FACT-CHECKING TOOLKIT

Vilnius,
November 2025

A consolidated guide to methodologies, tools, and resources for effective fact-checking

ATEITININKŲ FEDERACIJA, LITHUANIA
DECIDE PROJECT

Funded by the European Union

Table of Contents

u ime obitelji
World Youth Alliance Southeast Europe
PLATTFORM CHRISTDEMOKRATIE
ATEITININKAI

Funded by the European Union

Introduction

About Fact-Checking Ambassadors Programme

Fact-Checking is the process of verifying whether a claim, image, statistic, or story is accurate. It means tracing information back to credible primary sources, checking context, and clearly labeling what's true, false, misleading, or unproven—with transparent evidence.

Why does it matter so much today?

Info overload: We're exposed to thousands of posts daily; falsehoods spread faster than corrections.

AI-generated content: Deepfakes and synthetic text make fakes look convincing.

High-stakes decisions: Bad info can sway elections, health choices, and public safety.

Trust & accountability: Rigorous verification rebuilds trust in institutions and media.

Civic resilience: A fact-literate public is harder to manipulate.

The Fact-Checking Ambassadors Programme was created to strengthen media literacy, critical thinking, and civic resilience across Europe. It offers a structured yet adaptable framework to help participants recognise, verify, and responsibly communicate accurate information. The methodology identifies nine guidelines and realizes them through four integrated implementation parts.

This material was compiled using material presented during a live event - Media Literacy Workshop: How to recognise and Combat Disinformation - held in Vilnius, Lithuania, on October 10–12, 2025 and supplemented with public information sources based on the lecturers' recommendations. AI tools were used to process the information.

We thank Prof. Dr. Darius Pilkynas and Assoc. Prof. Liutauras Ulevičius from the Faculty of Communication of Vilnius University, Jovita Tautkevičiūtė-Kalinauskienė from the Debunk.org project for the theoretical material provided, and Arnas Jasinskas from the Federation of Futures for conducting a lively seminar on disinformation, whose insights were also useful in preparing this toolkit.

Funded by the European Union

Vision

The main vision of Fact-Checking Ambassadors Programme is to equip trusted community members to detect, verify, and responsibly communicate facts—so that communities become more resilient to mis/disinformation and can make better decisions.

Programme objectives

Knowledge: Participants can define mis/disinformation/propaganda; describe common tactics; explain ABCDE and OSINT basics.

Skills: Participants can run a 5–10 minute triage, verify images/videos/text, and document a transparent reasoning chain.

Attitudes: Participants engage respectfully, avoid confrontation, and prioritize clarity, empathy, and accuracy.

Practice: Each participant produces one short fact-check and one outreach action plan tailored to their community.

Multiplication: At least one local micro-training (60–90 min) delivered within 60 days after certification.

What this toolkit is for

This toolkit helps the ambassadors turn critical thinking into consistent daily practice: how to notice questionable claims, check them efficiently, and communicate findings responsibly inside your NGO and wider communities. It's task-first: short explanations, then concrete actions you can take in 5/10/30 minutes, plus examples and worksheets. Read it once to learn the map, then return when planning trainings, responding to incidents, or mentoring new ambassadors. Treat it as a practical field manual, not a theory reader, and lean on the quick cards and links. The outcome is a shared language, clear workflows, and a baseline for quality.

You'll use it to:

Assess claims and sources quickly, even under time pressure.

Choose the right verification tools and workflows.

Explain your reasoning clearly and respectfully to non-experts.

Support your NGO's communication with reliable, documented checks.

u ime obitelji
World Youth Alliance Southeast Europe
PLATTFORM CHRISTDEMOKRATIE
ATEITININKAI

Funded by the European Union

What is inside this toolkit

The toolkit anchors itself in clear concepts — misinformation, disinformation, and malinformation—and habits like lateral reading and provenance checks, inoculation/prebunking, and rigorous source triage, all aimed at making every check reproducible. It frames modern disinformation as a fast, scalable system where attention capture and identity cues matter more than medium, and stresses that transparent verification, archives, and provenance standards remain the best leverage.

A practical ABCDE lens shifts work from post-by-post reactions to system awareness: Actor, Behaviour, Content, Degree, Effect—used to choose if/where to intervene and how strongly. The section also covers why people believe falsehoods (confirmation/availability biases, motivated reasoning, classic propaganda devices) and urges truth-first, respectful messaging to avoid backfire. It explains AI's role in both generation and amplification and prescribes counter-moves: verify origin, use visual/audio forensics (keyframes, reverse searches), archive every step, and communicate uncertainty.

Time-boxed 5/10/30 workflows convert theory into repeatable action. The 5-minute SIFT triage prevents accidental amplification; the 10-minute flow produces a short verdict with a confidence label; the 30-minute check documents ABCDE, context, ethics, and a proportional response, with a public-ready explainer template.

"Tools & Techniques" detail the practical how-to: provenance & source triage, reverse-image and keyframe searches, quote/number verification, domain/website forensics for clones/typosquats, social-network triage, OSINT geo/chrono checks, campaign pattern detection tied to ABCDE, a vulnerability lens to tailor responses, and strict archiving/traceability. It also includes a "what to use when" crib and a Response Matrix to match Degree × Effect to least-amplifying actions. Ethics and duty-of-care guardrails—minimize harm, protect privacy, show uncertainty, and comply with data rules—plus a quality checklist make outputs safer and reproducible.

Finally, European case snapshots (e.g., Operation Doppelgänger, MH17, Slovakia 2023 deepfake audio, Zelensky deepfake, LT hybrid pushes) turn methods into muscle memory through 30-minute labs that combine verification, ABCDE mapping, and proportional response.

Funded by the European Union

Primer & Key concepts

Before checking anything, it's important to understand what type of information you're dealing with. Crisp definitions prevent confusion and keep discussions productive. Use them during planning and debriefs, and cross-link the terms in your internal docs. Watch for common mix-ups like intent (mis vs. disinformation) and the neglect of malinformation. Each entry should include a two-line example and a "see also" link to methods.

