Facta Simonidis
19, nr 1 (2026): 133–168

NAUKI O BEZPIECZEŃSTWIE
SECURITY STUDIES

Mikołaj Kabała
Jan Kochanowski University of Kielce
e-mail: mikolajkabala@interia.pl
https://orcid.org/0009-0001-6113-6548

Artificial Intelligence and Security in the Age of Realfakes

Sztuczna inteligencja a bezpieczeństwo w erze *realfake*

Summary

This article analyzes the significance of security in the context of AI-generated images and clarifies the concept of "realfake" as a category describing synthetic or hybrid visuals that mimic the evidentiary power of documentary imagery. In the study, 217 participants evaluated 18 stimuli – real photographs, DALL-E images, and images stylized in the Ghibli aesthetic – in terms of authenticity, sharing propensity, and perceived risk. A repeated-measures ANOVA with Greenhouse–Geisser correction revealed significant differences between conditions. Real photographs were rated as the most authentic, DALL-E images as the least authentic, and Ghibli-style images generated the highest sense of uncertainty and risk. The results justify combining detection, provenance solutions, and media literacy education focused not only on deception but also on the destabilization of the evidentiary environment.

Keywords: realfake; deepfake; artificial intelligence; synthetic media; information security; disinformation; repeated-measures ANOVA

Streszczenie

Artykuł analizuje znaczenie bezpieczeństwa w kontekście obrazów generowanych przez AI oraz doprecyzowuje pojęcie „realfake" jako kategorii opisującej syntetyczne lub hybrydowe wizualia naśladujące moc dowodową obrazu dokumentalnego. W badaniu 217 uczestników oceniało 18 bodźców – fotografie rzeczywiste, obrazy DALL-E i obrazy stylizowane na estetykę Ghibli – pod kątem autentyczności, skłonności do udostępniania i postrzeganego ryzyka. ANOVA z powtarzanym pomiarem i korektą Greenhouse'a–Geissera wykazała istotne różnice między warunkami. Fotografie rzeczywiste uznano za najbardziej autentyczne, obrazy DALL-E za najmniej autentyczne, a obrazy stylizowane na Ghibli generowały najwyższe poczucie niepewności

nadesłano: 17 września 2025 | zaakceptowano: 21 kwietnia 2026

ISSN: 1899-3109; eISSN: 2956-4085
https://doi.org/10.56583/fs.2927

i ryzyka. Wyniki uzasadniają łączenie detekcji, rozwiązań proweniencyjnych i edukacji medialnej ukierunkowanej nie tylko na oszustwo, lecz także na destabilizację środowiska dowodowego.

Słowa kluczowe: realfake; deepfake; sztuczna inteligencja; media syntetyczne; bezpieczeństwo informacyjne; dezinformacja; ANOVA z powtarzanym pomiarem

Introduction

The rapid diffusion of generative artificial intelligence has transformed the visual layer of disinformation. Earlier debates focused mainly on deepfakes – understood as technically sophisticated manipulations of faces, voices, or videos.[1] That vocabulary remains useful, but it is no longer sufficient. Recent scholarship increasingly frames deepfakes not only as a technical artifact, but also as a broader governance, communication, and organizational challenge.[2] Much of the current threat landscape involves not only identity swaps or face replacements, but also fully synthetic images, style-transferred scenes, hybrid composites, and AI-assisted visual fabrications that circulate as if they were documentary traces of real events. The central security problem is, therefore, not only falsification as such but the corruption of the evidentiary environment in which political, social, and institutional judgments are made.

Accordingly, the security significance of synthetic media lies not only in whether audiences can or cannot detect a fabricated item; it also lies in whether manipulated imagery degrades trust, multiplies verification costs, and complicates the attribution of responsibility in crisis communication, electoral conflict, and war-related reporting.[3] From that perspective, empirical reception is not a peripheral issue, but a core variable for assessing security risk.

Experimental research also suggests that the perception of synthetic media is shaped by more than simple true–false discrimination. AI-synthesized faces may be perceived as indistinguishable from real ones and even as more trustworthy.[4] Com-

1 Yisroel Mirsky and Wenke Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys* 54, no. 1 (2021): 7:1, 7:3–7:4, article 7, https://doi.org/10.1145/3425780.

2 Jan Kietzmann et al., "Deepfakes: Trick or Treat?," *Business Horizons* 63, no. 2 (2020): 136, https://doi.org/10.1016/j.bushor.2019.11.006.

3 Cristian Vaccari and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society* 6, no. 1 (2020): 1, https://doi.org/10.1177/2056305120903408.

4 Sophie J. Nightingale and Hany Farid, "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy," *Proceedings of the National Academy of Sciences* 119, no. 8 (2022): 2, https://doi.org/10.1073/pnas.2120481119.

parative studies indicate that neither unaided judgment nor automated classification fully resolves the detection problem.[5] Content warnings do not uniformly improve recognition[6] and, in some contexts, may themselves generate unintended side effects.[7] The interpretive effects of synthetic visuals are also conditioned by image modality, photorealism, and the distinction between photography and AI-generated depiction.[8]

This article develops a more rigorous conceptualization of "realfake" and situates it within the literature on synthetic media, democratic resilience, and information security. Methodologically, the inferential analysis treats repeated judgments from the same respondent as repeated measures rather than as independent observations. In reporting terms, the results section provides descriptive tables, confidence intervals, ANOVA details, effect sizes, *post hoc* comparisons, and graphical presentation.

The main objective of this article is to determine whether different classes of synthetic and non-synthetic images generate significantly different reception profiles and to clarify what those differences imply for conceptualizing realfake as a security-relevant category.

The main research problem may be formulated as follows: how do real photographs, DALL-E outputs, and Ghibli-style images differ in perceived authenticity, declared willingness to share, and perceived risk, and what do these differences reveal about the security logic of realfakes?

The main hypothesis is that these three stimulus classes produce significantly different reception patterns and that the most security-relevant effect of realfakes lies not only in successful deception, but also in the production of evidentiary uncertainty.

I argue that realfake should be defined functionally rather than purely technically: it is media that perform reality in order to influence interpretations of reality. Under that definition, the most consequential cases are not necessarily those that are perfectly deceptive; they may also be those that generate enough uncertainty to fragment shared standards of verification.

---

5 Matthew Groh et al., "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds," *Proceedings of the National Academy of Sciences* 119, no. 1 (2021): 7, https://doi.org/10.1073/pnas.2110013119.

6 Andrew Lewis et al., "Deepfake Detection with and Without Content Warnings," *Royal Society Open Science* 10, no. 11 (2023): 2, https://doi.org/10.1098/rsos.231214.

7 John Ternovski et al., "Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments," *Journal of Online Trust and Safety* 1, no. 2 (2022): 8, https://doi.org/10.54501/jots.v1i2.28.

8 Liv Hausken, "Photorealism Versus Photography: AI-Generated Depiction in the Age of Visual Disinformation," *Journal of Aesthetics & Culture* 16, no. 1 (2024): 9–10, https://doi.org/10.1080/20004214.2024.2340787.

1. From Deepfake to Realfake – a Security-Oriented Conceptual Framework

The term "deepfake" remains valuable, yet it is narrower than the problem addressed in this article. Recent scholarship increasingly treats synthetic and manipulated media as problems of democratic communication and public trust rather than as purely technical artifacts.[9] In security studies, they can also be located within the broader environment of cognitive confrontation and information warfare.[10] From the perspective of influence operations, synthetic media may function as instruments of persuasion and psychological manipulation rather than merely as carriers of falsehood.[11] The proposed category of "realfake" is therefore intended to capture not only face-swapping or audiovisual substitution, but also synthetic and hybrid visual materials that derive persuasive force from documentary conventions.

