WHEN ARTIFICIAL INTELLIGENCE GENERATED CONTENT BECOMES WEAPON-GRADE COMMUNICATION

Author(s) / Szerző(k):

Dávid Horváth
Ludovika University of Public Service
(Hungary)

E-mail:
david@netokracia.hu

Cite: Horváth, Dávid (2025): When Artificial Intelligence Generated Content Becomes Weapon-Grade Communication. *Mesterséges Intelligencia – interdiszciplináris folyóirat*, VII. évf. 2025/2. szám. 77-98.
Idézés: Doi: https://www.doi.org/10.35406/MI.2025.2.77

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

EP / EE: Ethics Permission / Etikai engedély: KFS/2025/MI0012

Reviewers:
Lektorok:
Public Reviewers / Nyilvános Lektorok:
1. Jakusné Harnos Éva (Ph.D.), Ludovika University of Public Service (Hungary)
2. László Teknős (Ph.D.), Ludovika University of Public Service (Hungary)

Anonymous reviewers / Anonim lektorok:
3. Anonymous reviewer (Ph.D.) / Anonim lektor (Ph.D.)
4. Anonymous reviewer (Ph.D.) / Anonim lektor (Ph.D.)

Abstract

This methodological paper explains how AI-supported, multimodal content can escalate into weapon-grade communication. After outlining the limits of detection-centric responses, it proposes an WGC-matrix that operationalizes credibility hijacking and cognitive overload at the intersection of intent, channel, and cognitive impact. The framework is validated through 12 case studies spanning text, image, audio, and video modalities.

Keywords: weapon-grade communication (WGC), generative AI, credibility hijacking, cognitive overload, hybrid warfare

Disciplines: military science; communication and media studies; criminology; security studies

Absztrakt

AMIKOR A MESTERSÉGES INTELLIGENCIA ÁLTAL GENERÁLT TARTALOM FEGYVERNEK MINŐSÜLŐ KOMMUNIKÁCIÓ

A tanulmány módszertani választ ad arra, miként válik a mesterséges intelligencia által generált multimodális tartalom fegyvernek minősülő kommunikációvá (FMK). A detekció-központú megközelítések korlátai után egy FMK-mátrixot javasolunk, amely a szándék–csatorna–kognitív hatás metszetében operacionalizálja a hitelesség-eltérítést és a kognitív túlterhelést. A keret 12 esettanulmányon kerül validálásra.

Kulcsszavak: fegyvernek minősülő kommunikáció (FMK), generatív mesterséges intelligencia, hitelesség-eltérítés, kognitív túlterhelés, hibrid hadviselés

Diszciplinák: hadtudomány; kommunikáció- és médiatudomány; kriminológia; biztonságpolitika

Content generated by artificial intelligence (AI) is forcing security policy and social science analysis onto a new path because, while the logic of influence is often familiar, the scale and pace of implementation are taking a qualitative leap. It is not just that more messages appear faster: the entire information ecosystem (platform architecture, recommendation systems, moderation, media routines, user psychology) begins to function as an operational space. In this sense, weapon-grade communication cannot be identified by pinpointing "individual lies," but rather by the fact that content production, the perception of authenticity, and the structural conditions for dissemination are shifting together.

The "Firehose of Falsehood" in the age of algorithms: from quantity to quality

According to the classic view of agenda-setting (AS), the media space not only reflects reality, but also thematizes it: it does not tell us what to think, but what to think about

(McCombs, Shaw, 1972). Since the 2010s, this has been happening in a platform environment where the agenda is also organized by algorithmic ranking systems and feedback mechanisms. The "firehose of falsehood" (FoF) propaganda model describes this as high-volume, multi-channel, rapid and repetitive content distribution, often aimed not at persuasion but at overload and disruption of the decision-making environment (Paul, Matthews, 2016). The question is therefore not "how much" false content there is, but when quantity turns into qualitative impact: thematization, emotional attunement, and ultimately the erosion of public trust.

At the same time, platform logic also indicates that mere scaling is not automatically successful: dissemination takes place through the filters of recommendation systems and network structures, so organic reach is not simply a function of the number of posts produced. It can be empirically measured that the network of connections and ranking simultaneously limits and shapes encounters with differing views (Bakshy, Messing, Adamic, 2015). This is where the methodological dilemma becomes clear: traditional frameworks often stick to direct statements and "persuasion," while AI-supported FoF logic often plays on the long-term destruction of common points of reference.

In the military science community, this is organically linked to the doctrinal idea of shaping the information environment (SIE): the goal is not to "make a single message true," but to shape the context in which the recipient is forced to interpret it (UK Ministry of Defence, 2013). The limitation of the diagnosis is therefore that traditional tools typically respond to content-related statements (fact-checking, source analysis, labeling), while the impact at the WGC level often manifests itself in the structure of the agenda, the rhythm of attention, and the erosion of trust.

When we can't even believe our own eyes: the crisis of multimedia credibility

In a multimodal disinformation environment, credibility is rarely determined by a single medium: text, images, audio, and video can all contribute to building a sense of authenticity. One of the central consequences of information disorder is that the boundaries between deliberate deception, misleading context, and manipulated content become blurred, while reactions are rapid and corrections are delayed (Wardle, Derakhshan, 2017). In this sense, credibility hijacking (CH) is not merely technical manipulation, but a communicative operation: the evidential nature (photo/video) lends borrowed authority to narratives that would otherwise be more easily dismissed as text.

The deepfake phenomenon (DF) encapsulates this shift: synthetic media not only spreads false claims, but also makes the status of evidence debatable, which is both a risk to democratic decision-making and a challenge to national security (Chesney, Citron, 2019). The blind spot is therefore twofold: the logic of "detectable fakes" is too narrow in a fast, first-impression-optimized dissemination cycle; and the crisis of credibility consists not

only in "believing a fake video," but also in the recipient later turning to the evidence-based media with general skepticism. In other words, the erosion of trust can still occur even if the specific fake is subsequently exposed.

Blind spots in current research models: why does misinformation slip through the filter?

The research blind spot often arises when we perceive the phenomenon as a short-term attitude change and underestimate the long-term recalibration of the cognitive environment. According to the illusory truth effect (ITE), mere repetition—even with existing knowledge—can increase credibility because processing fluency activates "familiarity = truth" shortcuts (Fazio, Brashier, Payne, Marsh, 2015). If we combine this with FoF tempo and platform rhythm, it becomes clear that AI does not produce false statements, but rather variants and repetitions that work together to lower the credibility threshold.