Types of false or misleading information:

Misinformation – false but shared without intent to harm.

Disinformation – false content shared with intent to deceive.

Malinformation – genuine information used out of context to cause harm (e.g., doxxing).

Other important concepts:

Verification vs. validation: Verification checks facts (what happened?). Validation checks inference and method (does the evidence support the claim?).

Lateral reading & provenance: Instead of staying on one page, open new tabs to check who the source is. Before engaging content, identify who/what/where/when/why.

Inoculation / prebunking: Warning people *before* they encounter manipulation. Brief, proactive warning + example of the manipulation technique to build resistance before exposure.

Source triage: who published, track record, credentials/claim to expertise, funding, who shared.

Societal vulnerabilities: polarization, low media literacy, inequality, weak cybersecurity, low institutional trust, weak platform regulation → higher FIMI risk.

Traceability: Every claim you publish must be reproducible by a colleague using your notes, links, and archived captures.

See Glossary for concise definitions and examples.

Funded by the European Union

A short history of modern disinformation

Pre-digital roots (late 19th–mid-20th c.). Long before the internet, states and political movements learned to weaponize rumors, forged documents, and mass media. The early 1900s saw industrial-scale propaganda in newspapers and posters; WWI and WWII professionalized techniques like emotional framing, selective statistics, and demonization of out-groups. Radio and newsreels lowered distribution costs and raised the stakes: if you could seize a broadcast tower, you could shape reality for a nation overnight. These periods teach two enduring lessons: mediums change, but psychological levers—fear, identity, grievance—are remarkably stable; and distribution control is power.

Cold War "active measures" (1945–1991). Soviet and Warsaw Pact services ran multi-year influence campaigns blending forgeries, front outlets, and planted stories. The most famous was the KGB/Stasi AIDS operation: beginning in the mid-1980s, operatives seeded articles claiming HIV was a U.S. bioweapon, laundering the claim through foreign newspapers and pseudo-experts until it echoed globally. After the Cold War, officials confirmed the operation's authorship; historians later clarified its codename as Operation Denver (popularly known as Operation INFEKTION). The campaign's public-health harm—fueling mistrust and risky behaviour—remains a cautionary tale about the real-world costs of disinformation.

Commercial internet & search era (1990s–2008). The web rewired gatekeeping. Search engines and early forums created new incentives: capture attention, rank high, monetize via ads. "Content farms" and low-friction blogs blurred lines between opinion, rumor, and reporting. Forums and chain emails socialized "peer-to-peer credibility," where familiarity often substitutes for verification. The pattern that emerges here—optimize for clicks, then backfill a narrative—will only intensify on social platforms.

Platform ascendance & algorithmic feeds (2008–2013). Facebook's News Feed, YouTube recommendations, and Twitter trending lists made distribution both personalized and opaque. Microtargeted ads let actors test-and-iterate messages at low cost. The platform governance model—reactive, at massive scale—struggled to keep up with coordinated networks, sockpuppets, and "engagement hacking" that exploits novelty, outrage, and identity.

Networked operations go mainstream (2014–2018). Russia's information operations around Ukraine and Western elections showcased cross-platform coordination. The Internet Research Agency (IRA) ran thousands of personas, pages, and groups to exploit social cleavages; U.S. investigations and the Senate Intelligence Committee documented the reach and tactics, while platforms began major takedowns of IRA properties. Parallel to this, the 2014 downing of MH17 over Ukraine triggered a rapid cycle of competing narratives, doctored imagery, and official obfuscation; open-source investigators (notably Bellingcat) countered with

geolocation, weapon-system tracing, and timeline reconstruction, a pivotal moment for OSINT's role in debunking state disinformation.

Encrypted chats, fringe platforms, and the "infodemic" (2019–2021). Messaging apps and alternative platforms lowered moderation visibility and increased virality in closed networks. During COVID-19, the WHO labeled the crisis an infodemic—an overabundance of information (true, false, and everything between) that created confusion and dangerous behaviours. The pandemic era normalized "evidence-shaped to fit identity," where communities assembled bespoke realities from memes, videos, and doctored screenshots faster than institutions could respond.

War and platforms meet OSINT at scale (2022–2023). Russia's full-scale invasion of Ukraine accelerated both manipulation and counter-manipulation. Influence networks ran cloned outlets and impersonation sites—most visibly Operation Doppelgänger, which spoofed European media and ministries to plant anti-Ukraine narratives and erode support. Investigators and governments exposed the technique mix: look-alike domains, copied page templates, seeded stories, and ad buys to push reach. Meanwhile, open-source communities and professional newsrooms fused methods—satellite imagery, sensor data, on-the-ground video—to rebut claims in near-real-time, shifting the information balance toward transparent verification.

Generative AI era (2023–2025). Cheap synthesis changed the cost curve. Text, images, voices, and videos can now be fabricated or subtly altered at scale; low-effort actors can run "quantity-over-quality" campaigns that feel authentic enough to pass a casual scroll. Real-world elections started to register concrete harms: in January 2024, AI-generated robocalls mimicking President Biden tried to deter turnout in New Hampshire; U.S. regulators responded by banning AI voices in robocalls and pursuing fines and charges against the organizers and carriers involved. At the same time, platforms and publishers began piloting provenance tools like C2PA/Content Credentials, which embed verifiable edit histories into media files; the approach is promising but depends on ecosystem adoption and user interface clarity.

Europe's regulatory turn (2022–2025). The EU complemented voluntary measures with binding rules. The Strengthened Code of Practice on Disinformation (2022) pushed demonetization and transparency commitments among platforms, ad networks, and fact-checkers, while evaluations highlight uneven compliance. The Digital Services Act (DSA) then set legal obligations and enforcement powers—fully applicable to platforms since 17 February 2024—including systemic risk assessments (e.g., disinformation), data access for vetted researchers, and significant fines for breaches. Researchers and civil society are now probing how much these measures reduce monetization and reach for disinformation in practice, while regulators test the DSA against real cases (including "Doppelgänger"-style operations).