For that reason, the article defines realfake as follows: a synthetic, manipulated, or hybrid representation intentionally designed to simulate the evidentiary force of documentary media and thereby shape perception, attribution, or decision-making about real-world events. This definition introduces four constitutive dimensions:

1. Indexical mimicry: the content is received not merely as imaginative fiction, but as a plausible trace of something that happened.
2. Operational orientation: the image matters because it can redirect attention, blame, fear, trust, or action.
3. Verification asymmetry: the fabricated item can spread more quickly than institutional verification can stabilize the record.
4. Ambiguity yield: the item may remain effective even when it does not fully convince, because it can still fragment consensus and delay judgment.

This formulation clarifies two boundaries. Not every deepfake is a realfake: clearly parodic or overtly aesthetic synthetic media may be technologically sophisticated without making a documentary claim. Conversely, not every realfake needs to be a deepfake in the narrow sense: "cheapfakes," staged imagery, context-shifted authentic photographs, or AI-assisted composites may all qualify if they mimic evidentiary

9 Maria Pawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society* 1 (2022): 12, https://doi.org/10.1007/s44206-022-00010-6.

10 Adrian Mitręga, "Wojna poznawcza we współczesnym środowisku bezpieczeństwa," *Annales Universitatis Paedagogicae Cracoviensis: Studia de Securitate* 13, no. 2 (2023): 122, https://doi.org/10.24917/26578549.13.2.7.

11 Pawel Zegarow and Ewelina Bartuzi, "Deepfake Influence Tactics through the Lens of Cialdini's Principles: Case Studies and the DEEP FRAME Tool Proposal," *Applied Cybersecurity & Internet Governance* 3, no. 2 (2024): 292, https://doi.org/10.60097/acig/201147.

reality. The point is therefore functional rather than taxonomic. "Realfake" names a class of security-relevant media operations whose defining feature is the simulation of documentary reality.

That distinction strengthens the article's disciplinary contribution. It connects the analysis to debates on transparency obligations under the AI Act,[12] the limits of ex-ante regulation,[13] and the emerging need to regulate deepfakes in international law.[14] In the Polish legal literature, attention is also being paid to the legality of creating and disseminating deepfakes after the adoption of the AI Act.[15] In security studies, "realfake" is thus treated here as an operational category rather than as a rhetorical synonym for "deepfake."

This conceptual move is not rhetorical ornament. It generates a testable implication. If realfake matters because it occupies a zone between obvious fabrication and unquestioned documentary truth, then audience responses should not collapse into a simple binary of "believed" versus "disbelieved." Instead, one should expect differentiated reception profiles and, in particular, a meaningful middle zone of hesitation. This expectation is consistent with recent arguments that the effects of AI mis- and disinformation should be analyzed through human reactions to ambiguity rather than through technology alone.[16] The empirical section below evaluates precisely that point.

2. Materials and Methods

2.1. Design and Sample

The study employed a cross-sectional, anonymous online survey with a within-subject repeated-measures design. Data were collected in April 2025 through open online

12 Mateusz Łabuz, "Regulating Deep Fakes in the Artificial Intelligence Act," *Applied Cybersecurity & Internet Governance* 2, no. 1 (2023): 267, https://doi.org/10.60097/acig/162856.

13 Mateusz Łabuz, "Deep Fakes and the Artificial Intelligence Act—an Important Signal or a Missed Opportunity?," *Policy & Internet* 16, no. 4 (2024): 784–785, https://doi.org/10.1002/poi3.406.

14 Dominika Kuźnicka-Błaszkowska and Nadiya Kostyuk, "Emerging Need to Regulate Deepfakes in International Law: The Russo-Ukrainian War as an Example," *Journal of Cybersecurity* 11, no. 1 (2025): 2, https://doi.org/10.1093/cybsec/tyaf008.

15 Julia Bernacka, "Problematyka prawna technologii deepfake – analiza legalności tworzenia i rozpowszechniania deepfake'ów po uchwaleniu AI Act," *Prawo i Więź* 58, no. 5 (2025): 673, https://doi.org/10.36128/hg1acq35.

16 Mateusz Łabuz and Christopher Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About? Shifting Away from Technology Toward Human Reactions," *Politics & Policy* 52, no. 4 (2024): 875, https://doi.org/10.1111/polp.12617.

distribution of the questionnaire link in Polish Facebook groups dedicated to survey participation. Participation was voluntary, unpaid, and anonymous; informed consent was obtained electronically before respondents began the questionnaire. The analytical sample comprised 217 respondents, including five participants under 18 years of age. All participants evaluated the same fixed sequence of stimuli. The order of presentation was identical for all respondents and was not randomized. The questionnaire consisted of four sections: demographic information (sex, age, education, and place of residence), image-authenticity judgments, ratings of declared willingness to share the images and their perceived disinformation potential, and general questions concerning AI-generated visual content. Exclusion criteria were limited to non-completion. Incomplete submissions were not retained by the survey platform and were therefore not included in the final dataset. The study was based on an anonymous, voluntary, non-invasive online survey. No medical intervention was involved, and no identifying data were collected. To enhance transparency and reproducibility, the stimulus audit trail, response matrix, agreement diagnostics, and full statistical output are provided in Appendices 1–4.

The study used a within-subject perception design. Each participant evaluated 18 images divided into three conditions: six real photographs, six DALL-E-generated images, and six Ghibli-style images. For each image, respondents answered three questions on five-point ordinal scales: perceived authenticity, declared willingness to share, and perceived risk.

The final dataset comprised 217 respondents, yielding 3,906 image-level evaluations (217 participants × 18 images). Because each participant assessed stimuli from all three conditions, the correct inferential structure is repeated measures. The statistical procedure, therefore, aggregates judgments to the participant level within each condition and tests condition effects using one-way repeated-measures ANOVAs. This directly addresses the methodological concern that repeated evaluations from the same person should not be treated as independent observations. Table 1 summarizes the main characteristics of the sample.

Table 1. Main characteristics of the sample (N = 217)

Category	*n*	%
Gender		
Female	163	75.1
Male	53	24.4

Continued on next page

Table 1. Main characteristics of the sample (N = 217) (Continued)

Category	*n*	%
Other / prefer not to say	1	0.5
Age group		
< 18 years	5	2.3
18–24 years	114	52.5
25–34 years	63	29.0
35–44 years	19	8.8
45–54 years	11	5.1
55–64 years	1	0.5
65 years and older	4	1.8
Education level		
Primary	2	0.9
Secondary	91	41.9
Higher education – Bachelor	74	34.1
Higher education – Engineer	13	6.0
Higher education – Master	36	16.6
Higher education – Doctorate and above	1	0.5

Source: Author's own work.

2.2. Stimuli and Procedure

The stimulus archive contained 18 numbered images grouped into three analytical classes: real photographs (n = 6), DALL-E outputs (n = 6), and Ghibli-style images (n = 6). In this article, "Ghibli-style" is retained as an operational label for the archived stimulus category in order to preserve consistency with the original material. Analytically, the category refers to stylized AI-generated images rather than photorealistic deepfakes. It should not be read as a claim about a single proprietary model or a formal benchmark family. What matters analytically is that these images represent a stylized synthetic mode that differs from the more recognizably artificial DALL-E outputs.

The preserved source files allow for the full reconstruction of condition membership and response structure, but not every platform-side generation parameter. Accordingly, the study does not invent missing metadata *post hoc*. Instead, it explicitly reports what is verifiable from the archive and treats the experiment as a perception study of archived stimuli rather than as a benchmark comparison of model architectures. A reproducibility register and stimulus audit trail are provided in Appendix 1.

This approach is methodologically preferable to the *post hoc* reconstruction of unverifiable prompt logs.

2.3. Measures

Three measures were analyzed.

