AI War Propaganda: How to Verify Deepfakes, Memes and Conflict Content

Maria Cattini | 08/05/2026 | AI

A video appears on X minutes after an alleged strike. It looks dramatic enough to travel fast: explosions, smoke, a military frame, captions written with certainty. Then come the counter-posts. One account says the footage is real. Another calls it AI. A third says the same clip comes from another war. By the time journalists, fact-checkers, or investigators begin checking it, the public has already absorbed the impression.

That is the new operational problem of war reporting. The fog of war is no longer only produced by distance, censorship, fear, or battlefield confusion. It is now thickened by synthetic images, meme propaganda, platform algorithms, and the public habit of scrolling before checking.

The Reuters Institute article by Gretel Kahn describes this shift through the current US-Iran conflict, where AI-generated content, fake drone footage, fabricated satellite images, edited clips, synthetic statements, memes, and parody material circulate alongside authentic documentation. The result is not just misinformation. It is a damaged evidence environment.

The Problem Is Not Only Fake Content

The easiest mistake is to treat every AI image as a simple verification problem: real or fake. That frame is too narrow.

Some content is designed to deceive. A fabricated satellite image, a fake military clip, or a deepfake victim photo can create a false account of events. But other material works differently. Meme videos, Lego-style animations, parody clips, and “slopaganda” may be visibly artificial, yet still push a worldview. They do not need to fool the viewer completely. They only need to leave an emotional trace.

That distinction matters for OSINT work. A fake image asks: “Did this event happen this way?” A propaganda meme asks: “What attitude is this content trying to normalize?”

The Reuters Institute piece describes how both the United States and Iran have used internet-native aesthetics to project narratives during the conflict. The US side is described as using military spectacle, supercuts, drone footage, film references, and video game imagery. Iranian-linked communication is described as leaning into AI-generated media, parody, deepfakes, stylized videos, and unrelated humorous clips.

For investigators and journalists, the issue is not whether memes are “serious.” They are serious because they shape attention before evidence can catch up.

How AI Changes the Evidence Environment

Traditional misinformation often relied on old footage with a new caption. That still happens. The new layer is synthetic media produced at speed and distributed through accounts that audiences may already trust.

The article highlights a more dangerous effect: authentic content becomes easier to dismiss. Mahsa Alimardani of WITNESS describes how verified images of civilian casualties from the Minab school strike were dismissed by some accounts as AI-generated or recycled based on aesthetic impressions rather than forensic method. Claims such as “the lighting is too good” or “it looks staged” spread before investigators confirmed authenticity.

This is the “liar’s dividend”: once synthetic media exists, bad actors can use its existence to reject real evidence. A genuine image can be attacked not because it fails verification, but because it feels too polished, too cinematic, too emotionally convenient.

That is where the public conversation often collapses. People begin treating suspicion as analysis. In OSINT, suspicion is only the starting point. It is not evidence.

A Practical Verification Workflow for AI-Heavy Conflict Content

The goal is not to “detect AI” in isolation. The goal is to assess a claim, preserve the material, separate visual doubt from evidence, and document what can and cannot be established.

1. Capture the claim before analyzing the media

Start with the exact claim attached to the image or video. Record the post URL, account name, timestamp, caption, quoted source, language, and engagement context. Save screenshots and, where lawful and appropriate, preserve the page.

Reason: the same visual can support different claims. A clip may be authentic but miscaptioned. A synthetic image may represent a real event falsely illustrated. A meme may not claim literal accuracy but may still serve propaganda.

Look for: who posted it first in your visible chain, what they claimed, whether they attributed it to an official source, media outlet, witness, or anonymous channel.

Expected result: you should end this step with a claim log, not a verdict.

Common mistake: downloading the image and forgetting the surrounding text. In conflict verification, the caption is often the payload.

2. Separate the object from the narrative

Treat the content in layers: the visual object, the caption, the account, the amplification pattern, and the political narrative.

Reason: AI-generated material may be fake as an image but still attached to a real event. Real footage may be attached to a false location. A parody video may be obviously artificial but still designed to dehumanize or normalize escalation.

Look for: signs that the post is making a factual claim, mocking an enemy, validating an official account, or undermining trust in other evidence.