This is accompanied by the source monitoring (SM) thesis: memory stores not only content but also source cues, but these fade over time, leaving "source-free content" (Johnson, Hashtroudi, Lindsay, 1993). AI-assisted production scales this very phenomenon: the same narrative returns on new platforms, in new languages, and in new styles, while it becomes increasingly difficult to reconstruct where the audience first encountered it. Traditional diagnostic tools are often structurally delayed: by the time the refutation is ready, the source references have already evaporated from the attention cycle.

From this point of view, AI "slip through the filter" is not only a technical detection issue, but also a methodological gap: current models often do not measure consistently enough (i) the ecosystem effect resulting from scaling, (ii) the chains of multimodal authentication, and (iii) the long-term cognitive consequences. This is where the WGC framework comes in: it does not promise "yet another taxonomy," but rather the ability to address the phenomenon simultaneously at the levels of communication theory, criminology, and military science—with the same set of questions, in a comparable manner, with falsifiable statements.

Based on the above diagnosis, AI-assisted influence is not simply "more disinformation," but a shift in the operating parameters of the information ecosystem: scaling, multimodal authentication, and delayed correction together produce the WGC-level effect. It follows that the conceptual framework of a single discipline is not sufficient to grasp the phenomenon: communication theory agenda and attention mechanisms, criminological operational patterns, and military SIE logic must be read within a common framework. The following chapter therefore provides a targeted literature review of the conceptual and methodological approaches used in international and domestic research to address this problem of "more disinformation" — and identifies the blind spots that need to be covered by the subsequent analysis matrix.

Literature map: from propaganda to generative warfare (overview of publications)

Artificial intelligence-supported, multimodal forms of weapon-grade communication (WGC) can be understood if we place the phenomenon not as an isolated technological novelty, but within the historical arc of propaganda, disinformation, and hybrid warfare research. In this arc, generative AI does not rewrite the rules of the game from scratch, but accelerates and extends the mechanisms already described: content production, the illusion of authenticity, and the scaling of dissemination (McCombs, Shaw, 1972; Paul, Matthews, 2016; Horváth, 2023a). The chapter organizes the literature review around three focal points: (i) the connections between classical schools of disinformation and hybrid warfare; (ii) the trap of technological determinism; (iii) the credibility crisis surrounding multimodality and "evidential media."

Classical schools of disinformation and the connection between them and hybrid warfare

Classic models of propaganda (Lasswell, 1927, 1948; Bernays, 1928; Jowett, O'Donnell, 2019) described communication in terms of a goal-channel-effect logic, where the source of influence is mostly institutionally identifiable. In this tradition, disinformation is an operational tool of propaganda: it deliberately distorts or falsifies information to reduce cohesion, weaken morale, and disrupt decision-making (Rid, 2020).

Cold War "active measures" (Kalugin, 1994; Barron, 1983; Andrew, Mitrokhin, 2000) indicated early on that communication is not merely a side effect, but an independent operational dimension. Fake documents, front organizations, and long-term narrative building created an information environment in which the recipient's "reality" could be gradually shifted.

21st-century hybrid warfare literature captures this relationship at the ecosystem level: military and non-military tools, overt and covert operations, and the toolkits of state and non-state actors form a mutually reinforcing system (Hoffman, 2007; Kramer, Speranza, 2017; NATO, 2016). In this framework, fake news is not a matter of press ethics, but a tactical tool that targets perception and decision-making processes (Horváth, 2022; 2023a; 2023b).

Generative AI acts as a qualitative accelerator at three points in this historical arc. (1) It radically reduces the cost and time of content production, making large-volume, multi-channel logics (FoF) easier, more sustainable, and more adaptable to operate (Paul, Matthews, 2016). (2) It reinforces the illusion of authenticity: illusory truth effect (Fazio et al., 2015) and source monitoring errors (Johnson, Hashtroudi, Lindsay, 1993) scale in multimodal, automated environments. (3) Communication in the hybrid space can increasingly be interpreted as a "weapon": it not only accompanies kinetic operations, but often triggers or prepares them (Horváth, 2024).

Subchapter synthesis (short): the classic propaganda–disinformation–hybrid warfare arc provides a stable conceptual basis for the WGC framework; the novelty of generative AI is not its intent, but its scale, speed, and automation of multimodal authentication.

The trap of technological determinism in current research

Discourses surrounding generative AI often slip into technological determinism: as if the tool itself explains disinformation, the crisis of trust, or political polarization. Classic critiques (Winner, 1980; Morozov, 2013) point out that technology operates in a socio-technical space, between business, political, and institutional incentives.

Platform research (Gillespie, 2018; Couldry, Hepp, 2017) also emphasizes that algorithmic ranking and moderation build invisible rules: they amplify certain content and hide others. Generative AI enters this environment: its training data corpora, product strategies, and regulatory gaps together shape the information ecosystem in which WGC-type tactics operate.

Cybersecurity-focused approaches often treat the issue as a detection problem (e.g., watermarking, AI detectors; Kirchenbauer et al., 2023), while some social science analyses treat AI as a homogeneous threat, making little distinction between military, criminological, and commercial logics. The trap of determinism is therefore twofold: (i) it overestimates the "omnipotence" of the tool, (ii) it underestimates the strategic actors and the operational configuration.

This study addresses this from the perspective of the WGC framework and the WGC matrix: it does not ask "what AI can do," but rather who uses it as a weapon, with what intentions, with what set of channels, and under what ecosystem conditions (Horváth, 2023c; Horváth, 2024). The emphasis thus shifts from content tagging and detection to the exploration of structures relevant from a security policy perspective.

Subchapter synthesis (short): the risk of generative AI is not due to the "magic power" of the tool, but to the combination of platform logic + actor incentives + operational goals; the WGC matrix positions itself to analyze this socio-technical configuration.

Multimodality and evidentiary media: research directions in credibility hijacking

For a long time, "evidential media" (photos, audio, video) were the pillars of credibility in the press, law, and political communication. Research in visual culture (Mitchell, 1994; Hariman, Lucaites, 2007) indicated early on that visual evidence is always a framed and interpreted sign, while meme culture (Shifman, 2014) demonstrated the logic of rapid recontextualization.

The legal and national security debate surrounding deepfakes and synthetic media is based on these insights: deepfakes pose a threat to privacy, democracy, and national security (Chesney, Citron, 2019). Furthermore, well-known elements of cognitive mechanisms are reinforced in multimodal environments: repetition can increase credi-

bility (Fazio et al., 2015, while source cues become worn out (Johnson, Hashtroudi, Lindsay, 1993).

Wardle and Derakhshan (2017) consistently treat different types of manipulated visual content within their framework of "information disorder": recontextualization, false captions, modified and generated images. The novelty of generative AI is that the "evidential" content is no longer necessarily a reframing of existing material, but often a document illusion created from scratch.