1. Perceived authenticity (1 = definitely fake, 5 = definitely authentic).
2. Declared willingness to share (1 = definitely would not share, 5 = definitely would share).
3. Perceived risk (1 = very low risk, 5 = very high risk).

For the supplementary accuracy analysis, two coding thresholds were used. Real photographs were counted as correctly identified when rated 4–5 on authenticity. Synthetic images were counted as correctly identified when rated 1–2. Ratings of 3 were treated as indeterminate rather than forced into a true-false dichotomy. This distinction is important because indeterminacy is itself substantively meaningful in the context of realfakes.

2.4. Analytical Strategy

The analytical pipeline consisted of four steps. First, participant-level means were computed separately for each condition and each outcome. Second, one-way repeated-measures ANOVAs tested whether condition affected authenticity, declared willingness to share, and perceived risk. Where Mauchly's test indicated a violation of sphericity, Greenhouse–Geisser correction was applied to the degrees of freedom. Third, Holm-corrected paired *t*-tests were used for *post hoc* comparisons. Fourth, Friedman tests were added as non-parametric robustness checks to ensure that the substantive pattern did not depend on a strict interval-level interpretation of the five-point scales.

The inferential strategy does not use a chi-square framework that would treat multiple ratings from the same respondent as independent observations. Instead, it applies respondent-level aggregation within conditions and a one-way repeated-measures ANOVA consistent with the structure of the data.

3. Results

3.1. Descriptive Results

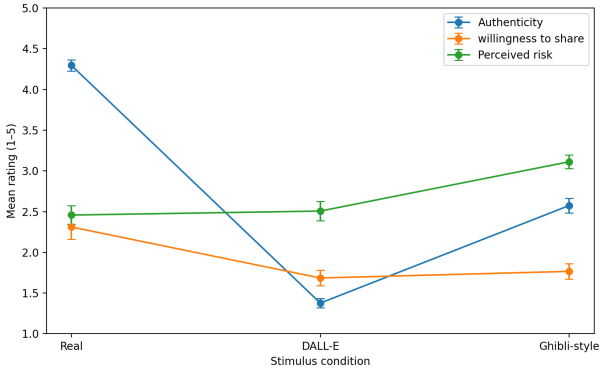
The descriptive pattern is clear. Real photographs were rated as highly authentic, DALL-E images as clearly inauthentic, and Ghibli-style images as intermediate. Regarding declared willingness to share, real photographs received the highest score,

followed by Ghibli-style images and then DALL-E outputs. On perceived risk, Ghibli-style images scored highest, suggesting that stylized synthetic visuals were not seen as harmless merely because they were not maximally photorealistic.

Table 2. Descriptive statistics by condition

Outcome	Condition	Mean	SD	95% CI
Authenticity	Real	4.29	0.52	[4.22, 4.36]
Authenticity	DALL-E	1.38	0.42	[1.32, 1.43]
Authenticity	Ghibli-style	2.57	0.67	[2.48, 2.66]
Declared willingness to share	Real	2.31	1.13	[2.16, 2.46]
Declared willingness to share	DALL-E	1.69	0.66	[1.60, 1.78]
Declared willingness to share	Ghibli-style	1.77	0.72	[1.67, 1.86]
Perceived risk	Real	2.46	0.85	[2.34, 2.57]
Perceived risk	DALL-E	2.51	0.88	[2.39, 2.62]
Perceived risk	Ghibli-style	3.11	0.62	[3.03, 3.20]

Source: Author's own work.



Source: Author's own work.

Figure 1. Visualization of the outcome profiles across conditions

3.2. Repeated-Measures ANOVA

Condition had a large and statistically significant effect on all three outcomes. Because the sphericity assumption was violated in each model, the Greenhouse–Geisser correction was applied. The strongest effect concerned perceived authenticity ($\eta^2p = 0.897$),

indicating that the three stimulus families were received in sharply different ways. Significant condition effects also emerged for declared willingness to share ($\eta^2p = 0.250$) and perceived risk ($\eta^2p = 0.273$).

Table 3. Repeated-measures ANOVA summary

Outcome	$SS_{condition}$	SS_{error}	$MS_{condition}$	MS_{error}	df ($GG_{corrected}$)	F	p	η^2p
Authenticity	933.63	106.67	466.81	0.25	1.84, 397.91	1890.55	< 0.001	0.897
Declared willingness to share	50.27	151.10	25.14	0.35	1.38, 299.00	71.93	< 0.001	0.250
Perceived risk	57.64	153.23	28.82	0.35	1.66, 359.02	81.26	< 0.001	0.273

Source: Author's own work.

Post hoc comparisons for authenticity showed a fully ordered pattern: real photographs were rated as more authentic than Ghibli-style images, which in turn were rated as more authentic than DALL-E outputs. This is substantively important because it shows that synthetic media do not form a single perceptual class.

Table 4. *Post hoc* contrasts for authenticity (Holm-corrected)

A	B	Mean difference	95% CI	t	p	Cohen's d_z
Real	DALL-E	2.92	[2.82, 3.01]	59.93	< 0.001	4.07
Real	Ghibli-style	1.72	[1.62, 1.82]	32.52	< 0.001	2.21
DALL-E	Ghibli-style	−1.20	[−1.28, −1.12]	−29.41	< 0.001	2.00

Source: Author's own work.

The mean scores for declared willingness to share followed the same order (Real > Ghibli-style > DALL-E). For perceived risk, Ghibli-style images were rated as riskier

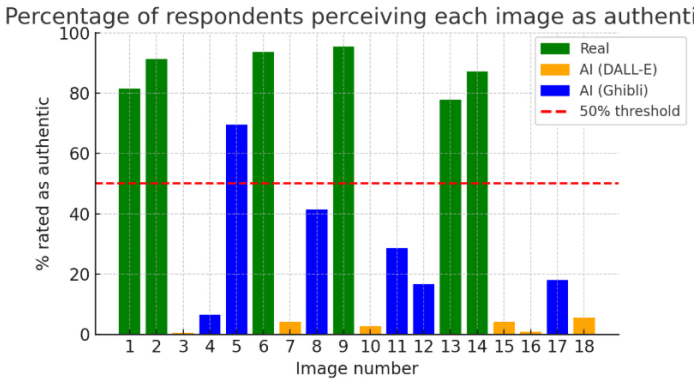
than both real photographs and DALL-E outputs. The full *post hoc* tables are provided in Appendix 1.

3.3. Ambiguity, Not Only Deception

A security-focused reading of the findings emerges when authenticity ratings are examined as distributions rather than only as means. Across all real photographs, 87.8% of ratings fell in the authentic range (4–5), whereas only 3.1% of DALL-E ratings did so. Ghibli-style images generated the largest zone of ambiguity, with 15.1% neutral judgments (rating = 3). At the level of individual stimuli, however, authenticity rates varied across images, and the most persuasive stimulus was judged authentic by 69.6% of participants (Figure 2). In other words, their effect was not simply to fool or fail; they more often produced hesitation.

That finding is theoretically consequential. In an information-security setting, hesitation can be operationally meaningful. A fabricated image need not persuade every recipient that an event certainly occurred. It may be enough to render the evidentiary field unstable, to slow verification, to trigger premature sharing by some users, or to create discursive space for strategic denial by others. This is precisely where the concept of "realfake" becomes analytically useful: it captures the security significance of media that erode certainty without necessarily achieving universal deception.

Supplementary detectability differences between stimulus classes were substantial and are reported descriptively in Figure 2 and Appendix 3. The inferential core of the study rests on the repeated-measures models reported above.



Source: Author's own work.

