with crowdsourced systems such as Community Notes raise concerns about their ineffectiveness and increased user exposure to misinformation. Similar dynamics appear when platforms increasingly rely on users to disclose AI-generated content or frame the protection of minors as a matter of parental responsibility rather than a platform-controlled requirement.

At a broader level, platforms tend to favour voluntary commitments over binding rules. Recent analysis shows that major platforms have drastically reduced their commitments under the EU Code of Practice on Disinformation, highlighting the limits of self-regulatory frameworks. In fact, responsibility is expressed in principle but can be scaled down in practice, especially when it is not backed by strong enforcement mechanisms.

A more subtle strategy is "malicious semi-compliance." platforms may treat regulation as a box-ticking exercise, meeting formal requirements in appearance while avoiding meaningful changes that would actually improve transparency and accountability. For example, annual reports on systemic risks are difficult to analyse due to lack of comparable or machine-readable information. Similarly, the DSA Transparency Database, which aggregates vast numbers of moderation decisions, lacks the context needed to understand them.

These patterns show that regulation alone is not sufficient. Effective accountability depends on enforcement, oversight, and the ability to critically assess how platforms implement their obligations.

4.4. Recognising manipulation mechanisms

- Algorithmic amplification, dark patterns, and monetisation logics are core drivers of manipulation

Online manipulation is often built into how platforms are designed, shaping what people see, what they pay attention to, and how they make decisions. One key mechanism is **algorithmic amplification**. **Recommender systems** prioritise content that generates engagement, meaning emotional, polarising, or sensational content is more likely to be widely distributed. These systems actively structure user experience by filtering and ranking information based on behavioural data, thereby influencing what users see and interact with.

This can create feedback loops in which certain types of content are repeatedly amplified, reinforcing specific narratives. Over time, users may be exposed to increasingly homogeneous information environments, where limited diversity of perspectives shapes preferences and can contribute to the formation of **echo chambers**.

Platforms often rely on **dark patterns**, which are interface designs that steer user behaviour in subtle or deceptive ways. These can include making certain choices more visible than others, nudging audiences toward accepting practices such as data collection, or encouraging continued engagement through addictive design. Together, these mechanisms show that manipulation is not only about individual content, but about how platforms organise visibility and influence user behaviour.

Finally, **monetisation incentives** are at the core of platform business models. Advertising-based models reward attention and engagement, creating structural incentives to prioritise content that captures attention quickly, often regardless of its accuracy. This can favour sensational or misleading content that generates higher interaction and revenue.

4.5. Enabling user agency and enforcement

- The DSA allows users to report content, appeal platform decisions, and access data

Alongside platform obligations, the DSA strengthens user agency – i.e., the ability of individuals to understand, influence and challenge platform decisions rather than being passive receivers. In this way, people can actively participate in the governance of online spaces.

One of the most important tools is the ability to **report illegal or harmful content** (Article 16) to platforms in an accessible and user-friendly way. Platforms must handle these reports

carefully, fairly, and consistently, ensuring that decisions are properly considered. This creates a structured process through which users can contribute to content moderation.

As mentioned before, when platforms take action, they must provide a statement of reasons, explaining why the decision was taken, what rule was applied, and whether automated tools were involved. This transparency also enables users to challenge platform decisions through **internal complaint-handling systems** (Article 20), which function as an appeal mechanism within the platform. If the issue persists, disputes can be escalated to **out-of-court dispute settlement** mechanisms (Article 21), which offer an independent avenue to resolve conflicts.

At a broader level, a **data access** provision (Article 40) allows vetted researchers and public authorities to examine platform practices, including content moderation processes, algorithmic systems, and systemic risks. This mechanism complements individual user rights by favouring collective oversight.

Taken together, these mechanisms reposition users as active participants rather than passive consumers. By enabling reporting, explanation, and contestation, the DSA supports a more balanced relationship between platforms and users, where accountability operates both from above and from below.