In my own WGC framework, I have called this "credibility hijacking": a tactic that uses existing trust structures (platforms, media, visual conventions) to elevate manipulated content to the status of evidence (Horváth, 2024). This leads to the main thesis of this chapter: disinformation now operates in a coordinated configuration of text, image, sound, and video, so analysis must also be based on a multimodal, ecosystem-level methodology.

Subchapter synthesis (short): deepfakes and synthetic media are not just "fake content," but a crisis of evidence status; the concept of credibility hijacking links this change to the WGC framework.

Solution attempt: the ecosystem model of "weapon-grade communication" (WGC) and the WGC matrix (introduction of the new method)

Based on the diagnosis in the previous chapters, the problem is not "AI knowledge" but the socio-technical configuration of AI's use as a weapon: the combination of actors, goals, channels, and effects, which is simultaneously a phenomenon of communication theory, criminology, and military science. Therefore, the WGC ecosystem model (Horváth, 2023c) is not a new taxonomy, but an analytical order: it breaks down the same phenomenon at the ecosystem–operational–tactical levels and organizes the interpretation specifically around the security policy question "what is at stake?"

The WGC matrix: the intersection of intent, channel, and cognitive effect

The purpose of the WGC matrix is to view AI-supported content not only as content (what does it claim?), but also as an operational product (why, how, with what consequences?). The three axes of the matrix are therefore: (i) intent (what is the intervention aimed at: destabilization, undermining legitimacy, intimidation, making money, etc.), (ii) channel configuration (open platforms, closed networks, advertising ecosystem, bot/fake profile infrastructure), and (iii) cognitive impact (what psychological and social consequences does it push towards: erosion of trust, paralysis of action, panic, cynicism). With this logic, agenda-setting or the "firehose of falsehood" is not a separate theoretical block, but a description of the intention–channel–effect relationship (see: McCombs–Shaw, 1972; Paul–Matthews, 2016).

While the S–F–C (Situation–Forecast–Control) grid describes the temporal and

management dynamics of cognitive warfare, the WGC matrix introduced in this study performs the structural operationalization of the starting point of this process, the Situation. The WGC matrix thus functions as a diagnostic "magnifying glass": it breaks down a static snapshot of the security environment (S) into dynamic operational elements (intent, channel configuration, cognitive effect), thereby laying the foundation for subsequent forecasting (Forecast) and intervention (Control).

According to the short protocol of methodological operationalization, coding is performed on the three axes of the WGC matrix (intent; channel configuration; cognitive effect). This chapter operationalizes two key outputs with indicator groups: (i) credibility hijacking, (ii) cognitive overload. The rating is done on a three-point scale (low/medium/high), and each value is linked to a documented source chain (platform report/cybersecurity report/OSINT/fact-check/official announcement). We use negative testing: if the intent or channel coordination cannot be substantiated by the source chain, the rating is downgraded or the claim is not upheld.

Operationalization
of credibility hijacking and cognitive overload

The matrix works when the "effect" is not an impression but a consistently marked group of indicators. This study highlights two WGC outputs that are particularly common in multimodal environments: credibility hijacking and cognitive overload.

Authenticity deviation refers to solutions that short-circuit the recipient's "sense of evidence" (e.g., iconic images, "evidential" videos, authoritative voices, institutional branding), while source reconstruction deteriorates; This fits in with the source monitoring theory, according to which source traces fade faster than content memory (Johnson, Hashtroudi, Lindsay, 1993).

By cognitive overload, I mean a state of quantitative scaling (variants, languages, repetition, channel synchronization) in which the recipient encounters not a statement but a persistent noise environment; Therefore, labeling does not occur on a "true/false" axis, but rather by recording the pattern of propagation, repetition, variant formation, and channel crossover.

As a practical designation (to save characters), three-level ratings can be given: low/medium/high for a) intensity of credibility hijacking, b) overload pressure, c) channel coordination; in each case, a source chain is linked to the categories.

Reliability and validity
of the method based on case studies

The reliability of the method is ensured by the fact that the case studies are based on the same analytical framework: the same questions must be answered (actor–intention–channel–effect–reaction) and coded using the same coding rules (Horváth, 2023c). Validity here is not a laboratory result, but a chain of

evidence: every statement is backed by a traceable chain of sources (platform report / cybersecurity report / OSINT / fact-check / official announcement), and the matrix also forces negative testing: if the intention cannot be verified or the channel coordination cannot be substantiated, the rating is downgraded. This also implies the practical meaning of falsifiability: the goal is not to "prove the WGC," but to find out where it does not hold.

Validation: case study-based application – selection logic and analytical framework

Validation is based on case studies: selection is not aimed at representativeness, but at methodological coverage. The logic behind this is: (i) multimodal spectrum (text–image–sound–video). The order of the main categories follows the increase in technological complexity and psychological impact: we move from one-dimensional signals requiring cognitive processing (text) to multimodal forms aimed at visceral perception and unconditional trust (video). (ii) different sets of actors (state influence and criminal exploitation), (iii) different stakes (electoral integrity, critical infrastructure, public trust), (iv) documentation (public reports and verifiable source chain). The next chapter is therefore not a simple "list of cases," but a test of the matrix. The order of the case studies within each category reflects the evolution of the threat: we move from initial technological demonstrations or individual-level "noise" towards system-level operations that pose a strategic risk.

Text generated by artificial intelligence

Case 1: The evolution of the Dragonbridge network from revenge porn to high politics: how Chinese spam became a strategic weapon

The Dragonbridge/Spamouflage prototype is an example of how a low-quality, "spam-like" text stream can become a persistent, ecosystem-level weapon of communication: it does not seek to win a single argument, but rather shapes the background noise of the digital public sphere. The network was described by Graphika as Spamouflage Dragon (Graphika, 2020) and later documented by the Google Threat Analysis Group as DRAGONBRIDGE, with massively removed channels and multilingual content (Google TAG, 2022; Google TAG, 2024); state interests were also confirmed by cybersecurity reports (Mandiant, 2022), and transnational pressure patterns also emerged (Rapid Response Mechanism Canada, 2023–2025). The role of AI here is scaling: translation, variation, rapid reproduction of templates, and "circumventing" moderation filters with many small narrative variations (Google TAG, 2024). In the WGC section, the intention is long-term information environment shaping; the channel is a multi-platform, multilingual, loosely connected network of accounts; the cognitive effect is uncertainty/erosion of trust through cognitive overload and thematization (see: McCombs–Shaw, 1972).