Figure 2. Percentage of respondents perceiving each image as authentic (authenticity = 4–5)

3.4. Robustness Checks

The substantive pattern was unchanged in non-parametric Friedman tests based on participant-level scores. All three outcomes remained significant at $p < 0.001$, confirming that the central findings do not depend on treating the five-point scales as interval measures. Supplementary participant-level accuracy analyses likewise showed that DALL-E outputs were most readily rejected as fake, real photographs were usually recognized as authentic, and Ghibli-style images produced the weakest decisive classification. Again, the salient point is not merely that one AI class was "better" or "worse," but that one class more effectively occupied a middle zone between immediate rejection and confident acceptance.

4. Discussion

First, the empirical differentiation between DALL-E-type and Ghibli-style images demonstrates that "AI-generated content" is too undifferentiated for security analysis. Recent literature shows that synthetic media matter not only because they can mislead but because they can erode trust and intensify uncertainty in public communication.[17] Other analyses suggest that the broader effect lies in audience reactions to ambiguity, not only in the technical properties of the artifact itself.[18] The present results extend this literature by indicating that a stylized yet plausible intermediate class can produce a more consequential reception profile than images that are obviously artificial.

Second, the results support a more precise conceptual distinction between "deepfake" and "realfake." The former is predominantly a production-centered label tied to synthetic substitution or fabrication, whereas the latter, as used here, denotes a broader operational category defined by documentary mimicry and security effects. This aligns the argument with work that treats manipulated and synthetic media as problems of information warfare, governance, and evidentiary destabilization rather than as purely technical curiosities.

Third, the intermediate position of Ghibli-style images suggests that the most consequential realfake is not necessarily the one that appears maximally realistic. It may instead be the one that produces the widest zone of hesitation. This finding is consistent with the idea that uncertainty itself is a security effect: it slows verification,

17 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 1.

18 Łabuz and Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About?," 882.

fragments judgment, and weakens the shared evidentiary basis on which institutions and publics rely when reacting to contested events.

These findings indicate that the security relevance of synthetic media cannot be reduced to successful deception alone. A substantial part of the risk lies in the production of interpretive uncertainty, which increases verification costs and weakens the stability of evidentiary judgments. The present findings indicate the practical relevance of AI-assisted disinformation-detection approaches that combine technical screening with contextual assessment rather than relying solely on binary authenticity judgments.[19] At the legal level, the distinction between fraudulent documentary simulation and openly synthetic creative production requires clearer normative articulation, particularly in light of post-AI-Act debates on the legality of creating and disseminating deepfakes.[20]

This interpretation also helps refine policy priorities. If the main problem were only complete deception, then better detectors might be sufficient. However, if the core challenge includes ambiguity yield, then the response must be broader: provenance infrastructures, workflow logging, platform labeling, institutional verification routines, and media literacy all become central. Appendix 4 develops that policy logic in detail.

5. Limitations

Several limitations should be stated explicitly:

First, the sample was non-probabilistic. The study should, therefore, be interpreted analytically rather than as nationally representative.

Second, the preserved archive does not contain every original generation log or platform-side parameter. The study addresses this by clearly distinguishing between reconstructible metadata and unavailable archival information rather than filling the gaps speculatively.

Third, the study concerns still images rather than video or audio, consequently the conclusions should not be generalized automatically across all synthetic media formats.

Fourth, the design captures immediate perceptual judgments, not longer-term behavioral effects in real platform ecologies.

---

19 Julia Puczyńska and Youcef Djenouri, "AI in Disinformation Detection," *Applied Cybersecurity & Internet Governance* 3, no. 2 (2024): 212, https://doi.org/10.60097/acig/200200.

20 Bernacka, "Problematyka prawna technologii deepfake," 677–678.

Fifth, because all participants viewed the stimuli in the same fixed order, order and carry-over effects cannot be excluded.

Sixth, the inferential analysis was based on respondent-level condition means rather than item-level mixed-effects models, which means that the reported effects capture robust differences within this stimulus set but do not fully model stimulus-specific variance.

Seventh, declared willingness to share was measured as self-reported intention rather than observed online behavior.

Eighth, the sample included a very small number of respondents under 18 years of age (n = 5), recruited through open online survey-participation groups. No identifying data were collected, and the questionnaire concerned only low-risk perceptual judgments.

These limitations, however, do not weaken the core contribution. The study still provides a methodologically cleaner and theoretically sharper account of how different classes of synthetic images are received and why that matters for security analysis.

Conclusion

The article makes two principal contributions. Empirically, the study demonstrates that real photographs, DALL-E outputs, and Ghibli-style images generate significantly different patterns of perceived authenticity, declared willingness to share, and perceived risk. Conceptually, it shows why the notion of "realfake" deserves a stable place in the vocabulary of security studies. A realfake is not merely fake content generated by advanced tools; it is synthetic or hybrid media that imitate the evidentiary force of reality and thereby intervene in the interpretation of reality.

The findings support the main hypothesis: the three stimulus classes generated significantly different reception profiles, and the most consequential security effect was not limited to false acceptance alone but also included ambiguity, hesitation, and evidentiary destabilization.

For the discipline of security studies, that distinction matters. The problem is no longer confined to identifying manipulated artifacts after the fact. It concerns the protection of evidentiary environments themselves. Future work should, therefore, connect perception studies such as this one with provenance infrastructures, platform governance, institutional verification routines, and crisis communication research. The strategic challenge of synthetic media is not simply falsification; it is the organized production of uncertainty within systems that still rely on visual evidence.

Appendix 1 – Expanded Methodological and Statistical Documentation

1. Purpose and Scope

This appendix provides expanded methodological and statistical documentation for the analyses reported in the main article. Its purpose is to make the analytical pipeline fully auditable at the level of design, aggregation, model selection, robustness checks, and classification logic. The appendix is aligned with the inferential structure of the main text and uses the respondent, not the individual image judgment, as the primary inferential unit.

Because each respondent evaluated all three stimulus classes, the design is a within-subject repeated-measures design. Accordingly, omnibus comparisons were estimated with a one-factor repeated-measures ANOVA, with Greenhouse–Geisser correction applied where sphericity was violated. Holm-corrected paired comparisons and non-parametric Friedman tests were used as confirmatory follow-ups. Prior research indicates that human discrimination of manipulated and AI-generated media is imperfect and strongly stimulus-dependent.[21]

2. Notation and Analytical Logic

Let:

- $i = 1, \ldots, N$ – index respondents, where $N = 217$;
- $c \in \{R, D, G\}$ – index conditions: Real, DALL-E, Ghibli-style;
- $j = 1, \ldots 6$ – index images within a condition;
- $m \in \{A, S, P\}$ – index outcomes: authenticity, declared willingness to share, and perceived risk.

Let $Y_{icj}^{(m)}$ denote the raw Likert response given by respondent i to image j in condition c on outcome m, with a response scale of 1–5.

For inferential analyses, responses were first aggregated to the respondent-by-condition level:

$$\overline{Y}_{ic}^{(m)} = \frac{1}{6} \sum_{j=1}^{6} Y_{icj}^{(m)}.$$

21 Mirsky and Lee, "The Creation and Detection of Deepfakes," 1–4; Nightingale and Farid, "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy," 2.

Thus, each respondent contributed one condition-level score per outcome for each of the three stimulus classes.

The corresponding repeated-measures model may be written as:

$$\overline{Y}_{ic}^{(m)} = \mu^{(m)} + \alpha_c^{(m)} + s_i^{(m)} + \varepsilon_{ic}^{(m)},$$

where:

- $\mu^{(m)}$ – is the grand mean for outcome m;
- $\alpha_c^{(m)}$ – is the fixed effect of condition;
- $s_i^{(m)}$ – is the respondent-specific subject effect;
- $\varepsilon_{ic}^{(m)}$ – is the residual error term.