What actually changed—and what didn't. The medium evolved from print to broadcast to feeds to foundation models, but the playbook still revolves around attention capture, identity signals, and narrative repetition. The cheapness of creation and the opacity of curation (recommendation systems, private groups, encrypted channels) tilted the field toward scale and speed. Yet transparent verification, open archives, and provenance standards give defenders leverage—especially when paired with pre-bunking and community-rooted messengers. Notably, the measurable effects of exposure to foreign influence content can be smaller than feared at the individual level; the societal hazard is the aggregate: narrative saturation that erodes shared facts and consumes institutional time.

Starting fact-checking

Before we get started on the fact-checking process, let's understand how misinformation works and why even intelligent people believe false things. We'll briefly discuss when and how to respond, and the ethics of fact-checking.

After aligning concepts, we will jump to the 5/10/30 workflows during live work. Consult Tools & Techniques for specific methods, and study the European Case Studies for end-to-end examples. When Keep the cycle alive by looping findings into future improvements. At the end, we present country playbooks for the countries participating in our project, which indicate how to localize the years discussed earlier and which reliable sources to use. You can create a playbook for any country based on the examples. To sum up:

1. Start here to align on concepts and ethics.
2. Jump to Workflows (5/10/30) for step-by-step actions.
3. Consult Tools & Techniques for specific verifications (images, numbers, quotes, networks).
4. Study European Case Studies to see the workflows end-to-end.
5. View the country playbook for operating in your specific location

How disinformation works

Disinformation is not just a false post or a fake image — it is a **system**. It spreads because of how information is created, shared, and reacted to online. Understanding how this system works helps ambassadors recognise patterns quickly and choose the right response.

1. It starts with an actor

Someone creates or amplifies misleading content. It can be:

- An individual with strong opinions
- A coordinated group or network
- A bot or automated account
- A media outlet with an agenda
- A state actor trying to influence opinion

The key question: *Who benefits if people believe this?*

2. It uses strategic behaviour

Disinformation rarely spreads by accident. Common behaviours include:

- Posting the same message across many accounts

- Using emotional language to provoke reactions
- Automating posts to increase volume
- Paying for ads or boosting content
- Redirecting users through suspicious links

These behaviours help false claims reach more people faster.

3. Content is shaped to trigger emotions

Disinformation works best when it makes people feel:

- Outrage
- Fear
- Pride
- Threatened or unsafe

Strong emotions cause people to share before thinking. Images, short videos, and simple slogans are often used because they bypass critical thinking.

4. The goal is attention, not accuracy

Disinformation spreads because platforms reward:

- Engagement
- Clicks
- Shares
- Comments

The more emotional or sensational the content, the more the algorithm pushes it. False content doesn't need to be convincing — only attention-grabbing.

5. Degree: It spreads through networks

A single post may not cause harm, but networks can:

- Repeat the same message from different accounts
- Coordinate posting at specific times
- Use influencers to give it credibility
- Move the same narrative across platforms (FB → TikTok → Telegram)

This repetition makes the message feel familiar and therefore "true."

6. Effect: It aims to influence beliefs or behaviour

Disinformation tries to:

- Damage trust in institutions
- Increase polarization
- Discourage voting or participation
- Spread fear or uncertainty
- Manipulate public opinion or behaviour

The impact may be small for each person, but large at the community level.

7. Why people believe it

People are vulnerable when content:

- Confirms their existing beliefs
- Fits their worldview
- Seems to come from someone "like them"
- Shows a dramatic story or shocking image

This is normal human psychology — not a flaw.

8. The good news: simple checks work

Most disinformation can be stopped early by:

- Checking the source
- Searching for the original image or video
- Looking for earlier versions of the claim
- Pausing before sharing
- Asking: *What's the intention behind this post?*

Disinformation relies on speed. Fact-checkers rely on clarity. Slowing down even for 10 seconds breaks the cycle.

ABCDE lens is a *practical way to "see the system" behind a single claim, ABCDE shifts you from post-by-post reactions to system awareness: actor, behaviour, content, degree, and effect. Run it before committing resources so triage and strategy align. Don't fixate only on content—signals about actors and potential effects often determine proportional responses. Each use should end with a brief recommendation on what to do next.*

A – Actor: Who originates/amplifies the content? Individual, coordinated network, bot, media outlet, NGO, state actor? What is their track record and declared interests?

B – Behaviour: Posting cadence, cross-platform coordination, astroturfing, brigading, bot-like signals, monetization patterns.

C – Content: What is being claimed? Exact wording, data used, edits, visuals, emotional triggers, missing context.

D – Degree: How widespread and impactful? Is it niche, trending, or mainstream? Which audiences are targeted?

E – Effect: What harm could result (safety, public health, civic trust)? What counter-measures are proportionate?

Use ABCDE to decide whether to intervene, how strongly, and where (public post, private message, organizational statement, media request).

Human psychology: why smart people believe false things

Believing false information doesn't mean someone is uneducated or not intelligent. In fact, smart people can be just as vulnerable — sometimes even more so. This happens because disinformation works with human psychology, not against it. Biases, identity, and emotion shape how people receive corrections. Check your messages against common pitfalls and classic propaganda devices so they land without backfiring. Avoid shaming and repeating the false claim in headlines. Aim for truth-first, respectful explanations that invite dialogue.

- Cognitive biases:
    - Confirmation bias (we search for evidence we already agree with)
    - Availability bias (recent or vivid stories feel more likely)
    - Motivated reasoning (we defend our identity, not the facts)
- Propaganda devices (classic playbook): Name Calling, Glittering Generalities, Transfer, Testimonial, Plain Folks, Card Stacking, Bandwagon.
- Emotional dynamics: Fear, outrage, and identity pride increase shareability and reduce scrutiny.

Practical implication: Ambassadors should pair verification with empathetic communication—leading with clarity, not contempt. Focus on sharing verified truth, not proving others wrong. Avoid shaming — it closes conversation. Use clear, simple explanations. Address emotions, not only facts.