Distribution: YouTube/Facebook/X/TikTok + blogs/forums; massive, varied posts and comments, even with low organic reach, as "constant background noise."

Coding: Intent = strat. thematization/ erosion; Channel = multi-platform–multi-language–fake profile network; Cognitive effect = overload/apathy/erosion of trust; Coordination = medium–high.

Case 2: "Hi Mom, I'm in trouble!" – a grammatically flawless attempt at cheating, or the role of generative AI in digital bank robbery

The AI-based phishing prototype is a case where weapon-grade communication is not spectacular at the campaign level, but rather "scatters" and destroys digital trust at the micro level. According to several recent studies, the persuasiveness of spear phishing messages written by generative models can match or even exceed that of human experts, while the cost and time of production are radically reduced (arXiv, 2024; Electronics, 2024; MDPI, 2025). AI plays a threefold role here: error-free text writing tailored to the brand voice; variant generation (against filters); personalization based on OSINT/ previous leaks. In WGC terms, the intention is to gain access/money in the short term and to erode crisis communication and institutional credibility in the long term; the channel is "everyday" infrastructure (e-mail/SMS/messaging/private messages); the cognitive effect is based on urgency, the exploitation of automatisms based on authority mimicry and trust-building, followed by the establishment of lasting suspicion. In terms of national security, the method can slip into the realm of hack-and-leak if it targets institutions/elections (Canadian Centre for Cyber Security, 2024).

Distribution: email + SMS + messenger + private social media messages; brand imitation, urgent "identification/update" narratives, redirection to clone sites (OTP Bank, 2024).

Coding: Intent = access/money → erosion of trust; Channel = multi-channel direct message; Cognitive effect = urgency/ authority/automatic trust → general suspicion; Coordination = low–medium (depending on campaign type).

Case 3: The anatomy of operation Doppelgänger: when europe's most read newspapers begin to spread russian propaganda

Doppelgänger is a type of textual WGC where the weapon is not quantity but the appearance of credibility: the user reads a pro-Russian narrative in a "familiar media voice." The campaign was named and described by EU DisinfoLab after a network of cloned sites imitating European media outlets was identified in 2022 (EU DisinfoLab, 2022–). The sustainability of the operation is supported by investigative and government analyses: fake articles are published on a daily basis across multiple channels; USCYBERCOM has also described the structure as complex web deception (Correctiv, 2024; USCYBERCOM, 2024). The role of AI here is "style and language alignment": large quantities of quickly produced texts that conform to journalistic style and only shift the narrative at key points (sanctions, support, war fatigue). In WGC terms, the intention is to undermine European cohesion and the legitimacy of NATO/EU policies; the

channel is domain and brand cloning + targeted social media dissemination; the cognitive effect is to disrupt source monitoring and reinforce the "media is not reliable" narrative.

Distribution: cloned news sites + X/Facebook accounts + closed messaging groups; link sharing with search engine and sharing-optimized texts.

Coding: Intent = legitimacy destruction/ cohesion weakening; Channel = domain/ brand cloning + targeted distribution; Cognitive effect = source confusion/trust erosion/war fatigue; Coordination = high.

Subchapter synthesis (short): The common lesson of text-based case studies is that generative AI weaponizes the nature of "text as infrastructure": it not only produces content, but also creates a scalable, variable, and multilingual narrative presence that places a lasting strain on the public's attention and monitoring capacity.

Dragonbridge/Spamouflage exemplifies this model in terms of quantitative superiority and multilingual variant production ("low quality, high volume" background noise), while Doppelgänger is its counterpart: here, the decisive resource is not quantity, but authenticity diversion (brand/domain cloning, press style imitation), i.e., confusing source identification. Generative phishing of the "Hi Mom..." type completes the picture from the perspective of micro-level persuasion: with the disappearance of linguistic errors, the former lay filters (poor spelling, foreign expressions) lose their validity, and the erosion of trust seeps into everyday transactional communication. Together, the three cases illustrate the chain whereby textual WGC is capable of simultaneously shaping the agenda, falsifying sources, and triggering direct action (clicking/referral/abstention), while the long-term "gain" is the erosion of trust through cognitive overload.

Image generated by artificial intelligence

Case 4: Puffer-Jacket Balenciaga Pope:: how he convinced millions with a single AI-mem that Pope Francis had changed his style

The case of the "Balenciaga Pope" prototype is one where the illusion of photorealism becomes an independent WGC component: the image does not state a "fact," yet it massively activates the reflex of visual evidence. The visual was launched in March 2023 (Reddit) and then transferred to major platforms with a loss of context; many interpreted it as a press photo, and recognition of its generated origin typically only occurred after it went viral (CBS News, 2023; The Guardian, 2023; Reuters Fact Check, 2023). The role of AI here is twofold: to produce a photo effect with a low entry threshold and to blur the line between meme and news photo, which in the long run weakens the norm of the photo as evidence (Horváth, 2023).

Summary (WGC brief): Evidence: viral spread + fact checks. Role of AI: photo effect from prompt, acceleration of context loss. Intent: attention/thematization (without harm, but

with a norm-destroying effect). Channel: platform feeds. Cognitive effect: visual evidence heuristics, delayed suspicion. Distribution: Reddit → X/Facebook/Meta. Coordination: low.

Case 5: A few seconds of stock market chaos that wiped out hundreds of billions of dollars with a fake photo

The "explosion photo" near the Pentagon is no longer a pop culture anomaly, but a condensed visual shock that activates market and security reflexes: a single image caused a brief real market turmoil before the refutation stabilized (The Guardian, 2023; Al Jazeera, 2023). The visual most likely originated from a generative image system; the news photo scheme worked, while typical artifacts appeared in the details (UC Berkeley School of Information, 2023). The function of AI here is low-cost panic-mongering: a quickly produced, believable composition that triggers heuristics (danger, urgency, national security) even without a narrative (Paul and Matthews, 2016). At the WGC level, the intent is to distort risk perception in the short term; the channel is platform credibility markers and accounts posing as "news sources"; the cognitive effect is to increase the likelihood of false alarms and overreactions.

Summary (WGC brief): Evidence: official refutation + press analyses + artifact notes. AI role: generating news photo patterns to create "shock." Intent: distortion of risk perception, activation of panic reflex. Channel: credibility markers and viral network. Cognitive effect: false alarm, overreaction, short-term uncertainty. Dissemination: X → media coverage → refutation. Coordination: moderate.