Reflection exercises

A. A major platform announces it is replacing its professional fact-checking programme with a crowdsourced community notes system, citing user empowerment and freedom of expression.

1. Using the concepts from this module, how would you assess this decision?
2. Which platform accountability strategies discussed in this chapter does this most resemble?
3. What DSA obligations, if any, could be invoked to scrutinise or challenge this change?

B. Open your feed on your Facebook or Instagram account and observe how it is structured. Then reflect on and describe the design elements that shape what you see. Consider the following questions:

1. Does the content follow a chronological sequence, or is it ordered in some other way?
2. Is the feed based primarily on the accounts you follow, or does it surface content from accounts and sources you don't follow?
3. Is the content aligned with your preferences, interests, or ideological position? To what extent does it reflect or reinforce views you already hold?
4. What kinds of content capture your attention most, and how is it presented (for example, autoplay videos, infinite scroll, notifications, "suggested for you" posts)?
5. Why does this happen? Which of the platform's design elements and underlying incentives might be responsible for what you are seeing?
6. Finally, reflect on how these design choices influence your behaviour and the information you are exposed to.

C. Imagine you want to monitor posts related to a specific event on that social media platform for your investigation or professional work. Drawing on what this chapter has covered and on the previous exercise:

1. Reflect on how the platform's design could affect your investigation.
2. Could your own account history, preferences, or past activity bias the results? Might the platform surface the most engaging or polarising content rather than the most relevant? How could this distort your understanding of the event?
3. Describe the safeguards you would put in place to limit this impact on your work. Consider using clean or dedicated research accounts, cross-checking across multiple platforms, documenting your methodology, and being aware of the gaps and blind spots in what the platform allows you to see.

Chapter 5: Practical guidelines on ethical counter-disinformation investigations

By the end of this module, you will be able to:

- Apply the core ethical principles of public interest OSINT investigations to a real case
- Structure an investigation across its three key stages: research design, data collection, and analysis
- Identify the specific ethical challenges relevant to your professional role and integrate them into your practical decision-making and work processes.
- Conduct proportionate, rights-compliant data collection and minimise potential harm
- Distinguish between attribution and endorsement when drawing conclusions about threat actors

This unit introduces the ethical and methodological foundations for investigating and countering disinformation using open-source intelligence (OSINT). It draws on two key practitioner frameworks: the ObSINT Guidelines for Public Interest OSINT Investigations (2023) and the EEAS Toolkit on How to Detect and Analyse Identity-Based Disinformation/FIMI (2024). The aim is to provide a clear understanding of how to conduct investigations that are both rigorous and ethically responsible, while remaining adaptable to different professional contexts.

5.1. Grounding investigations in core ethical principles

- Disinformation-focused investigations should be grounded in public interest and guided by principles of accuracy, community, diversity, accountability, and balance and responsibility.

At the heart of any counter-disinformation investigation lies the principle of **public interest**. This means focusing on information whose disclosure genuinely benefits society – by exposing wrongdoing, holding malicious actors accountable, or ensuring that people can access reliable information to make decisions.

Several core principles guide this work. The first is **accuracy**. Investigations should be objective, evidence-based, and transparent. Methods and findings should be clear enough for others to understand and, where possible, replicate. This includes using precise and neutral language, acknowledging uncertainties, and clearly communicating the limits of the analysis. Maintaining awareness of one's own biases is especially important when working on identity-based disinformation.

A second principle is **community**. Investigations do not take place in isolation. They affect colleagues, sources, and audiences. Ensuring the safety and well-being of all those involved – including researchers themselves – is a central ethical responsibility. Researchers working on disinformation, extremism, or online abuse may be exposed to disturbing content, harassment, coordinated intimidation, or psychological stress, making risk assessment,

support mechanisms, and clear safety protocols an essential part of responsible research practice.