Expected result: a structured assessment: “The image is unverified,” “the claim is unsupported,” “the account frames it as evidence,” or “the content appears propagandistic rather than evidentiary.”

Common mistake: writing “fake” and stopping there. Fake content can still have real influence.

3. Check provenance, but do not overestimate it

Provenance means understanding where a file came from and whether it has been altered. In theory, standards such as C2PA can help create a chain of custody. In practice, the Reuters Institute article notes that such signals are not yet widely deployed enough to solve fast-moving conflict verification:

many phones do not sign images at capture, many models do not embed provenance, and many platforms do not show provenance signals to users.

Reason: absence of provenance is not proof of fabrication. Presence of metadata is not proof of truth.

Look for: original upload context, platform labels, visible edits, repost trails, compression artifacts, and whether credible investigators have preserved earlier versions.

Expected result: a provenance note with limits clearly stated.

Common mistake: treating “no metadata” as suspicious on its own. Social platforms often strip metadata.

4. Triangulate with independent signals

Do not rely on a single artifact. Compare the claim with other available material: different angles, local reports, official statements, known landmarks, timing, weather, shadows, language, uniforms, terrain, or damage patterns.

Reason: Manisha Ganguly, cited in the Reuters Institute article, warns against relying on single artifacts or sources. AI-generated satellite imagery and other synthetic visuals can cosmetically validate official narratives if treated as standalone proof.

Look for: consistency across independent sources, not repetition from the same network. Ten reposts of the same image are not ten sources.

Expected result: a confidence level, not a theatrical debunk.

Common mistake: confusing virality with corroboration.

5. Evaluate the amplification path

Ask how the content moved. Did it travel through state-linked media, official accounts, influencers, anonymous aggregators, diaspora networks, meme pages, or news outlets?

Reason: platforms are not passive pipes. The Reuters Institute article cites Emerson T. Brooking’s point that algorithms are central to this kind of information conflict because discovery is not based only on who users follow or what they say they want.

Look for: sudden engagement spikes, repeated captions, coordinated phrasing, official reposts, and whether platform incentives reward the most emotional version.

Expected result: a distribution map showing why the content mattered.

Common mistake: focusing only on authenticity while ignoring reach.

6. Publish uncertainty without creating more fog

When verification is incomplete, say exactly what is known, what is unknown, and what would change the assessment. Avoid aesthetic claims such as “looks AI” unless backed by method.

Reason: the Minab example shows how aesthetic suspicion can become a weapon against authentic documentation. Bad verification language can damage the evidentiary space it tries to protect.

Look for: wording that separates evidence from interpretation.

Expected result: a responsible public note: “We have not verified the location,” “the image has not been authenticated,” “the event is reported, but this visual cannot yet be linked to it,” or “the claim

relies on synthetic or unverified imagery.”

Common mistake: publishing a confident denial before the evidence is checked.

Practical Scenario: The Synthetic Satellite Image

A state-linked outlet posts a satellite image claiming a military target was destroyed. The image spreads under the authority of a news brand. Some users treat it as OSINT proof.

A responsible workflow would not begin with a dramatic thread. It would begin with preservation: capture the post, image, caption, timestamp, account identity, and engagement. Then separate the claims: Is the outlet claiming the strike happened? Is the image presented as satellite proof? Is the image original, reposted, or unattributed?

Next, compare the visual with other available imagery and reporting. If no independent confirmation exists, the correct result is limited: the post cannot be used as evidence of the strike. If later forensic review shows the image is AI-generated, the finding should say so while avoiding a broader unsupported claim about the entire event.

The expected result is not “nothing happened.” The expected result is: “This specific visual does not prove what the post claims.”

That distinction protects the investigation from becoming propaganda in reverse.

The Bigger Risk: War as Entertainment

Memes and AI videos also change the emotional frame of conflict. Sam Dubberley of Human Rights Watch warns in the Reuters Institute article that videogame imagery, Lego videos, and memeification can make war feel less real, potentially desensitizing audiences to civilian harm.

For OSINT practitioners, that creates a second duty. Verification is not only about proving whether a file is authentic. It is also about refusing the platform logic that turns civilian harm into shareable spectacle.