Case 6: Millions of shares for the "perfect" war photo: when reality is too bloody, we draw it pretty

The "All Eyes on Rafah" case is the third stage of generated images: it is not classic disinformation, yet it is WGC -relevant because it scales the toolkit of moral framing and reputational pressure. In May 2024, a declared AI-generated visual spread widely and briefly became a "mandatory" reference point in debates about the humanitarian dimension of the conflict (Al Jazeera, 2024). The power of the image lies not in its replication of reality, but in its "symbolic condensation" of real suffering: it is shareable, emotionally charged, and, through its story format, elevates the visual space of private profiles to the conflict agenda. Debates about the blurring of the line between photo and illustration and the issue of credibility costs have intensified in retrospect (ABC, 2024). At the WGC level, the intention is to shift the moral agenda; the channel is the platform ecosystem and the remixable template; the cognitive effect is normative coercion and source forgetting, which can be quickly followed by counter-images and counter-campaigns (Newsweek, 2024).

Summary (WGC brief): Evidence: platform spread data + press debate. AI role: symbolic condensation, remixability, rapid global scaling. Intent: moral framing, mobilization, reputational pressure. Channel: Meta/X/

TikTok, especially Instagram stories. Cognitive effect: normative coercion, source amnesia, credibility cost risk. Distribution: mass story sharing, cross-platform remixing. Coordination: low–medium.

Subheading synthesis (short): The common denominator of image-based case studies is that generative AI, through photorealism, begins to challenge the status of "visual evidence": the viewer believes what they see before checking the source. The puffy-jacketed Balenciaga pope is the threshold case: as a pop-cultural, "harmless" viral, he normalizes the fact that a believable photo is no longer a guarantee, thus paving the way for later, more targeted operations. The fake photo of the Pentagon explosion already demonstrates the systemic consequences: a single visual impulse can activate market and security reflexes for a short time, thus radically reducing the cost of image-based panic. All Eyes on Rafah shows the third dimension: the generated image does not necessarily function as a "lie," but with moral framing and condensation into a memetic symbol, it can shape reputational pressure and the agenda, while forgetting the source and blurring the line between photo and illustration can generate credibility costs. Together, the three cases indicate that visual WGC is not just about falsification: the decisive factor is that visual content spreads faster than correction and, in the long run, rewrites the routines of controlling the collective image of reality.

Artificial intelligence-generated voice

Case 7: The €220,000 phone call when the boss's voice orders a bank robbery

The case of CEO voice cloning demonstrates the "precision" use of WGC against micro-level, high-value targets: a single call can activate organizational obedience patterns. In a case that made headlines, fraudsters imitated the voice of a senior executive at an energy company and instructed the target to make an urgent transfer; the voice (accent, intonation, familiar tone) was enough to carry out the transaction. Artificial intelligence here "arms" classic social engineering: the source of communication (the boss) becomes falsifiable, and thus deception occurs through the most confidential channel. Its military relevance arises where the corporate decision-making hub is linked to critical infrastructure: a wrong decision can lead not only to financial losses but also to operational disruptions and supply chain distortions, and thus converge with state/quasi-state interests (Horváth, 2023; Horváth, 2024).

Summary (WGC brief): Evidence: incident description + cybersecurity interpretation. AI role: voice model from a short sample, reproduction of a "command" voice. Intent: money/access → distortion of critical decisions. Channel: targeted phone call (outside moderation visibility). Cognitive effect: authority and urgency reflex, activation of organizational obedience. Distribution: not mass; individual, targeted calls. Coordination: low–medium.

Case 8: Weaponizing telephone harassment: "Don't go vote!" – said the Biden clone voice

The 2024 Biden robocall case shows how AI-generated voice crosses from the experimental space into the WGC domain, directly threatening election integrity. Ahead of the New Hampshire primary, automated calls imitating the voice of Joseph R. Biden urged voters to "not go vote" but to "save their vote for November." The voice (tone, rhythm, intonation) acted as an anchor of authority: the recipient tends to identify the recognizable voice with authenticity, even if the content itself is suspicious. The technological innovation is not the robocall as a channel, but the scalability of voice cloning: short sample → reproducible "presidential voice" → mass, variable messaging with minimal human input. In terms of communication theory, this is the voice-based equivalent of action-oriented, repeated messaging (firehose logic); in criminological terms, it is an attempt to manipulate voter behavior; and from a military science perspective, it can be linked to hybrid goals of destabilizing democratic institutions (see: McCombs–Shaw, 1972; Paul and Matthews, 2016; Horváth, 2022; Horváth, 2024).

Summary (WGC brief): Evidence: official investigation + press/fact-check confirmation. AI role: voice cloning and scalable robocall variants. Intent: to reduce participation, undermine the process. Channel: telephone network + automated calling system. Cognitive effect: authority reflex, false authenticity, timing pressure. Dissemination: robocall → online press + social media reaction/refutation. Coor-dination: moderate.

Case 9: 48 Hours before the polls close: the audio recording that poisoned Slovakia

The deepfake audio recording before the 2023 Slovak parliamentary elections is an example of targeted, regional application: even in a small country, it can interfere with the electoral process if the timing follows the logic of an "October surprise." Close to the campaign silence period, an audio recording was released in which Michal Šimečka and Monika Tódová allegedly discuss "manipulating" the election; its rapid spread was later interpreted by fact-checking and expert analyses as AI-generated manipulation. The technical role here is no longer merely the reproduction of sound: authenticity is reinforced by the imitation of discursive style (phrasing, tempo, "local" speech patterns), i.e., the model produces a "character" who credibly enters into a politically charged dialogue. The core of communication theory is the undermining of source credibility (what did I actually hear?), which quickly generates confusion and distrust; criminologically, this is targeted intervention in the electoral process; militarily, it is the "contamination" of the Central and Eastern European information space (McCombs–Shaw, 1972; Paul and Matthews, 2016; UNESCO, no year).

Summary (WGC brief): Evidence: rapid viral spread + fact-checking + expert evaluations. AI role: voice cloning + discursive style imitation. Intent: undermining trust, weakening legitimacy, shock before the campaign

silence period. Channel: social media spread + forwarding chains. Cognitive effect: source uncertainty, reputational damage, rapid polarization. Distribution: platforms + messengers, forwarding in closed groups. Coordination: medium–high.

Subchapter summary (short): The common denominator of the three voice cases is that the attack does not require a new platform, but rather weaponizes everyday voice channels: robocalls, social media forwarding, targeted phone calls. The medium of voice is critical because the heuristic of "recognizable voice = authenticity" kicks in faster than source verification, especially under time pressure. As a result, refutation is often structurally delayed: some of the damage/distortion occurs before the correction.