Understanding these psychological points helps you communicate in a way people can actually hear. Truth spreads best when it feels respectful, human, and relatable.

Deciding when and how to respond

Match response intensity to real-world risk and reach so you avoid amplifying minor content. Combine "degree" and "effect" to choose between private clarification, a short public post, a full explainer, or coalition messaging. Be wary of quote-tweeting or restating the falsehood prominently. Set a monitoring plan that specifies who checks what, where, and how often.

- Intervene when: the narrative is gaining traction, harms are plausible, your NGO is named, or your community is targeted.
- Match response to risk:
    - Low risk, low reach: quiet clarification to the individual.
    - Medium risk: short public post with a clear correction and one authoritative link.
    - High risk / coordinated: full explainer, coalition messaging, media engagement, and proactive prebunking.
- Avoid oxygen traps: Don't restate the false claim in headlines; lead with the truth, then address the falsehood.

Ethics & duty of care

Guardrails protect dignity, safety, and trust while you verify and communicate. Before publishing, ask whether you're minimizing harm, being transparent about uncertainty, protecting volunteers, and complying with legal and brand rules. Avoid doxxing, unnecessary personal detail, and GDPR issues; log harassment and escalate threats. Use an ethics checklist, a safety protocol, and a brief compliance note.

- Respect & dignity: Critique claims, not people.
- Proportionality: Consider unintended amplification; choose channels accordingly.
- Transparency: Share methods, sources, and uncertainty.
- Safety: Protect volunteers' identities where needed; log harassment; escalate threats.
- Compliance: Follow your organisation's brand, legal, and data-protection rules.

Funded by the European Union

Workflows

The 5/10/30-minute workflows are our time-boxed playbooks for doing fact-checking well under real-world pressure: a 5-minute SIFT triage to avoid accidental amplification, a 10-minute concise check that produces a short, linkable verdict with a confidence label, and a 30-minute full check that documents evidence, context (ABCDE), ethics, and a proportional response. We include them to turn good intentions into repeatable action—so every ambassador, regardless of experience, can follow the same steps, reach consistent quality, leave a transparent audit trail, and choose the least-amplifying response that still protects our community's trust.

In the end of this section we also provide short tutorials that can be used effectively in real-time to respond to disinformation and that can be useful in the learning process.

Here are the essential 5/10/30-minute workflows for your toolkit.

5 minutes — SIFT Quick Triage (before you share/forward)

Goal: Decide "ignore / save for later / quick clarify" without falling for traps.

1) Stop. Breathe. Don't amplify. Capture the post/link/screenshot (URL + timestamp). If it's a Story/Reel, screen-record and note the time.

2) Investigate the Source. Who published it? What's their track record/expertise? How are they funded? Who shared it onward and why? If it claims to be a known outlet or journalist, check the official contact page and compare domains (spot typosquats/clones/impersonation).

3) Find Better Coverage. Open two independent tabs (lateral reading). Search the core claim phrasing + site: operators. Look for primary or authoritative sources (official stats, court/authority statements, reputable outlets).

4) Trace to Origin. Scroll to the earliest version you can find. If visual: run Google Lens/TinEye quickly—do you see earlier dates or different contexts?

5) Decide + Note Risk. If it smells like a campaign element (bots, identical copy bursts, influencer hot-takes, "too polished" clone site), do not engage publicly yet. Jot: claim type (event/number/quote/image/video), confidence (low/unknown), and whether this might hit known societal vulnerabilities (polarization, low media literacy, low trust, etc.).

Outputs (≤60 words): one-paragraph triage note with link(s), earliest appearance, quick source verdict, and next action (ignore / ask privately / 10-min check).

Funded by the European Union

10 minutes — Concise Verification (publishable micro-finding)

Goal: Produce a short, linkable result with a confidence label.

0) Setup. Make a mini log: claim, URL(s), screenshot(s), timestamp(s). Archive key links (Wayback or Perma.cc). Classify: mis / dis / mal / unverified.

1) Classify claim type & pick the right micro-tools. Image: Save file → Google Lens/TinEye → compare earliest use & higher-res matches → check for crops/edits.

- Video: Extract 3–5 keyframes (right-click frame, or a keyframe tool) → reverse search frames. Scan for AI tells: lip-sync/A-V mismatch, glassy gaze, too-smooth skin, deformed hands, overly clean/monotone voice.
- Quote/"X said Y": Search exact phrase in quotes + site: of official pages or reputable outlets.
- Number/statistic: Identify the primary dataset; check the latest official publication (method, time range, caveats).
- Account/source: Check domain age/registrar, TLS certificate, about/contact page, bylines. Compare to the legitimate outlet's site (detect cloned layouts/typosquats). Glance at posting history for copy-paste bursts and coordinated hashtags (behaviour signal).

2) Cross-validate. Open at least two independent corroborations (authority + reputable reporting). If none: say unverified and what would verify it.

3) Degree & Effect (fast). Is it spreading beyond a niche? Who's targeted? Any geo-blocking/redirect chains/ads visible? Note if it taps known community vulnerabilities.

4) Micro-writeup. Two sentences: what's claimed; what you found (with strongest link). One sentence: confidence + reason ("earliest instance from 2019 shows different context"; "no record in primary dataset"). Ethics check: avoid amplifying the falsehood in the headline; don't include unnecessary personal data.

Outputs:

- 1-paragraph verdict (True / False / Misleading / Unverified) with confidence.
- 3 links (origin, best alternative coverage, primary/official).
- Label Response: quiet clarification / short public post / escalate to 30-min check.

Common pitfalls to avoid: emotionally charged framing, stats without sources, "professional-looking" but masquerading websites, and identity-triggering propaganda devices.

Funded by the European Union

30 minutes — Full Check (reproducible, public-ready)

Goal: A transparent, archived explainer with proportional response guidance.

0) Admin & Safety. Start a note (date, team, channels monitored). Archive all key URLs (Wayback/Perma). If harassment or doxxing risk exists, switch to safer comms; log incidents (duty of care).