The best defensive posture is slow by design: preserve first, separate claims, triangulate, state limits, and resist aesthetic verdicts. In an AI-thickened fog of war, the investigator’s job is not to react faster than propaganda. It is to make the evidence harder to corrupt.
A video appears on X minutes after an alleged strike. It looks dramatic enough to travel fast: explosions, smoke, a military frame, captions written with certainty. Then come the counter-posts. One account says the footage is real. Another calls it AI. A third says the same clip comes from another war. By the time journalists, fact-checkers, or investigators begin checking it, the public has already absorbed the impression.

That is the new operational problem of war reporting. The fog of war is no longer only produced by distance, censorship, fear, or battlefield confusion. It is now thickened by synthetic images, meme propaganda, platform algorithms, and the public habit of scrolling before checking.

The Reuters Institute article by Gretel Kahn describes this shift through the current US-Iran conflict, where AI-generated content, fake drone footage, fabricated satellite images, edited clips, synthetic statements, memes, and parody material circulate alongside authentic documentation. The result is not just misinformation. It is a damaged evidence environment.

The Problem Is Not Only Fake Content

The easiest mistake is to treat every AI image as a simple verification problem: real or fake. That frame is too narrow.

Some content is designed to deceive. A fabricated satellite image, a fake military clip, or a deepfake

victim photo can create a false account of events. But other material works differently. Meme videos, Lego-style animations, parody clips, and “slopaganda” may be visibly artificial, yet still push a worldview. They do not need to fool the viewer completely. They only need to leave an emotional trace.

That distinction matters for OSINT work. A fake image asks: “Did this event happen this way?” A propaganda meme asks: “What attitude is this content trying to normalize?”

The Reuters Institute piece describes how both the United States and Iran have used internet-native aesthetics to project narratives during the conflict. The US side is described as using military spectacle, supercuts, drone footage, film references, and video game imagery. Iranian-linked communication is described as leaning into AI-generated media, parody, deepfakes, stylized videos, and unrelated humorous clips.

For investigators and journalists, the issue is not whether memes are “serious.” They are serious because they shape attention before evidence can catch up.

How AI Changes the Evidence Environment

Traditional misinformation often relied on old footage with a new caption. That still happens. The new layer is synthetic media produced at speed and distributed through accounts that audiences may already trust.

The article highlights a more dangerous effect: authentic content becomes easier to dismiss. Mahsa Alimardani of WITNESS describes how verified images of civilian casualties from the Minab school strike were dismissed by some accounts as AI-generated or recycled based on aesthetic impressions rather than forensic method. Claims such as “the lighting is too good” or “it looks staged” spread before investigators confirmed authenticity.

This is the “liar’s dividend”: once synthetic media exists, bad actors can use its existence to reject real evidence. A genuine image can be attacked not because it fails verification, but because it feels too polished, too cinematic, too emotionally convenient.

That is where the public conversation often collapses. People begin treating suspicion as analysis. In OSINT, suspicion is only the starting point. It is not evidence.

A Practical Verification Workflow for AI-Heavy Conflict Content

The goal is not to “detect AI” in isolation. The goal is to assess a claim, preserve the material, separate visual doubt from evidence, and document what can and cannot be established.

1. Capture the claim before analyzing the media

Start with the exact claim attached to the image or video. Record the post URL, account name, timestamp, caption, quoted source, language, and engagement context. Save screenshots and, where lawful and appropriate, preserve the page.

Reason: the same visual can support different claims. A clip may be authentic but miscaptioned. A synthetic image may represent a real event falsely illustrated. A meme may not claim literal accuracy but may still serve propaganda.

Look for: who posted it first in your visible chain, what they claimed, whether they attributed it to an official source, media outlet, witness, or anonymous channel.

Expected result: you should end this step with a claim log, not a verdict.

Common mistake: downloading the image and forgetting the surrounding text. In conflict verification, the caption is often the payload.

2. Separate the object from the narrative

Treat the content in layers: the visual object, the caption, the account, the amplification pattern, and the political narrative.

Reason: AI-generated material may be fake as an image but still attached to a real event. Real footage may be attached to a false location. A parody video may be obviously artificial but still designed to dehumanize or normalize escalation.

Look for: signs that the post is making a factual claim, mocking an enemy, validating an official account, or undermining trust in other evidence.

Expected result: a structured assessment: “The image is unverified,” “the claim is unsupported,” “the account frames it as evidence,” or “the content appears propagandistic rather than evidentiary.”