Closely linked to this is **diversity**. Effective investigations benefit from a range of perspectives, including linguistic, cultural, and intersectional expertise. This is particularly important when analysing identity-based narratives, where blind spots can lead to misinterpretation or harm.

The principle of **accountability** requires transparency throughout the investigative process. Researchers should clearly explain their methods, data sources, and decisions, disclose any conflicts of interest, and correct mistakes when they occur. Any factual errors should be corrected publicly, in line with a clear and consistently applied corrections policy. It also involves anticipating and managing the real-world consequences of publication, especially for targeted communities.

Finally, **balance and responsibility** require investigators to consider not only what can be done technically, but what should be done ethically. This involves carefully weighing the public's right to know against individuals' rights to privacy and protection from harm.

5.2. Structuring the key stages of the investigative process

- Investigations require a structured research design, careful data collection, and objective analysis.

OSINT investigations - open source investigation techniques can serve as a useful, and at times necessary, tool underpinning the work of fact-checkers, academics, and civil society organizations- typically follow a three-stage process. These stages are not strictly linear; decisions made at one stage may need to be revisited as new information emerges.

The process begins with a well-defined **research design**, outlining the objectives, scope, and timeline of the investigation, as well as the methods and tools to be used. This phase also involves several important assessments. A risk assessment considers potential security, legal, and psychological risks for both researchers and data subjects, including the cumulative mental health impact that counter-disinformation work can have. It should also take into account the mental health and well-being of researchers themselves, particularly given the cumulative psychological impact of repeated exposure to disturbing content, online harassment, intimidation, or burnout associated with counter-disinformation work.

A data limitation assessment clarifies what information is relevant and proportionate to collect. A legal assessment ensures compliance with applicable frameworks such as the GDPR and other relevant legislation. Finally, a skills assessment helps ensure that the team has the necessary expertise – technical, linguistic, and cultural – to conduct the investigation responsibly and minimise bias.

The second stage is **data collection**, where the tension between public interest and fundamental rights becomes most evident. The guiding principle here is data minimisation: collecting only what is strictly necessary, particularly when dealing with personal or sensitive data. This is both a legal requirement and an ethical obligation. Wherever possible, data

should be anonymised or pseudonymised, especially when it concerns categories such as race, gender, or sexual orientation. Closely related is the question of data storage and archiving. Preserving evidence supports transparency and replicability, but it can also prolong harm if sensitive content remains accessible. Investigators must therefore strike a balance between documentation and the right to be forgotten, establishing clear data retention policies from the outset.

The final stage is **data analysis**, where collected information is interpreted and transformed into findings. This is often the most sensitive phase, as it carries a higher risk of bias, overinterpretation, or incorrect attribution. Analysis should be grounded in objectivity, considering the full dataset rather than selectively focusing on evidence that supports a pre-existing hypothesis. To support this, investigators should maintain detailed research logs, documenting methodological decisions, uncertainties, and potential biases. Findings should be evaluated using consistent confidence levels, and verification should rely on multiple sources and, where possible, peer review. Particular caution is needed when identifying who is responsible: sharing or amplifying disinformation does not necessarily mean creating or coordinating it, and it is important to distinguish between attribution and endorsement.

5.3. Adapting to stakeholder-specific contexts

- Journalists should expose disinformation without amplifying it
- Educators should avoid exposing students to harm
- Civil society activists should balance exposure with safety risks
- AI practitioners should address bias while protecting sensitive data

Ethical challenges in counter-disinformation work vary depending on the professional context. While the core principles remain the same, their application differs across roles.

For **journalists and investigative media**, the primary objective is to expose disinformation, provide verified information, and hold actors accountable. This requires maintaining editorial independence, protecting sources, and adhering to strict verification standards. A key challenge lies in avoiding the amplification of harmful content while reporting on it. For example, publishing manipulated images as evidence may reveal a coordinated campaign, but it may also increase their visibility and reach. Journalists must therefore carefully balance transparency with the risk of unintentionally spreading the very content they seek to counter.