Artificial intelligence-generated video

Case 10: This is not Morgan Freeman: demonstration deepfake and the proof video crisis

The "This is not Morgan Freeman" demonstration deepfake video is the "zero milestone" of moving image manipulation: it is not an operational campaign, but a technological demonstration that simultaneously teaches the audience that "AI can already do this" and normalizes the idea that faces and voices can be reproduced at any time. The video declares from the outset that we are watching a deepfake (Avid Open Access, n.d.; HP, 2024), yet its effect is WGC-like: it shatters the layperson's sense of security that video and audio are "proof." Subsequent empirical results support this phenomenon: the debate surrounding deepfakes is not only technical, but also concerns a crisis of "epistemic trust" (Twomey, 2023), and we perform poorly in recognizing them even with instructions (Somoray, Miller, 2023).

In terms of WGC, this case is not about direct influence, but about changing the environment: according to the logic of "shaping the information environment," social perception thresholds and credibility criteria are rewritten, making subsequent political/commercial/criminal applications more effective (McCombs, Shaw, 1972; Paul, Matthews, 2016; Horváth, 2023).

Summary (WGC brief): Evidence: demonstration video + professional/educational references + empirical recognition results. AI role: photorealistic face + voice cloning + lip-syncing. Intent: normalization/"showcasing capabilities" → preparing for the erosion of credibility standards. Channel: public platforms (YouTube, professional portals, embeds). Cognitive effect: weakening of the status of the video as evidence, epistemic uncertainty. Distribution: mostly organic. Coordination: low.

Case 11: Karikó Katalin on the crypto exchange: deepfake advertisements causing billions in damages

The second pattern of deepfake videos is the world of investment and crypto fraud: fake offers are "legitimized" with the faces and voices of well-known public figures. According to cybersecurity and regulatory

reports, these campaigns are increasingly operating with AI-based videos, linked to organized boiler room infrastructure (Unit 42, 2024). Official warnings and reports also confirm this phenomenon (e.g., Government of Western Australia, 2025; ACCC Scamwatch, 2024), pointing out that video is no longer just an illustration, but the main tool for gaining trust.

At the technical-operational level, this is a multimodal pipeline: LLMs write the text variants, voice cloning reproduces the "expert" voice, and the deepfake engine puts together the facial expressions and synchronization. The "firehose" also works at the format level here: the same lie returns as a video, post, story, or advertisement (Paul, Matthews, 2016; Horváth, 2023). Criminologically, the pattern is linked to organized crime, money laundering chains, and call centers; at the systemic level, it undermines trust in financial communication. Its military relevance lies in the fact that it increases the vulnerability of critical financial infrastructures to information attacks and can generate longer-term instability.

Summary (WGC brief): Evidence: cybersecurity/authority reports + identified network patterns. AI role: cloning of celebrity/authority faces and voices + text variant generation + multimodal coordination. Intent: making money and gaining access → erosion of trust in the financial sector. Channel: video ads and short videos (Meta, TikTok, YouTube, messaging apps). Cognitive impact: authority heuristic, cognitive overload, "everything is suspicious" trust defense. Distribution: repetitive campaigns reinforced by recommendation systems. Coordination: high.

Case 12: Even poor-quality lies can kill: anatomy of Zelensky's fake "Lay down your arms!" video

The deepfake attributed to Zelensky (March 2022) made the military science stakes of moving image manipulation explicit: this is no longer a game of reputation, but an attempt to distort the chain of command, morale, and situation assessment (see: Euronews, 2022; Smalley, 2022). Although technically immature, it went down in history as the first deepfake video attributed to a head of state to appear in an active war zone. The key to the chronology is resilience: on the Ukrainian side, pre-bunking warnings, rapid debunking, and an authentic refutation video (The Guardian, 2022), which reduced the immediate impact, while at the same time the example immediately entered international professional and legal discourse (Kuźnicka-Błaszkowska, 2025).

In the WGC framework, the video combines the TEXT and IMAGE case types: the narrative of surrender + iconic visual frame (presidential setting, familiar background) + the illusion of moving image authenticity together target the most sensitive audience. The relevant lesson is that early, poor quality is not harmless: technology improves quickly, but the social immune system is slower (Desjardins, 2021).

Summary (WGC brief): Evidence: incident reports + removal/refutation + literature/legal processing. AI role: face and facial

expression manipulation + lip-syncing + imitation of the visual frame of a "presidential speech." Intent: surrender/resistance narrative → undermining morality and command structure. Channel: hacked news portals + social platforms (Facebook/X) + possible stream slippage. Cognitive effect: misjudgment of the situation, panic/breakdown, erosion of trust. Distribution: multi-channel, with rapid removal. Coordination: medium–high.

Subheading synthesis (short): The three video cases follow the same pattern: (1) demonstration deepfakes as a "basic motif" of normalization (crisis of evidence videos), (2) deepfake advertisements used in organized crime as industrial scaling (multimodal firehose), (3) war deepfakes as direct operational relevance (morale/command/credibility). Their common WGC point is that moving images simultaneously convey the illusion of visual evidence and audio authenticity, so refutation is structurally delayed: the effect often runs its course before debunking, even if the manipulation is easily recognizable in hindsight.

Conclusion: new ways of defense

The essence of multimodal influence accelerated by generative AI is not merely "more false statements," but a structural shift in the ecosystem of attention and credibility: the agenda (what we talk about), dissemination (how it reaches us), and source perception (who we believe) all become manipulable at once (McCombs, Shaw, 1972; Paul, Matthews, 2016; Johnson et al., 1993). In this framework, the new way to defend ourselves is not a single "better detector," but an analytical and decision-support approach that measures risk at the intersection of intent, channel, and cognitive effect, thus capturing it not in binary terms (true/false) but in terms of operational relevance.

What does the matrix offer that detection-centric thinking does not?

The detection-focused approach typically responds to the "authenticity" of content, while the WGC matrix focuses on the function of the attack: what it aims to achieve, through which channel ecosystem, and what cognitive vulnerability it targets (Paul, Matthews, 2016; Wardle, Derakhshan, 2017). The framework therefore works even when the attack is not a "classic lie" but rather context and identity falsification or "evidential media" authenticity diversion (Chesney, Citron, 2019; Johnson et al., 1993). The added value is practical: it makes cases with different modalities (text–image–sound–video) comparable and helps to plan defenses not at a single point (content control) but across the entire chain (distribution logic, attention cycle, source labeling, institutional response time) (Bakshy et al., 2015; Fazio et al., 2015).