1) Map ABCDE. Actor: Who originates and amplifies? Any signs of state-aligned media, persona farms, paid trolls, botnets, influencers? Verify legitimate journalists/outlets via official pages (no impersonation).

- Behaviour: Typosquatted/cloned domains, copied templates, redirect chains, paid ads, geo-blocking, synchronized posting, comment brigades, manipulated chatbot "flooding". Document with screenshots + timestamps.
- Content: Exact wording, visuals, edits/crops, emotional cues, missing context. Compare to earliest appearances from Lens/TinEye or keyframe searches.
- Degree: Reach/velocity/audiences; look for monetization hooks; note platforms involved.
- Effect: Likely harms (public safety, civic trust, reputation). Choose least-amplifying effective channel for response.

2) Evidence stack

- Primary/official: datasets, legal/agency docs, on-record statements.
- Independent reporting/experts: reputable outlets, recognized specialists.
- Technical checks: WHOIS/domain age, TLS certs; C2PA/Content Credentials (if present); EXIF/metadata (if available); style/time/weather/landmark checks for geolocation (Street View/Mapillary, sun/shadow).
- Campaign elements: hostile goal framing, trolls/bots, algorithm gaming, influencers, and journal/science masquerade—capture examples.

3) Vulnerability lens. Note context: polarization, low media literacy, inequality, weak cybersecurity, low trust, weak regulation. If multiple "yes," prioritize prebunk + media-literacy tips in the response.

4) Analysis & conclusion. Write a short explainer with:

- Headline that leads with truth (not the falsehood).
- 3-bullet summary (what's true/false, what's missing, why it matters).
- Method (what you checked, how).
- Verdict + confidence (True / False / Misleading / Unverified + High/Med/Low).

- Sources & archives (origin, strongest evidence, alternatives).
- Proportional response (quiet outreach / short public correction with 1 authoritative link / full post + coalition coordination / media request).
- Safety & ethics note (privacy respected, uncertainty stated, no doxxing).

5) Publish & monitor. Choose channel(s) that minimize amplification but reach the affected audience. Add a monitoring plan: who watches which platforms for 48–72h; when to update/retract.

Public explainer template

- Title: Lead with verified truth (avoid repeating the false claim).
- Summary (3 bullets): [A], [B], [C].
- What we checked: [Claim, where it appeared, earliest instance].
- How we checked: [Tools: Lens/TinEye/keyframes/WHOIS/dataset X].
- What we found: [Concise narrative with links].
- Verdict & confidence: [e.g., Misleading — Medium confidence].
- Why it matters: [Effect on community/decision].
- Sources & archives: [urls + archive links].
- Notes on limitations & next steps: [What would raise confidence].
- Contact: [NGO/AF comms address].

Tools and Techniques

These tools and techniques are the practical "how" of our toolkit—the repeatable methods that turn skepticism into reliable verification. They cover the full arc of a claim: provenance and source triage; image, video, and audio checks (incl. AI-manipulation cues); quote and data verification; domain/website forensics for typosquats and clones; social-network triage; OSINT geo/chrono-location; campaign-pattern detection using ABCDE; the vulnerability lens to tailor responses; and ethics, safety, and archiving for traceability. We provide them so every ambassador can act quickly and consistently, avoid accidental amplification, produce transparent, reproducible findings, and choose the least-amplifying response that still protects our community's trust.

Provenance & Source Triage

When to use: First contact with any claim.

How: Check publisher, track record, claim to expertise, funding, and who amplified it. Open independent tabs (lateral reading) and compare the domain to the outlet's official site to catch impersonation and typosquats.

Outputs: One-line source verdict + link(s) to official contact/about page.

Pitfalls: Credibility by design (professional-looking but fake), appeal to authority without verifiable credentials.

Visual Verification (images)

When to use: Photos, screenshots, memes, "evidence pictures."

How: Save the file → run Google Lens and TinEye → find earliest appearances and higher-res matches → check for crops/edits or reused images in new contexts.

Outputs: Earliest match URL + date, 1–2 comparison links, verdict (original/edited/reused).

Pitfalls: Relying on a single engine; ignoring language/region filters; forgetting to archive.

Video & Audio Verification (including AI cues)

When to use: Clips, reels, speeches, voice notes.

How: Extract 3–5 keyframes → reverse search each; compare audio and mouth movement; listen for over-clean, monotone timbre. Use AI-video tells list: lip–speech mismatch, A/V desync, glassy gaze/no blinking, too-smooth skin, deformed hands/fingers, robotic delivery.

Outputs: Keyframe matches, short A/V consistency note, verdict + confidence.

Pitfalls: Judging by "vibe"; ignoring room acoustics/lighting mismatches; forgetting that genuine low-quality video can look "AI-ish."

Quote/Claim Verification

When to use: "X said Y," viral screenshots of posts, headlines.

How: Search the exact phrase in quotes; check official channels (press pages, verified accounts); compare timestamps. For screenshots, hunt the original post or archived copy before trusting a cropped image.

Outputs: Link to original or authoritative denial/confirmation; verdict + confidence.

Pitfalls: Fake screenshots; satire out of context; quote fragments that flip meaning.

Data, Numbers & Graphs

When to use: Polls, crime stats, health numbers, "study shows."

How: Identify the primary dataset; verify the time window, method, denominators, and exclusions; recompute a simple check if possible. Note uncertainty and margin of error.

Outputs: Primary source link, one-sentence method note, corrected figure (if needed).

Pitfalls: Cherry-picked ranges, percent vs. percentage-points, model projections treated as observations.

Website & Domain Forensics

When to use: Suspected cloned outlets, typosquats, redirect chains.

How: Compare domain spelling and TLD to the real outlet; check WHOIS/creation date, TLS certificate, bylines and section slugs; follow links to see if they loop via redirectors or geo-blocked routes. Log ads that appear only on the clone (behaviour signal).

Outputs: Side-by-side domain facts (age/issuer/paths), screenshots of layout/bylines, verdict (legit/clone).