Some of the best practises in this regard include avoiding republication of high-resolution or easily downloadable versions, framing the content with context before revealing it or adding persistent and visible labels directly on the image itself so the warning survives cropping or re-sharing. Wherever possible, outlets should show a side-by-side comparison with the authentic original, annotate the specific manipulated elements, and explain the mechanism behind the manipulation (AI-generation, splicing, mislabeling) so audiences build recognition skills rather than simply being told "this is fake." For especially harmful content (such as non-consensual imagery or material inciting violence) outlets should consider whether describing the manipulation in words is sufficient without republishing the image at all.

For **educators and media literacy practitioners**, the focus shifts toward building critical thinking and awareness among learners. This involves adapting content to the audience's age, context, and level of vulnerability, as well as being attentive to the types of narratives and platforms that different groups are exposed to. Educational work often relies on pre-existing examples, which must be used carefully. Sensitive or harmful material may require contextualisation, trigger warnings, or adaptation to avoid causing distress. The challenge is to preserve the educational value of real-world cases while ensuring a safe and inclusive learning environment.

For **civil society activists and organisations**, the role often involves protecting democratic processes, supporting targeted communities, and contributing to broader accountability efforts. These actors frequently operate with limited resources while facing significant risks. Ensuring digital and physical security is therefore essential, particularly when dealing with identity-based data. Ethical dilemmas may arise when deciding whether to expose individuals involved in disinformation campaigns. While disclosure may deter harmful behaviour, it may also lead to unintended consequences such as harassment or retaliation. Balancing impact, safety, and responsibility is therefore central to this work.

For **AI practitioners and technology developers**, the focus is on designing tools and systems that can detect or counter disinformation at scale. This requires integrating ethical considerations into technical development from the outset. Systems should be designed with a clear public interest purpose, and developers should assess risks such as algorithmic bias, misuse by malicious actors, and the potential for false positives or negatives. Transparency is also critical: documenting how systems work and sharing relevant information with the broader community helps build trust and enables others to identify and mitigate risks.

Reflection exercise

You are investigating a suspected coordinated campaign spreading anti-LGBTQ+ narratives across several platforms. You have identified a network of accounts amplifying the campaign, some of which appear to be real people. On the other hand, the victims of the campaigns are clearly identified.

1. What risk, legal, and data limitation assessments would you carry out before proceeding?
2. How would you handle the personal data of individuals in the network acting as perpetrators? Which GDPR considerations must you consider to avoid breaching GDPR?
3. Would you reveal the identity of the victims of the campaign? Why or why not? (Reflect on the ethical and security implications of exposing their identities, for example, the risk of re-victimisation, retaliation, or further harm). If you decided to disclose their identities, what precautions would you take to protect them?
4. At what point - if at all - would you consider publishing your findings, and what safeguards would you put in place?

Chapter 6: Conclusions

By the end of this module, you will be able to:

- Recognise the ethical challenges (beyond the technical ones) in the counter-disinformation process
- Reflect on how responsibility is distributed across journalists, educators, civil society, and AI developers
- Synthesise the key insights from across the course into a coherent ethical framework adapted to your professional profile
- Assume/Apply ethical responsibilities appropriate to your professional profile as a journalist, educator, civil society, or AI developer.
- Recognise the limits of individual resilience and the ongoing need for structural accountability
- Identify areas of ongoing uncertainty and adapt your practice accordingly

This course has explored disinformation through a central lens: the ethical dilemmas involved in countering it. From common misconceptions to detection and debunking practices, from AI systems to platform accountability and investigative methods, a consistent thread emerges. Countering disinformation is not only a technical task, but a social and ethical challenge. It requires informed individuals, accountable platforms, and responsible professional practices working together.