Critical infrastructure, election integrity, public trust: implications for application

In the case of critical infrastructure, the stakes are often not mass persuasion, but confidence-eroding disruption: even a small

amount of "evidential" content can be enough to cause customer panic, service and call center overload, and reputational damage. The advantage of the matrix is that it operationalizes the damage mechanism (cognitive overload, source amnesia, illusory truth effect) from early signs, rather than treating it as a retrospective narrative explanation (Fazio et al., 2015; Johnson et al., 1993).

In election integrity, "last mile" vulnerability—especially voice and video manipulation with a short reaction window—is a risk where legal and platform-level frameworks (e.g., DSA) are necessary but not sufficient on their own: the decisive factors are operational timing and psychological impact management, i.e., the ability to quickly pre-bunk/debunk and manage channel coordination (EU, 2022; UNESCO, 2024).

At the level of public trust, the challenge is to ensure that the public does not merely doubt individual statements, but also the minimum conditions of shared reality. The matrix therefore also "translates" for strategic decision-making: it classifies the content event as a security risk and identifies where the greatest systemic damage is likely to occur (Horváth, 2022; Horváth, 2024, 2025).

Limitations and further research directions

The proposed matrix is not a detector test or an automated truth-checking system: it is based on documented sources, a reconstructable source chain, and transparent coding, and therefore depends on data availability and analyst consistency (Wardle, Derakhshan, 2017). A further limitation is that direct measurement of cognitive effects (e.g., source amnesia, illusory truth) is often only possible with proxy indicators, which requires strict validation discipline (Johnson et al., 1993; Fazio et al., 2015).

The next steps should be: (i) developing inter-coder reliability procedures and a labeling standard supported by a list of examples, (ii) comparing the matrix with platform regulation compliance logics and case types (EU, 2022), and (iii) expanding the targeted, comparable case study corpus in the CI–choice–public trust triangle so that the model can maintain falsifiable statements in different operational environments (see: UNESCO, 2024; Europol, 2022).

References

ABC News (2024). What is 'All Eyes on Rafah'? *ABC News* (Australia). Downloaded: 2025.11.26. Web: https://www.abc.net.au

ACCC Scamwatch (2024). *Investment scams involving deepfakes*. ACCC Australia. Downloaded: 2025.11.26. Web: https://www.scamwatch.gov.au

Al Jazeera (2023). Fake Pentagon explosion photo goes viral, markets dip. *Al Jazeera*. Downloaded: 2025.11.26. Web: https://www.aljazeera.com

Al Jazeera (2024). 'All Eyes on Rafah': The AI image shared by millions. *Al Jazeera*. Downloaded: 2025.11.26. Web: https://www.aljazeera.com

Andrew, C. és Mitrokhin, V. (2000). *The Sword and the Shield: The Mitrokhin Archive and the Secret History of the KGB*. Basic Books, New York.

arXiv (2024). *Generative AI and Phishing / Spear-phishing studies*. arXiv preprint. Downloaded: 2025.11.26. Web: https://arxiv.org

Avid Open Access (n. d.). *This is not Morgan Freeman - Deepfake demonstration.* Avid Open Access. Downloaded: 2025.11.26. Web: https://avidopenaccess.org

Bakshy, E., Messing, S. és Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. DOI: https://www.doi.org/10.1126/science.aaa1160

Barron, J. (1983). KGB Today: The Hidden Hand. *Reader's Digest Press*, New York.

Bernays, E. L. (1928). *Propaganda*. Horace Liveright, New York.

Canadian Centre for Cyber Security (2024). *Cyber threats to democratic processes*. Government of Canada. Downloaded: 2025.11.26. Web: https://cyber.gc.ca

CBS News (2023). Pope Francis 'puffer jacket' fake photo goes viral. CBS News. Downloaded: 2025.11.26. Web: https://www.cbsnews.com

Chesney, R. és Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6), 1753–1819. DOI: https://www.doi.org/10.15779/Z38RV0D15J

Correctiv (2024). *Doppelgänger investigation*. Correctiv.org. Downloaded: 2025.11.26. Web: https://correctiv.org

Couldry, N. és Hepp, A. (2017). *The Mediated Construction of Reality*. Polity Press, Cambridge.

Desjardins, J. (2021). The rapid evolution of deepfake technology. *Visual Capitalist*. Downloaded: 2025.11.26. Web: https://www.visualcapitalist.com

Electronics (2024). AI in Phishing detection and generation. *Electronics*, 13(5). DOI: https://www.doi.org/10.3390/electronics13050000

EU DisinfoLab (2022). Doppelgänger - Media Cloning Campaign. *EU DisinfoLab Report*. Downloaded: 2025.11.26. Web: https://www.disinfo.eu

Euronews (2022). Deepfake video of Zelensky calling on Ukrainians to surrender. *Euronews*. Downloaded: 2025.11.26. Web: https://www.euronews.com

European Union (2022). Digital Services Act (DSA) – Regulation (EU) 2022/2065. *EUR-Lex*. Downloaded: 2025.11.26. Web: https://eur-lex.europa.eu

Europol (2022). Facing the Challenge of Deepfakes in Cybercrime and Disinformation. Europol Innovation Lab, *The Hague*. Downloaded: 2025.11.26. Web: https://www.europol.europa.eu

Fazio, L. K., Brashier, N. M., Payne, B. K. és Marsh, E. J. (2015). Knowledge Does Not Protect Against the Illusory Truth Effect. *Journal of Experimental Psychology: General*, 144(5), 993–1002. DOI:

https://www.doi.org/10.1037/xge0000098

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven.

Google Threat Analysis Group (2022). *DRAGONBRIDGE activity. Google TAG Report*. Downloaded: 2025.11.26. Web: https://blog.google/threat-analysis-group

Google Threat Analysis Group (2024). *DRAGONBRIDGE updates. Google TAG Report*. Downloaded: 2025.11.26. Web: https://blog.google/threat-analysis-group

Government of Western Australia (2025). *ScamNet alerts on deepfake investment scams*. WA Government. Downloaded: 2025.11.26. Web: https://www.scamnet.wa.gov.au

Graphika (2020). Spamouflage Dragon: Spamouflage network. *Graphika Report*. Downloaded: 2025.11.26. Web: https://graphika.com

Hariman, R. és Lucaites, J. L. (2007). *No Caption Needed: Iconic Photographs, Public Culture, and Liberal Democracy*. University of Chicago Press, Chicago.

Hoffman, F. G. (2007). *Conflict in the 21st Century: The Rise of Hybrid Wars*. Potomac Institute for Policy Studies, Arlington.

Horváth, D. (2022). Pandemic and Infodemic – Which One Is More Dangerous? *Acta Universitatis Sapientiae, Communicatio*, 9(1), 125–139. DOI: https://www.doi.org/10.2478/auscom-2022-0009

Horváth, D. (2023a). Az álhír fogalma és helye a hibrid hadviselésben. *Nemzet és Biztonság*, 2023/1, 45–60.