Pitfalls: Assuming HTTPS implies authenticity; missing subtle domain swaps (rn vs. m).

Social-Network Triage

When to use: Accounts, threads, trending posts.

How: Archive the post; scan account history for copy-paste bursts, time-zone patterns, sudden follower spikes, or identical comment strings (bots/brigades). Check hashtag choreography and cross-posting to other platforms.

Outputs: 3–5 screenshots with timestamps; note on coordination signals; quick Degree estimate (reach/velocity).

Pitfalls: Over-attributing coordination to fandoms; ignoring language variants.

OSINT Geolocation & Chronolocation

When to use: "This happened here/now" claims.

How: Match landmarks via Google Maps/Street View/Mapillary/OpenStreetMap; compare skylines, road furniture, shop signs. Estimate time with sun/shadow direction, weather archives, or event calendars.

Outputs: Matched coordinates + proof images; time estimate (if feasible).

Pitfalls: Renovations/new signage since imagery capture; mirrored/cropped visuals.

Campaign Pattern Detection (ABCDE-linked)

When to use: Recurrent narratives or multi-asset pushes.

How: Log Actor (state-aligned media, persona farms, influencers), Behaviour (impersonation emails, typosquats, cloned layouts, ads, redirect chains, geo-blocking), Content (identical talking points), Degree (cross-platform spread), Effect (target audience, plausible harm).

Outputs: One-page pattern sheet + proportional response recommendation.

Pitfalls: Treating one post as the problem when the infrastructure is the point.

Vulnerability Lens (context-aware risk)

When to use: Before publishing or engaging communities.

How: Run checklist: polarization, media-literacy levels, socio-economic inequality, weak cybersecurity, low institutional trust, weak regulation. If several "yes," prefer prebunk, careful tone, and local messengers.

Outputs: 2-line context note that justifies response choice.

Pitfalls: Blaming the audience; ignoring community knowledge holders.

Ethics, Safety & Compliance

When to use: Every check before publication.

How: Apply the Ethics Checklist: minimize harm; avoid unnecessary personal data; show uncertainty; use least-amplifying channel; log harassment and escalate if needed; respect copyright/consent/GDPR.

Outputs: Tick-box record in the case file.

Pitfalls: Restating the falsehood in headlines; doxxing by accident (screenshots with PII).

Archiving & Traceability

When to use: From first click.

How: Save original URLs, screenshots with timestamps, and archive copies (Wayback/Perma). Name files consistently. Keep a short methods log so a colleague can reproduce your result.

Outputs: Compact case folder (origin → method → verdict).

Pitfalls: Moving targets (deleted posts); missing timezones.

Funded by the European Union

European Case Studies

These European Case Studies are concrete, real-world scenarios—like Operation Doppelgänger's cloned outlets, MH17's OSINT reconstruction, the Slovakia pre-election deepfake audio, the Zelensky "surrender" deepfake, and Lithuania's hybrid bot-and-scam pushes—that let ambassadors practice the exact skills in our toolkit under believable conditions. Each one pairs the narrative with the exact signals, tools, and workflows you teach (SIFT • 10-minute check • 30-minute ABCDE), and leans on Debunk-style practice. We include them to turn abstract methods into muscle memory: spotting behaviour signals (cloned domains, redirects, bots), running fast visual/audio checks (Lens, TinEye, keyframes, AI-tells), mapping ABCDE to choose a proportional response, and documenting findings with archives and confidence labels. By rehearsing with European examples close to our context, teams learn to act quickly, avoid accidental amplification, and protect community trust with transparent, reproducible verification.

"Operation Doppelgänger" — cloned outlets & typosquats targeting the EU

Narrative. A long-running influence operation publishes articles on look-alike domains that mimic European media and institutions, then boosts them via ads, redirects, and influencer pickup.

Signals & tools. Domain forensics (typos/TLD swaps, WHOIS age, TLS issuer), layout/byline/contact-page mismatches; ad/redirect chains; geo-blocking; archiving.

Workflow mapping. SIFT (spot impersonation); 10-min (domain/TLS + earliest image appearances via Lens/TinEye); 30-min ABCDE (Actor = infrastructure + amplification; Behaviour = clones/ads/redirects; Degree = cross-platform spread; Effect = reputational/civic harm).

What to teach. "Provenance first" beats hot takes; treat the infrastructure (clone network) as the story, not just one post.

Read more. EU DisinfoLab overviews and PDFs; CORRECTIV's routing analysis of the redirect chain.

MH17 (2014) — OSINT geolocation & timeline reconstruction

Narrative. Competing claims followed the downing of flight MH17 over Eastern Ukraine. Open-source investigators traced a Buk launcher's route, linking imagery, serial markings, roadsigns, and timestamps.

Signals & tools. Reverse-image/video checks, Street View/Mapillary, signage/landmark matching, shadow/time checks, archive captures; structured case notebook.

Workflow mapping. 10-min (keyframes → reverse search; triangulate earliest posts); 30-min ABCDE + evidence stack (primary/official findings, reputable reporting, OSINT).

What to teach. How reproducible methods and archiving create courtroom-grade confidence.

Read more. Bellingcat's report(s); later references in European proceedings and media.

Slovakia 2023 — pre-election deepfake audio drop

Narrative. Days before the vote (during media silence), AI-generated audio appeared, posing as a conversation about rigging the election, spreading rapidly on social platforms and messaging apps.

Signals & tools. Audio forensics basics (room tone, cuts, prosody); source/provenance gaps; sudden cross-platform pickup; influencer amplification without primary source; platform policy blind spots for audio.

Workflow mapping. SIFT (don't amplify; capture originals; note silence-period risk); 10-min (exact-phrase search; seek original upload; compare with verified voice samples; log uncertainty); 30-min (Degree & Effect → proportional response; prebunk guidance for partners).

What to teach. Why audio deepfakes are harder to police than video—and how to communicate uncertainty clearly and fast.

Read more. Wired explainer; Bloomberg coverage; incident databases summarizing the drop and timing.