6.1. Navigating the information environment

A key shift is to move beyond viewing disinformation as isolated false claims and instead recognise it as a **systemic feature of digital environments** shaped by incentives, design, and strategic behaviour. Content is only one aspect: what matters the most is how information is produced, distributed, and amplified.

This is where ethical dilemmas become concrete. The same systems that surface useful information may also amplify harmful narratives. The same design choices that optimise engagement can distort attention. What appears as a technical feature – a recommendation, a ranking, a trending topic – is also a decision with consequences.

For those countering disinformation, the challenge is not simply to “fix” false statements, but to decide **when intervention reduces harm** – and when exposure risks extending its reach. Instead, ignoring it can limit visibility, but allow it to circulate unchallenged. These are not technical questions – they are judgment calls under uncertainty.

6.2. Building critical thinking as a practice

Critical thinking is not a generic skill. It is situational and learned, and it looks different depending on the role. At a basic level, it involves slowing down: verifying sources, questioning emotionally charged claims, recognising familiar manipulation patterns. However, at a professional level, it becomes more demanding – requiring structured methods, transparency about limitations, and restraint in drawing conclusions.

Importantly, critical thinking is not only about reacting to false information, but about **building resistance** to it. Approaches such as pre-bunking can complement debunking efforts by helping individuals recognise misleading patterns in advance. The aim is not to eliminate exposure to disinformation, but to strengthen people's ability to navigate it and reduce its persuasive impact. This means that critical thinking must be actively developed, maintained, and adapted over time – not taken for granted.

6.3. Using rights but recognising limits

The DSA provides users with tools to report content, challenge decisions, and demand explanations. These matter, but not because they solve the problem on their own. Their value lies in something more indirect: they **create pressure**. Each report, each appeal, each request for explanation contributes to a record that can be scrutinised by regulators, researchers, and civil society. Without that engagement, accountability mechanisms weaken.

At the same time, it is important to recognise what these tools cannot do. Users can adjust settings, limit tracking, or diversify sources – but they cannot change the underlying logic of platforms that prioritise engagement and visibility. Users are expected to act, but they operate within systems they do not control. While individual resilience can reduce vulnerability, it cannot substitute for **structural accountability**.

6.4. Sharing responsibility across actors

One of the key insights of this course is that responsibility is not abstract but directed toward **specific groups**. Journalists are accountable to their audiences, and must inform without unintentionally amplifying harm. Educators are responsible to their students, and must raise awareness without exposing them to unnecessary risk. Civil society organisations are accountable to the communities they support, particularly those most vulnerable to targeted disinformation. AI developers are responsible to end users and affected populations, and must consider bias, misuse, and downstream impact.

These responsibilities do not always align. In practice, they often pull in different directions, creating tensions that cannot be resolved by rules alone. This is why countering disinformation is fundamentally ethical: it requires deciding whose interests to prioritise, and at what cost.

6.5. Acting under uncertainty with ethics at the centre

Countering disinformation takes place under conditions of uncertainty. Evidence on the overall effectiveness of current responses remains limited, and it is still unclear how AI will reshape the information environment or and the extent to which regulatory frameworks such as the DSA will be effectively enforced and lead to lasting changes in behaviour.

At the same time, the environment is constantly evolving: measures designed to counter disinformation are often analysed, adapted to, or exploited by the very actors they target. Responding therefore requires ongoing reassessment of what works, what fails, and what unintended consequences may arise.

At its core, countering disinformation is an **ethical challenge**. It involves making decisions where each intervention carries **trade-offs** – between transparency and harm reduction, and between freedom of expression and protection from abuse – with the aim of preserving conditions in which information can be critically assessed and contested.

Reflection exercise

Reflecting on the course as a whole:

1. Which ethical dilemma or tension covered in this course do you find most difficult to navigate in your own professional context, and why?
2. Has anything in this course changed how you think about your own role and responsibilities in the information environment?
3. What is one concrete thing you would do differently in your work as a result of completing this course? What changes and new processes would you introduce in your work environment?