The omnibus null hypothesis for each outcome was:

$$H_0^{(m)} \; : \; \mu_R^{(m)} = \mu_D^{(m)} = \mu_G^{(m)}.$$

3. Archival Limits and Reproducibility Boundaries

This appendix reports only those procedural elements that can be documented from the preserved study materials with confidence. The available archive contains the final stimulus set, the questionnaire structure, the rating scales, and the respondent-level data required to reproduce the inferential analyses. It does not support a complete reconstruction of every model-side generation decision, intermediate iteration, or prompt-level version history for the synthetic images.

The revised appendix, therefore, does not claim a level of procedural reconstruction that the archive itself cannot support. Instead, the study is described more precisely as a perception study of an archived stimulus set. This choice increases methodological honesty and prevents the supplement from overstating the reproducibility of generative steps that were not fully preserved.

A stronger future replication should preserve, alongside the final stimuli, a complete generation ledger containing prompt history, model/version identifiers, parameter settings, iteration sequence, seed information where available, and explicit curation rules for final image selection.

4. Sample Structure and Observation Count

The analytical structure of the study is summarized in Table 5. The inferential unit in repeated-measures analyses was the respondent-level condition mean, not the single image judgment.

Table 5. The analytical structure of the study (N = 217)

Component	Value
Respondents included in final analysis	217
Conditions per respondent	3
Images per condition	6
Total images evaluated per respondent	18
Total image-level judgments	3,906
Outcomes rated per image	3
Scale range	1–5

Source: Author's own work.

5. Condition-Level Descriptive Statistics

Table 6 reports respondent-level descriptive statistics by condition for each outcome. All values are based on aggregated respondent-by-condition means and correspond to the descriptive estimates reported in the main article. For each condition mean, the 95% confidence interval was computed as:

$$CI_{95}\left(\overline{Y}_c\right) = \overline{Y}_c \pm t_{0.975,\ N-1}\left(\frac{s_c}{\sqrt{N}}\right),$$

where s_c is the sample standard deviation for the respondent-level condition mean.

Table 6. Descriptive statistics by condition (respondent-level means, N = 217)

Outcome	Condition	Mean	SD	95% CI
Authenticity	Real	4.29	0.52	[4.22, 4.36]
	DALL-E	1.38	0.42	[1.32, 1.43]
	Ghibli-style	2.57	0.67	[2.48, 2.66]
Declared willingness to share	Real	2.31	1.13	[2.16, 2.46]
	DALL-E	1.69	0.66	[1.60, 1.78]
	Ghibli-style	1.77	0.72	[1.67, 1.86]
Perceived risk	Real	2.46	0.85	[2.34, 2.57]
	DALL-E	2.51	0.88	[2.39, 2.62]
	Ghibli-style	3.11	0.62	[3.03, 3.20]

Source: Author's own work.

These descriptive statistics already indicate a clear separation between the three stimulus classes on perceived authenticity. They also show that declared willingness to share was highest for real photographs and lower for both synthetic categories, whereas perceived risk was highest for the Ghibli-style condition.

6. Repeated-Measures ANOVA

Sphericity was evaluated with Mauchly's test. Because the assumption was violated for all three outcomes, Greenhouse–Geisser-corrected degrees of freedom were used for inferential interpretation.

For each outcome, the F statistic was computed in the standard form:

$$F = \frac{MS_{condition}}{MS_{error}}.$$

Partial eta squared was computed as:

$$\eta_p^2 = \frac{SS_{condition}}{SS_{condition} + SS_{error}}.$$

Greenhouse–Geisser-corrected degrees of freedom were defined as:

$$df_{1,GG} = \varepsilon_{GG}(k-1),$$

$$df_{2,GG} = \varepsilon_{GG}(k-1)(N-1),$$

where $k = 3$ conditions and ε_{GG} is the Greenhouse–Geisser epsilon.

Table 7. Mauchly's test of sphericity and Greenhouse–Geisser correction

Outcome	Mauchly's W	Mauchly p	$df_{1,GG}$	$df_{2,GG}$	ε_{GG}
Authenticity	0.792	< 0.001	1.84	397.91	0.921
Declared willingness to share	0.757	< 0.001	1.38	299.00	0.692
Perceived risk	0.647	< 0.001	1.66	359.02	0.831

Source: Author's own work.

Table 8. Repeated-measures ANOVA summary (Greenhouse–Geisser corrected)

Outcome	$SS_{condition}$	SS_{error}	$MS_{condition}$	MS_{error}	df ($GG_{corrected}$)	F	p	$\eta^2 p$
Authenticity	933.63	106.67	466.81	0.25	1.84, 397.91	1890.55	< 0.001	0.897
Declared willingness to share	50.27	151.10	25.14	0.35	1.38, 299.00	71.93	< 0.001	0.250
Perceived risk	57.64	153.23	28.82	0.35	1.66, 359.02	81.26	< 0.001	0.273

Source: Author's own work.

Greenhouse–Geisser-corrected degrees of freedom are reported because Mauchly's test indicated a violation of sphericity for each outcome.

The omnibus condition effect was statistically significant for all three outcomes. The largest effect was observed for authenticity judgments ($\eta_p^2 = 0.90$), indicating a very strong separation between conditions on perceived authenticity. Condition effects were also substantial for declared willingness to share and perceived risk, though clearly smaller than for authenticity.

7. Holm-Corrected Pairwise *Post Hoc* Comparisons

Pairwise comparisons were estimated with paired-samples *t*-tests on respondent-level condition means. Holm correction was applied separately within each outcome family.

For each contrast:

$$\overline{D}^{(m)}_{C_1C_2} = \frac{1}{N}\sum_{i=1}^{N}(\overline{Y}^{(m)}_{ic_1} - \overline{Y}^{(m)}_{ic_2}),$$

$$t = \frac{\overline{D}^{(m)}_{C_1C_2}}{\frac{s_D}{\sqrt{N}}},$$

$$d_z = \frac{\overline{D}^{(m)}_{C_1C_2}}{s_D} = \frac{t}{\sqrt{N}},$$

where s_D is the standard deviation of paired differences.

The 95% confidence interval for the paired mean difference was:

$$CI_{95}\left(\overline{D}^{(m)}_{C_1C_2}\right) = \overline{D}^{(m)}_{C_1C_2} \pm t_{0.975,\ N-1}\left(\frac{s_D}{\sqrt{N}}\right).$$

Table 9. Holm-adjusted pairwise contrasts by outcome

Outcome	Contrast	Mean difference	95% CI	*t*	Holm-adjusted *p*	d_z
Authenticity	Real vs. DALL-E	2.92	[2.82, 3.01]	59.93	< 0.001	4.07
	Real vs. Ghibli-style	1.72	[1.62, 1.82]	32.52	< 0.001	2.21
	DALL-E vs. Ghibli-style	−1.20	[−1.28, −1.12]	−29.41	< 0.001	−2.00
Declared willingness to share	Real vs. DALL-E	0.62	[0.49, 0.75]	9.44	< 0.001	0.64
	Real vs. Ghibli-style	0.54	[0.41, 0.68]	8.01	< 0.001	0.54
	DALL-E vs. Ghibli-style	−0.08	[−0.16, 0.00]	−2.03	0.045	−0.14
Perceived risk	Real vs. DALL-E	−0.05	[−0.18, 0.09]	−0.68	0.496	−0.05
	Real vs. Ghibli-style	−0.65	[−0.79, −0.50]	−8.83	< 0.001	−0.60
	DALL-E vs. Ghibli-style	−0.60	[−0.73, −0.46]	−8.69	< 0.001	−0.59

Source: Author's own work.

Positive differences indicate higher scores for the first-named condition. Holm correction was applied within each outcome. d_z denotes the paired-samples standardized mean difference.

The *post hoc* structure clarifies the substantive shape of the condition effect. Authenticity ratings separated all three conditions strongly. For declared willingness to

share, real photographs were shared more readily than either synthetic category, while the difference between DALL-E and Ghibli-style images was small. For perceived risk, real and DALL-E images did not differ reliably, whereas Ghibli-style images were rated as significantly riskier than both.