Horváth, D. (2023b). Az álhír és a hibrid hadviselés kapcsolata. In *A hadtudomány és a 21. század*. Ludovika Egyetemi Kiadó, Budapest. pp. 88–102.

Horváth, D. (2023c). A fegyvernek minősülő kommunikáció (FMK) ökoszisztéma-modellje és FMK-mátrix. *Hadtudományi Szemle*, 16(3), 21–45.

Horváth, D. (2024). Az álhír, vagyis a fegyvernek minősülő kommunikációs taktika mint aktuális hadtudományi kihívás. In Gazdag, F., Padányi, J. és Tálas, P. (eds.). *A hadtudomány aktuális kérdései* 2023. Ludovika Egyetemi Kiadó, Budapest. pp. 151–168.

Horváth, D. (2025). A COVID–19-válság szerepe a fegyvernek minősülő kommunikáció új dimenzióinak kialakulásában. In *Lélektan és hadviselés – interdiszciplináris folyóirat*, 2025/2. szám (megjelenés alatt).

HP (2024). Deepfake risks and demonstrations. *HP Wolf Security*. Downloaded: 2025.11.26. Web: https://threatresearch.ext.hp.com

Johnson, M. K., Hashtroudi, S. és Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3–28. DOI: https://www.doi.org/10.1037/0033-2909.114.1.3

Jowett, G. S. és O'Donnell, V. (2019). *Propaganda & Persuasion*. SAGE Publications, Thousand Oaks.

Kalugin, O. (1994). *The First Directorate: My 32 Years in Intelligence and Espionage Against the West*. St. Martin's Press, New York.

Kirchenbauer, J. és tsai (2023). *A Watermark for Large Language Models*. arXiv preprint. Downloaded: 2025.11.26. Web: https://arxiv.org/abs/2301.10226

Kramer, F. D. és Speranza, L. D. (2017). *Meeting the Russian Hybrid Challenge*. Atlantic Council, Washington D.C.

Kuźnicka-Błaszkowska, D. (2025). Legal implications of wartime deepfakes. *International Journal of Law and IT* (Kézirat).

Lasswell, H. D. (1927). *Propaganda Technique in the World War*. K. Paul, Trench, Trubner & Co., London.

Lasswell, H. D. (1948). The Structure and Function of Communication in Society. In Bryson, L. (Ed.). *The Communication of Ideas*. Harper and Row, New York. pp. 37–51.

Mandiant (2022). DRAGONBRIDGE/ Spamouflage activity. *Mandiant Threat Intelligence*. Downloaded: 2025.11.26. Web: https://www.mandiant.com

McCombs, M. E. és Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 36(2), 176–187. DOI: https://www.doi.org/10.1086/267990

MDPI (2025). Emerging threats in AI phishing. *Future Internet*, 17(1). DOI: https://www.doi.org/10.3390/fi17010000

Mitchell, W. J. T. (1994). *Picture Theory: Essays on Verbal and Visual Representation*. University of Chicago Press, Chicago.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs, New York.

NATO (2016). *Hybrid Warfare / Hybrid Threats*. NATO Archives. Downloaded: 2025.11.26. Web: https://www.nato.int

Newsweek (2024). 'All Eyes on Rafah' AI image sparks controversy. *Newsweek*. Downloaded: 2025.11.26. Web: https://www.newsweek.com

OTP Bank (2024). *Tájékoztatás adathalász kísérletekről. OTP Bank hivatalos közlemény*. Downloaded: 2025.11.26. Web: https://www.otpbank.hu

Paul, C. és Matthews, M. (2016). *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation, Santa Monica. DOI: https://www.doi.org/10.7249/PE198

Rapid Response Mechanism Canada (2023–2025). *RRM Reports on Foreign Interference*. Government of Canada. Downloaded: 2025.11.26. Web: https://www.canada.ca

Reuters Fact Check (2023). Fact Check: Image of Pope Francis in puffer jacket is AI generated. *Reuters*. Downloaded: 2025.11.26. Web: https://www.reuters.com

Rid, T. (2020). *Active Measures: The Secret History of Disinformation and Political Warfare*. Farrar, Straus and Giroux, New York.

Shifman, L. (2014). *Memes in Digital Culture*. MIT Press, Cambridge.

Smalley, I. (2022). *The first weaponized deepfake in war?* Tech Policy Press. Downloaded: 2025.11.26. Web: https://techpolicy.press

Somoray, K. és Miller, D. J. (2023). Detection of deepfake videos. *Computers in Human Behavior*, 145. DOI: https://www.doi.org/10.1016/j.chb.2023.107765

The Guardian (2022). Deepfake of Zelensky surrender removed from social media. *The Guardian*. Downloaded: 2025.11.26. Web: https://www.theguardian.com

The Guardian (2023). Fake photo of Pentagon explosion causes brief market dip. *The Guardian*. Downloaded: 2025.11.26. Web: https://www.theguardian.com

The Guardian (2023). Pope Francis in a puffer jacket: AI-generated image fools the internet. *The Guardian*. Downloaded: 2025.11.26. Web: https://www.theguardian.com

Twomey, J. (2023). The Epistemic Threat of Deepfakes. *Synthese*, 201(3). DOI: https://www.doi.org/10.1007/s11229-023-04000-x

UC Berkeley School of Information (2023). *Analysis of the fake Pentagon explosion image artifacts*. UC Berkeley. Downloaded: 2025.11.26. Web: https://www.ischool.berkeley.edu

UK Ministry of Defence (2013). *Joint Doctrine Note 2/13: Information Operations*. Ministry of Defence, Shrivenham. Downloaded: 2025.11.26. Web: https://assets.publishing.service.gov.uk/media/5a7bf3f3ed915d74e33f2cbb/20130828-JDN_2_13_Information_Operations.pdf

UNESCO (2024). *Elections and the Threat of AI-generated Disinformation: Case Study – Slovak Parliamentary Elections 2023*. UNESCO. Downloaded: 2025.11.26. Web: https://www.unesco.org

Unit 42 (2024). Deepfake scams and boiler room operations. *Palo Alto Networks / Unit 42*. Downloaded: 2025.11.26. Web: https://unit42.paloaltonetworks.com

USCYBERCOM (2024). Complex Web Deception / Doppelgänger. *US Cyber Command*. Downloaded: 2025.11.26. Web: https://www.cybercom.mil

Wardle, C. és Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe, Strasbourg. Downloaded: 2025.11.26. Web: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1), 121–136.