Common mistake: writing “fake” and stopping there. Fake content can still have real influence.

3. Check provenance, but do not overestimate it

Provenance means understanding where a file came from and whether it has been altered. In theory, standards such as C2PA can help create a chain of custody. In practice, the Reuters Institute article notes that such signals are not yet widely deployed enough to solve fast-moving conflict verification: many phones do not sign images at capture, many models do not embed provenance, and many platforms do not show provenance signals to users.

Reason: absence of provenance is not proof of fabrication. Presence of metadata is not proof of truth.

Look for: original upload context, platform labels, visible edits, repost trails, compression artifacts, and whether credible investigators have preserved earlier versions.

Expected result: a provenance note with limits clearly stated.

Common mistake: treating “no metadata” as suspicious on its own. Social platforms often strip metadata.

4. Triangulate with independent signals

Do not rely on a single artifact. Compare the claim with other available material: different angles, local reports, official statements, known landmarks, timing, weather, shadows, language, uniforms, terrain, or damage patterns.

Reason: Manisha Ganguly, cited in the Reuters Institute article, warns against relying on single artifacts or sources. AI-generated satellite imagery and other synthetic visuals can cosmetically validate official narratives if treated as standalone proof.

Look for: consistency across independent sources, not repetition from the same network. Ten reposts of the same image are not ten sources.

Expected result: a confidence level, not a theatrical debunk.

Common mistake: confusing virality with corroboration.

5. Evaluate the amplification path

Ask how the content moved. Did it travel through state-linked media, official accounts, influencers, anonymous aggregators, diaspora networks, meme pages, or news outlets?

Reason: platforms are not passive pipes. The Reuters Institute article cites Emerson T. Brooking's point that algorithms are central to this kind of information conflict because discovery is not based only on who users follow or what they say they want.

Look for: sudden engagement spikes, repeated captions, coordinated phrasing, official reposts, and whether platform incentives reward the most emotional version.

Expected result: a distribution map showing why the content mattered.

Common mistake: focusing only on authenticity while ignoring reach.

6. Publish uncertainty without creating more fog

When verification is incomplete, say exactly what is known, what is unknown, and what would change the assessment. Avoid aesthetic claims such as "looks AI" unless backed by method.

Reason: the Minab example shows how aesthetic suspicion can become a weapon against authentic documentation. Bad verification language can damage the evidentiary space it tries to protect.

Look for: wording that separates evidence from interpretation.

Expected result: a responsible public note: "We have not verified the location," "the image has not been authenticated," "the event is reported, but this visual cannot yet be linked to it," or "the claim relies on synthetic or unverified imagery."

Common mistake: publishing a confident denial before the evidence is checked.

Practical Scenario: The Synthetic Satellite Image

A state-linked outlet posts a satellite image claiming a military target was destroyed. The image spreads under the authority of a news brand. Some users treat it as OSINT proof.

A responsible workflow would not begin with a dramatic thread. It would begin with preservation: capture the post, image, caption, timestamp, account identity, and engagement. Then separate the claims: Is the outlet claiming the strike happened? Is the image presented as satellite proof? Is the image original, reposted, or unattributed?

Next, compare the visual with other available imagery and reporting. If no independent confirmation exists, the correct result is limited: the post cannot be used as evidence of the strike. If later forensic review shows the image is AI-generated, the finding should say so while avoiding a broader unsupported claim about the entire event.

The expected result is not "nothing happened." The expected result is: "This specific visual does not prove what the post claims."

That distinction protects the investigation from becoming propaganda in reverse.

The Bigger Risk: War as Entertainment

Memes and AI videos also change the emotional frame of conflict. Sam Dubberley of Human Rights Watch warns in the Reuters Institute article that videogame imagery, Lego videos, and memeification can make war feel less real, potentially desensitizing audiences to civilian harm.

For OSINT practitioners, that creates a second duty. Verification is not only about proving whether a file is authentic. It is also about refusing the platform logic that turns civilian harm into shareable spectacle.

The best defensive posture is slow by design: preserve first, separate claims, triangulate, state limits, and resist aesthetic verdicts. In an AI-thickened fog of war, the investigator's job is not to

react faster than propaganda. It is to make the evidence harder to corrupt.