8. Non-Parametric Robustness Checks

To verify that the substantive pattern was not an artifact of parametric assumptions, Friedman tests were estimated for each outcome across the three conditions.

The Friedman statistic was computed as:

$$\chi_F^2 = \frac{12}{Nk(k+1)} \sum_{c=1}^{k} R_c^2 - 3N(k+1),$$

where R_c denotes the sum of within-respondent ranks for condition c, $N = 217$, and $k = 3$.

Kendall's W was computed as:

$$W = \frac{\chi_F^2}{N(k-1)}.$$

Table 10. Friedman test results

Outcome	$\chi^2(2)$	p	Kendall's W
Authenticity	336.90	< 0.001	0.78
Declared willingness to share	119.34	< 0.001	0.28
Perceived risk	58.13	< 0.001	0.13

Source: Author's own work.

Friedman tests confirm the same directional conclusion as the repeated-measures ANOVA models. These non-parametric results support the same ranking pattern observed in the parametric analyses and, therefore, strengthen the robustness of the main findings.

9. Participant-Level Decisive Classification Accuracy

To supplement mean-rating analyses, a classification-oriented index was computed at the respondent level. This index captures whether participants correctly classified images in a decisive manner rather than remaining neutral.

For each image, responses were coded as follows:

- Real images were coded as correctly classified if authenticity was rated 4–5;
- Synthetic images were coded as correctly classified if authenticity was rated 1–2;
- A response of 3 was coded as neutral/indeterminate, not correct.

Let $C_{icj} = 1$ when the response to image j in condition c is both decisive and correct, and $C_{icj} = 0$ otherwise. The respondent-level decisive accuracy score for each condition was:

$$A_{ic} = \frac{1}{6} \sum_{j=1}^{6} C_{icj},$$

where C_{icj} takes the value 1 for a correct decisive classification and 0 otherwise.

Table 11. Participant-level decisive accuracy

Condition	Mean accuracy	SD	95% CI
Real	0.878	0.249	[0.844, 0.911]
DALL-E	0.925	0.159	[0.904, 0.946]
Ghibli-style	0.550	0.350	[0.503, 0.597]
Overall	0.784	0.164	[0.762, 0.806]

Source: Author's own work.

Table 12. Decisive classification profile by condition (%)

Condition	Correctly classified decisively	Neutral / indeterminate	Misclassified decisively
Real	87.8	9.2	3.1
DALL-E	92.5	5.1	2.5
Ghibli-style	55.0	29.9	15.1
Overall	78.4	14.7	6.9

Source: Author's own work.

Percentages may not sum to exactly 100 because of rounding. This classification-oriented analysis reinforces the rating-based results. Real and DALL-E images were usually classified decisively and correctly, whereas the Ghibli-style condition generated markedly more hesitation and substantially more decisively incorrect judgments. From a classification standpoint, the Ghibli-style condition therefore occupied the most ambiguous position in the stimulus set.

10. Image-Level Heterogeneity within the Ghibli-Style Condition

Because the Ghibli-style condition occupied an intermediate position in the omnibus and *post hoc* analyses, an additional descriptive breakdown was computed at the level of individual images. These results are descriptive only and should not be interpreted as replacing the respondent-level inferential framework.

For each Ghibli-style image, authenticity responses were grouped into three bands:

- Judged fake: responses 1–2;
- Neutral: response 3;
- Judged authentic: responses 4–5.

The corresponding image-level proportions were defined as:

$$p_{j,fake} = \frac{n_{j,1-2}}{N} * 100,$$

$$p_{j,neutral} = \frac{n_{j,3}}{N} * 100,$$

$$p_{j,authentic} = \frac{n_{j,4-5}}{N} * 100.$$

Table 13. Image-level authenticity profile within the Ghibli-style condition (%)

Image	Judged fake (1–2)	Neutral (3)	Judged authentic (4–5)
G1	19.4	11.0	69.6
G2	41.9	15.7	42.4
G3	55.1	15.7	29.2
G4	71.3	11.1	17.6
G5	38.9	20.8	40.3
G6	61.1	13.0	25.9

Source: Author's own work.

These values are descriptive and are reported to show within-condition heterogeneity, not to redefine the inferential unit of the study.

The table shows that the Ghibli-style category was not internally uniform. Some items were frequently treated as authentic, while others were much more often recognized as synthetic. This heterogeneity helps explain why the condition-level mean for Ghibli-style stimuli falls between the real and DALL-E categories rather than clustering tightly with either one. That pattern is compatible with prior findings showing that human detection of synthetic media varies substantially across individual stimuli rather than operating as a stable all-or-nothing capacity.[22]

11. Methodological Conclusion

Taken together, the analyses reported in this appendix support four methodological conclusions.

First, the inferential logic of the study must be based on repeated measures, because each respondent evaluated all three stimulus classes. Treating single judgments as independent observations would not reflect the actual design.

Second, the substantive pattern is robust across parametric and non-parametric specifications. The condition effect appears consistently for authenticity, declared willingness to share, and perceived risk.

Third, the Ghibli-style condition is the analytically decisive category. It was less authentic than real photography, more authentic than DALL-E imagery, and much more likely than either comparison category to generate hesitation or misclassification.

Fourth, the appendix should remain methodologically modest about what can and cannot be reconstructed from the preserved archive. The inferential analyses are reproducible from the retained dataset, but the full generative history of the synthetic images is not. That distinction strengthens, rather than weakens, the credibility of the supplement.

Appendix 2 – Truth or Deepfake – Perception Study Responses

This appendix serves a documentary and verification function. It identifies the anonymized respondent-level response matrix underlying the descriptive summaries and inferential analyses reported in the main article and Appendix 1.

22 Sergi D. Bray et al., "Testing Human Ability to Detect 'Deepfake' Images of Human Faces," *Journal of Cybersecurity* 9, no. 1 (2023): 1, https://doi.org/10.1093/cybsec/tyad011.

The complete response file contains the full set of questionnaire outputs in raw tabular form, including respondent characteristics, image-level ratings, and general post-survey items. Specifically, it includes demographic variables, evaluations of the study stimuli across the core analytical dimensions used in the article, and the additional survey items reported in the extended methodological documentation. No identifying personal data were collected, and the dataset is fully anonymized.

Because the full response matrix is extensive and technical in nature, it is not reproduced here in full as a typeset appendix. Instead, it has been submitted separately to the editorial office as supporting material for transparency and verification, and, if required, for further editorial or reviewer inspection. Graphical, tabular, and inferential syntheses of these data are presented in the main article and Appendix 1.

Accordingly, the function of Appendix 2 is not to introduce additional interpretations but to document the existence, scope, and availability of the underlying response material on which the reported empirical results are based.

Appendix 3 – Extended Case Studies and Literature Analysis

1. Purpose of the Appendix

This appendix provides an extended interpretive discussion of two Ghibli-style stimuli selected as analytically revealing cases within the most ambiguity-prone stimulus class identified in the study. It does not introduce new inferential analyses. Rather, it develops the image-level interpretation of patterns already reported in the main article and in Appendix 1, with particular attention to how selected stimuli illuminate the broader argument about realfake as a security-relevant phenomenon. For consistency with the main-article numbering, this appendix refers to the selected stimuli as Image 5 and Image 8.

The purpose of this appendix is, therefore, limited and precise. It is not intended to overgeneralize from single images but to show why certain synthetic stimuli were more destabilizing than others and how those differences can be understood in light of existing literature on deepfakes, AI-generated faces, and synthetic political content.