Zelensky "surrender" deepfake (2022) — rapid debunk in wartime

Narrative. A fabricated video of President Zelensky urging surrender briefly aired online (and reportedly via compromised outlets), but was quickly countered by an authentic statement and platform takedowns.

Signals & tools. AI-video tells: lip/A-V mismatch, glassy gaze, too-smooth skin; keyframe reverse search; provenance checks; official-channel verification.

Workflow mapping. 10-min (keyframes → reverse search; A/V consistency note; verify on official channels); 30-min (clear explainer leading with truth; archive links; confidence label).

What to teach. Preparedness + fast, truth-first messaging can blunt impact even when a deepfake lands.

Read more. France24 debunk; Euronews context; analysis of the rapid counter-message playbook.

Lithuania (2023–2024) — large-scale hybrid scam + disinformation push

Narrative. Debunk.org documented a coordinated attack mixing scam pages, inauthentic accounts, and pro-Kremlin narratives aimed at Lithuanian audiences; later reports track bot networks and themed pushes (e.g., Kaliningrad-related content).

Signals & tools. Network signals (copy-paste bursts, time-zone patterns), bot-like activity, campaign branding, cross-posting; provenance + archiving; vulnerability lens (low trust, regional tensions).

Workflow mapping. SIFT (archive; source triage); 10-min (account history; cross-platform checks; earliest appearances); 30-min ABCDE (Actor/Behaviour = botnets + troll pages; Degree = reach/velocity; Effect = community risk) → response matrix and prebunking.

What to teach. Join dots from posts → network → narrative → harm, then pick the least-amplifying effective response.

Read more. Debunk.org investigation(s) and follow-ups on bot networks and targeted topics.

Funded by the European Union

Country Playbooks

The purpose of these country playbooks is to localize methods, sources, and examples for each target context. We chose the countries where the activities of this project were carried out - Austria, Croatia and Lithuania.

Each playbook includes top 10 authoritative sources (gov/agency data portals; fact-checking partners), common narratives & seasonal cycles (e.g., elections, energy, migration, public health), local platform patterns (messaging apps vs. open networks; language variants), legal/compliance specifics (data, media, election-silence windows) and three local case vignettes with ready exercises.

Lithuania (LT)

Top authoritative sources (for quick linking):

- Official statistics: Statistics Lithuania / State Data Agency (OSP portal) – www.osp.stat.gov.lt
- Elections: Central Electoral Commission (VRK). – www.vrk.lt
- Health: Ministry of Health (SAM) - www.sam.lrv.lt
- Cybersecurity: National Cyber Security Centre (NCSC) – www.nksc.lt
- Fact-checking/OSINT references: Debunk.org (regional analyses) – www.debunk.org

Common narratives & seasonal cycles: border/migration incidents; energy/price shocks; NATO presence; election-period rumours; health policy scares. Use prebunk templates when VRK calendars publish milestones.

Watchpoints & platform patterns: cloned media domains and Doppelgänger-style articles; geo-blocked redirects; Facebook/Telegram cross-posting. Run domain/TLS/WHOIS checks + reverse-image (Lens/TinEye) for reused visuals.

Legal/compliance notes: check VRK pages for current election info before posting corrections; archive (Wayback/Perma) all links used in a check.

Croatia (HR)

Top authoritative sources:

- Official statistics: Croatian Bureau of Statistics (DZS) – www.dzs.gov.hr
- Elections: State Electoral Commission (DIP/SEC) - www.izbori.hr
- Health: Ministry of Health – www.zdravlje.gov.hr
- Cybersecurity: National CERT (CERT.hr / CARNET) - www.CERT.hr

Common narratives & seasonal cycles: EU topics (Schengen/euro), migration routes, energy/tourism, election-season rumours. Expect copy-paste bursts across Facebook pages and portals.

Watchpoints & platform patterns: typosquats of national outlets; Telegram/FB groups recycling old photos with new captions. Use 10-minute flow: keyframes → reverse search; domain age/TLS check.

Legal/compliance notes: use SEC pages for procedures and timelines when claims touch voting or turnout; archive screenshots with timestamps in case posts are deleted.

Austria (AT)

Top authoritative sources:

- Official statistics: STATISTICS AUSTRIA - www.statistik.at
- Elections: Federal Electoral Board (Bundeswahlbehörde) via Interior Ministry - www.bmi.gv.at
- Health: Federal Ministry of Social Affairs, Health, Care & Consumer Protection. - www.sozialministerium.gv.at
- Cybersecurity: CERT.at (national CERT) - www.cert.at

Common narratives & seasonal cycles: energy prices & climate policy; EU regulation; migration; cross-border stories with DE/IT/CZ. Cloned pages sometimes spoof ministries or national broadcasters.

Watchpoints & platform patterns: redirect chains and ad placements on look-alike news pages; local Facebook groups sharing screenshots from German pages. Validate source provenance first; map Behaviour (clones/ads/redirects) before debating Content.

Legal/compliance notes: election information is centralized—link to Bundeswahlbehörde pages for rules and official notices.

Conclusion

This toolkit turns good intentions into reliable practice. By pairing clear concepts with the 5/10/30 workflows, ABCDE triage, and a disciplined ethics and archiving routine, we give every ambassador a repeatable path from “this looks suspicious” to a transparent, proportionate response. The tools and techniques—source triage, visual and audio verification, domain forensics, OSINT geo/chrono checks, and campaign-pattern detection—are deliberately simple to start and rigorous enough to stand up in public. European case studies, country playbooks, and quick cards translate methods into muscle memory so teams can act fast without amplifying harm.

What matters most is consistency and care. Lead with truth, show your method and uncertainty, and choose the least-amplifying channel that still protects people. Capture originals, archive links, and leave a trail a colleague can reproduce tomorrow. Use the vulnerability lens to shape tone and timing; apply the response matrix to match action to risk; and measure what you do with lightweight KPIs so quality improves month by month. As platforms, tactics, and AI evolve, this toolkit is designed to evolve too: add new cases, update your quick cards, refresh country sheets, and run regular drills. If we keep learning together—and keep our work traceable, respectful, and clear—we strengthen trust in our communities and make it harder for manipulation to stick.