2. Case Study I: Image 5

Among the Ghibli-style stimuli, Image 5 produced the strongest false-authenticity effect. A clear majority of respondents treated it as authentic, while only a minority decisively rejected it as synthetic. This makes Image 5 the clearest example in the

present study of a synthetic image that achieved high documentary plausibility despite not belonging to the real-photograph condition.

This pattern is analytically important because the image appears to benefit from a particularly efficient combination of visual familiarity and perceptual simplicity. First, the composition is strongly face-centered and does not overload the viewer with contextual details that would need to be verified. Second, the stimulus draws on a recognizable public-face schema, which may have encouraged rapid impression-based recognition rather than slower analytical checking. Third, the image does not contain immediately salient distortions that would force a strong rejection response. In this sense, its persuasive force appears to result less from sheer technical perfection than from the effective management of ambiguity.[23]

This interpretation is consistent with prior research suggesting that synthetic human faces can be experienced as highly plausible and, under some conditions, even as more trustworthy than real ones.[24] It is also consistent with broader findings that human performance in distinguishing AI-generated from authentic imagery remains strongly stimulus-dependent rather than uniformly high or low across all examples.[25]

3. Case Study II: Image 8

Image 8 generated a different but equally important response pattern. Unlike Image 5, the key feature of this stimulus was not one-sided acceptance but a near-split between authenticity judgments and rejection judgments, with an additional undecided group. This makes Image 8 analytically valuable not because it simply "passed" as real, but because it activated competing interpretations of the same visual material.

That distinction is crucial. For one part of the audience, the public figure, the political setting, and the documentary style were sufficiently coherent to support authenticity judgments. For the other, those same cues appear to have triggered suspicion and rejection. The result is not merely partial deception but interpretive instability. In other words, the image did not only test whether respondents would

23 Andreea Pocol et al., "Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media," in *Advances in Computer Graphics* (Springer Nature Switzerland, 2023), 427–428, https://doi.org/10.1007/978-3-031-50072-5_34.

24 Nightingale and Farid, "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy," 2.

25 Zeyu Lu et al., "Seeing Is Not Always Believing: Benchmarking Human and Model Perception of AI-Generated Images," 2023, 1–2, https://doi.org/10.48550/ARXIV.2304.13023.

Source: Synthetic adaptation prepared for the experiment; base photograph: Małgorzata Foremniak 2 (20337023435) (cropped), author: Fryta 73, via Flickr/Wikimedia Commons, licensed under CC BY-SA 2.0, https://commons.wikimedia.org/wiki/File:Ma%C5%82gorzata_Foremniak_2_%2820337023435%29_%28cropped%29.jpg.

Figure 3. Ghibli-style synthetic portrait based on a public photograph of actress Małgorzata Foremniak

believe it; it also tested whether they could still agree on what counts as credible visual evidence once a familiar political scene was rendered in a synthetic form.

This mechanism matters especially in politically salient contexts. Existing research on synthetic political media has shown that the principal danger is not limited to outright deception alone. It also includes increased uncertainty, reduced evidentiary confidence, and greater distrust toward mediated political communication.[26] Seen in this light, Image 8 is important precisely because it demonstrates how a synthetic, news-like political scene can divide perception even when it does not produce overwhelming false acceptance.

4. Relation to the Broader Literature

Taken together, the two case studies support a more mature interpretation of the main findings.

26 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 2–4.

Source: Synthetic adaptation prepared for the experiment; base photograph: Szymon Hołownia 2021, author: Tomasz Kaczor, via Wikimedia Commons, licensed under CC BY-SA 4.0, https://commons.wikimedia.org/wiki/File:Szymon_Ho%C5%82ownia_2021.jpg.

Figure 4. Ghibli-style synthetic image based on a public photograph of politician Szymon Hołownia

First, synthetic-image risk should not be reduced to the question of whether users can or cannot detect falsification in general. The present data suggest that susceptibility varies substantially across individual stimuli, even within the same stylistic category. Some synthetic images are rejected quickly, some are accepted at high rates, and some divide the audience between belief and suspicion. This means that the security significance of synthetic imagery depends not only on the general class of content but also on the micro-structure of individual examples.

Second, the present results support a shift from a purely artifact-centered account toward an audience-centered account. The practical problem is not only the formal quality of the synthetic image itself, but also the interaction between visual plausibility, contextual cues, prior expectations, and the viewer's threshold for doubt.[27] This perspective is especially important for the concept of realfake because it captures both the false acceptance of synthetic content and corrosive doubt toward content that appears visually coherent.

Third, the comparison between Image 5 and Image 8 suggests that different forms of risk may arise from different types of synthetic imagery. Face-centered portrait

27 Łabuz and Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About?," 875–877.

stimuli may benefit from perceptual simplicity and familiarity, thereby increasing the probability of false acceptance. Politically salient scene-based stimuli may instead maximize interpretive instability, in which neither belief nor disbelief becomes dominant. Both mechanisms are security-relevant, but they do not operate identically and should not be collapsed into a single model of deception.

5. Concluding Observations

The case-study evidence strengthens rather than complicates the central argument of the article. The most consequential problem is not only that some synthetic images are believed. It is also that synthetic visual content can weaken the reliability of visual judgment by producing ambiguity where viewers ordinarily expect documentary clarity.

For that reason, realfake should be understood not merely as a problem of false representation but also as a problem of evidentiary destabilization. Its security relevance lies both in the possibility of deception and in the erosion of shared confidence in visual proof. The present appendix should, therefore, be read as an interpretive extension of the main statistical results, showing at the image level how the broader quantitative pattern may operate through perception, context, and trust.

Appendix 4 – Countermeasures and Policy Recommendations

1. Purpose and Scope

This appendix extends the argument developed in the main article by translating its conceptual and empirical conclusions into a layered framework of countermeasures and policy recommendations. The central premise is that realfakes should not be approached solely as a problem of fabricated content. They should also be understood as a problem of evidentiary destabilization: synthetic or hybrid media may not only mislead recipients directly, but may also weaken confidence in authentic materials, increase verification costs, and burden decision-making in environments where time, trust, and attribution matter simultaneously.[28]

28 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 3; Łabuz and Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About?," 878.

For that reason, effective responses should not be judged by whether they eliminate all synthetic or manipulated content, which is an unrealistic objective, but by whether they reduce ambiguity, preserve traceability, shorten verification time, and support accountable institutional judgment. A robust response to realfakes must, therefore, be layered. Technical, organizational, communicative, educational, and legal measures should reinforce one another rather than operate in isolation.[29]

2. Technical Countermeasures

At the technical level, provenance-based mechanisms are more defensible than purely reactive detection when institutions require auditable records of origin and transformation. Content credentials, cryptographic signing, secure capture environments, preservation of original files, metadata retention, and file-integrity checks can help establish the source of a piece of content and whether it has been altered after creation or acquisition.[30]

At the same time, provenance should not be conflated with truth. Even where provenance information is available, it does not by itself resolve contextual deception, selective framing, misleading montages, or the strategic recirculation of genuine materials in false interpretive contexts. Provenance strengthens origin authentication and integrity tracking, but it does not eliminate the need for contextual verification and human judgment.[31]

Accordingly, detection systems should be treated as triage instruments rather than "truth machines." Their practical value lies in supporting prioritization, anomaly flagging, and review workflows. They should be used in conjunction with reverse-image searches, metadata inspection, source comparison, contextual corroboration, and expert assessment rather than as standalone arbiters of authenticity.[32]

A further implication concerns documentation standards. Where synthetic or hybrid materials are used in research, journalism, strategic communication, or institutional analysis, it is advisable to preserve a generation and modification ledger

29 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (U.S. Department of Commerce, January 2023), 12–13, https://doi.org/10.6028/nist.ai.100-1.

30 Coalition for Content Provenance and Authenticity, *C2PA Technical Specification*, v. 2.1, 2024, 1–4, https://spec.c2pa.org/specifications/specifications/2.1/index.html.