Funded by the European Union

Appendices

Glossary

ABCDE Framework — Actor, Behaviour, Content, Degree, Effect—a structured way to analyze operations.
Example: Map the network (Actor), coordination (Behaviour), narratives (Content), reach (Degree), and impact (Effect).

Archiving (Wayback, Memento) — Saving snapshots of webpages to preserve evidence and compare changes.
Example: Wayback shows the headline was edited after publication.

Attribution — Assessing who is likely behind an operation; often caveated and evidence-based.
Example: Language patterns and hosting link a site to a known influence firm.

Bias & Heuristics — Cognitive shortcuts (confirmation, availability, motivated reasoning) that skew judgment.
Example: You preferentially notice examples that fit your prior beliefs.

Bot / Automation — Accounts that auto-post or amplify content at unnatural speed/scale.
Example: A profile posts every 2 minutes, 24/7, in multiple languages.

Chain of Custody — Documenting how evidence was obtained and stored to maintain integrity.
Example: Recording the download time, URL, and hash for a video.

Chronolocation — Confirming when an image/video was captured using sun position, weather, or event clues.
Example: Shadow angles and a weather archive show the video is from April 2023.

Claim — A statement that asserts something is true or false; the unit you verify.
Example: A post says: "The city banned all gas stoves starting January 2026."

Context Collapse — When information is shared beyond its original context, changing its meaning.
Example: A satirical headline reposted as if it were real news.

Coordinated Inauthentic Behaviour (CIB) — Organized use of fake accounts/pages to mislead about identity or purpose.
Example: A cluster of pages shares identical posts at the same minute daily.

Correction / Update — A transparent note that fixes errors or adds new info post-publication.
Example: "Updated on 4 Nov 2025: corrected date from Oct 12 to Oct 21."

Cross-Reference — Corroborating a fact with independent, credible sources.
Example: Confirming numbers via the official dataset and a separate audit report.

Deepfake / Synthetic Media — Audio, image, or video generated or altered by AI to realistically mimic real people or events.
Example: An AI-cloned voice recording falsely 'admitting' to a crime.

Disinformation — False or misleading information deliberately created or spread to deceive.
Example: A network of pages fabricates polling numbers to discourage voting.

DSA / Code of Practice — EU Digital Services Act and voluntary Code guiding platform responsibility and transparency.
Example: Platforms publish ad libraries and risk assessments per DSA.

Evidence — Information that supports or refutes a claim. Can be documents, data, images, video, or expert testimony.
Example: A city council vote record and the official ordinance text.

Geolocation — Confirming where an image/video was captured by matching landmarks, signs, or terrain.
Example: Identifying a skyline and street signs to place a video in Kaunas.

Hash / Checksum — A unique fingerprint of a file to prove it hasn't changed.
Example: MD5/SHA256 matches before and after transfer.

Lateral Reading — Opening new tabs to check who the source is before reading closely.
Example: You first check a site's 'About' page and external profiles.

Malinformation — Genuine information shared out of context or with harmful framing to cause damage.
Example: Leaking a person's home address to incite harassment.

Metadata / EXIF — Embedded technical data about a file (device, time, GPS). Often stripped by platforms.
Example: Original photo shows GPS coordinates and capture time.

Misinformation — False or misleading information shared without intent to deceive.
Example: A friend shares an outdated map with wrong evacuation zones.

Narrative vs. Claim — A narrative is a broader storyline; a claim is a specific checkable statement.
Example: Narrative: 'Elections are rigged.' Claim: '10,000 dead voters cast ballots in City X.'

OSINT (Open-Source Intelligence) — Collecting and analyzing publicly available data to verify facts.
Example: Using company registries and flight trackers to confirm a person's travel.

Prebunking / Inoculation — Teaching manipulation tactics and counter-arguments before people encounter them.
Example: A lesson on 'false dilemmas' reduces later susceptibility.

Primary Source — Original material created at the time of an event or directly by the subject.
Example: The ordinance PDF on the city's official website.

Propaganda Devices — Classic techniques: name-calling, glittering generalities, transfer, testimonial, plain folks, card stacking, bandwagon.
Example: An ad uses a celebrity 'testimonial' without evidence.

Proportionality — Matching the strength of your conclusion to the quality/quantity of evidence.
Example: "Unverified" when sources conflict; avoid overclaiming.

Reliability Labels — Platform or third-party indicators for outlets (e.g., public broadcaster, state-affiliated).
Example: A video shows a 'state-controlled media' label under the channel name.

Reverse Image Search — Finding earlier or original versions of a picture to check origin and context.
Example: Tracing a protest photo to a 2016 event, not 'today.'

Safety & Harm Assessment — Judging potential risks (privacy, doxing, retraumatization) before publishing.
Example: You blur faces of minors in protest footage.

Satire / Parody — Humorous or exaggerated content that imitates real news or people.
Example: A site publishes a spoof article about 'flat Earth' rocket launches.

Secondary Source — A summary, analysis, or report based on primary sources.
Example: A news article describing the ordinance and quoting officials.

SIFT Method — Stop, Investigate the source, Find better coverage, Trace to the original.
Example: Before retweeting, you stop and find the original press release.

Sockpuppet / Astroturfing — False identities feigning grassroots support.
Example: A company runs 'citizen' pages praising its product and attacking critics.

Source Transparency — Clear info about who runs a site/account, funding, corrections policy, and contact details.
Example: The outlet lists its editorial team and a corrections log.

Truth Sandwich — Communications tactic: state the truth → address the myth → restate truth with evidence.
Example: "Vaccines are safe and effective. A viral post falsely claims X. Here's the data showing safety."

Verification Workflow (5/10/30) — Triage depth based on time available: 5-min quick check, 10-min deeper, 30-min full.
Example: In 5 minutes you check headline, source, and original link; in 30 you contact experts.

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