31 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 12–13.

32 Puczyńska and Djenouri, "AI in Disinformation Detection," 225–226.

recording the source material, model or tool used, key transformation steps, timestamps, file lineage, and final classification. Such documentation does not prevent manipulation, but it materially improves reviewability, accountability, and subsequent audits.[33]

3. Organizational Countermeasures

At the organizational level, the most important response is the introduction of explicit verification protocols for high-impact content. Any material capable of influencing public safety, political order, reputational standing, or operational decisions should pass through a documented escalation path that distinguishes between suspected synthetic content, unverified authentic content, and content verified as authentic. This distinction is important because realfake-related harm may arise not only from outright falsehood but also from the interval during which authenticity remains uncertain.[34]

Such protocols should identify the responsible reviewer, the minimum required checks, the threshold for escalation, the mode of communication uncertainty, and the rules for evidence retention. In institutional practice, this means that verification is not left to intuition or *ad hoc* improvisation. Instead, it becomes part of a repeatable security routine embedded in the organization's decision-making architecture.[35]

This logic is particularly important for public authorities, newsrooms, research teams, and platform operators. In all of these settings, the relevant question is not merely whether a piece of content is false, but whether the institution can demonstrate that it handled the material in a procedurally defensible manner. From that perspective, traceability is itself a security value.[36]

4. Communicative and Educational Countermeasures

At the communicative level, public resilience should not be reduced to the expectation that recipients will simply learn to "spot fake images." That model is insufficient for contemporary conditions, because high-quality synthetic content may be difficult

33 Coalition for Content Provenance and Authenticity, *C2PA Technical Specification*, 2–4.

34 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 2.

35 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 27.

36 National Institute of Standards and Technology, 22.

or impossible to identify through visual inspection alone. More durable resilience is built through process-oriented media literacy: source evaluation, corroboration, motive analysis, distribution-pattern assessment, temporal context, and the disciplined communication of uncertainty.[37]

For that reason, the strategic objective should not be overconfidence in individual detection ability, but better judgment under ambiguity. Public institutions, educators, and media organizations should, therefore, favor formulations such as "unverified," "pending authentication," or "source not yet confirmed" where evidentiary certainty has not yet been established. This does not signal institutional weakness, on the contrary, it signals procedural integrity and reduces the risk of premature or reputationally damaging claims.[38]

Educational responses should similarly avoid treating synthetic media solely as a technical novelty. They should be integrated into broader instruction on evidentiary reasoning, cognitive bias, source criticism, and the distinction between authenticity, credibility, and interpretive accuracy. In the context of realfakes, these distinctions are not semantic details but are operationally relevant.[39]

5. Legal and Regulatory Considerations

At the legal and regulatory level, policy should combine transparency duties, accountability rules, and implementation standards. Within the European framework, the AI Act introduces transparency obligations for certain categories of AI-generated or AI-manipulated content, including deepfakes. This is important because it moves disclosure from a largely voluntary norm toward a more structured legal expectation.[40]

However, disclosure alone is not a complete response. Labels may be absent, technically stripped, ignored by recipients, or insufficient for assessing whether a given item is materially misleading in context. A regulatory framework that relies only on disclosure risks addressing the form of synthetic content without fully addressing the

37 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 10.

38 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 27.

39 Łabuz and Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About?," 884.

40 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), Official Journal of the European Union, July 12, 2024, 82, art. 50, L, 2024/1689, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

harm arising from its use, circulation, repackaging, or strategic embedding in broader disinformation campaigns.[41]

A more defensible policy approach, therefore, requires at least four elements. First, institutions should be required to disclose synthetic or materially manipulated content where the legal framework demands it. Second, they should preserve relevant provenance and review information wherever high-impact use cases are involved. Third, operational environments should implement internal procedures governing verification, retention, escalation, correction, and appeals. Fourth, procurement and governance standards should favor systems and workflows that support auditability rather than opacity.[42]

This point is especially relevant in public-sector settings. If public bodies are expected to make or support decisions in environments increasingly affected by synthetic media, they should not rely on informal practices alone. They need operational rules that integrate legal compliance, evidentiary discipline, and technical accountability into a single governance model.[43]

Recent legal scholarship also suggests that the regulatory challenge posed by deepfakes and adjacent synthetic-media practices increasingly exceeds national responses. The problem is not only domestic misinformation but also cross-border circulation, attribution uncertainty, platform-mediated amplification, and the uneven pace of legal adaptation. From that perspective, synthetic-media governance should be understood as both a domestic regulatory issue and an emerging field of broad international coordination.[44]

6. Policy Recommendations by Institutional Domain

For public authorities, the priority should be secure intake procedures, chain-of-custody discipline, documented verification protocols, and calibrated crisis communication. Their task is not to promise perfect authenticity judgments in every case but to demonstrate that decisions were made through auditable and proportionate procedures.[45]

41 Bernacka, "Problematyka prawna technologii deepfake," 683.

42 Coalition for Content Provenance and Authenticity, *C2PA Technical Specification*, 113.

43 Bernacka, "Problematyka prawna technologii deepfake," 682–683.

44 Kuźnicka-Błaszkowska and Kostyuk, "Emerging Need to Regulate Deepfakes in International Law," 7.

45 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 22–24.

For media organizations, the priority should be source authentication workflows, the retention of original files and metadata where possible, provenance-aware publication standards, and careful wording in situations of partial uncertainty. Journalistic resilience depends not only on detecting manipulations but also on resisting the pressure to communicate certainty before it has been earned.[46]

For online platforms, the priority should be interface-level disclosure, provenance-friendly infrastructure, escalation procedures for high-risk manipulated content, and retention practices that support subsequent review and accountability. Platform governance should not treat all manipulated media as equivalent. The difference between satire, benign transformation, evidentiary distortion, reputational attack, and crisis-amplifying misinformation is normatively and operationally significant.[47]

For educational institutions, the priority should be process-oriented media literacy rather than purely visual detection drills. Students and adult learners alike should be trained to ask who produced a piece of content, the channel through which it traveled, what corroboration exists, which incentives may shape its circulation, and what degree of uncertainty remains.[48]

For researchers, the priority should be full methodological transparency whenever synthetic or hybrid materials are used in study design, experimentation, or dissemination. This requires clear documentation of how such materials were generated, selected, processed, and incorporated into the research design, along with the intended context of use, relevant assumptions, and known limitations. Such practice is not merely a matter of technical neatness. It is a condition of defensible reproducibility, interpretive clarity, and accountable research practice.[49]

7. Concluding Observations

In strategic terms, the most defensible response to realfakes is not the promise of perfect detection. It is the construction of institutions that remain functional under conditions of evidentiary stress. Countermeasures should, therefore, be evaluated

46 Vaccari and Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," 10.

47 Kuźnicka-Błaszkowska and Kostyuk, "Emerging Need to Regulate Deepfakes in International Law," 7.

48 Łabuz and Nehring, "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About?," 884.

49 National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 26.

by their ability to preserve traceability, support reasoned verification, communicate uncertainty responsibly, and slow the speed at which manipulated or ambiguously authenticated media can distort judgment.

Under this approach, resilience is not reduced to a software filter or a legal label. It becomes an integrated security practice combining technical provenance, institutional protocol, public reasoning, and regulatory accountability. This is the level at which countermeasures become proportionate to the actual problem identified in this study.

References

Legal Acts

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, July 12, 2024. L, 2024/1689, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

Literature

Bernacka, Julia. "Problematyka prawna technologii deepfake – analiza legalności tworzenia i rozpowszechniania deepfake'ów po uchwaleniu AI Act." *Prawo i Więź* 58, no. 5 (2025): 671–694. https://doi.org/10.36128/hg1acq35.

Bray, Sergi D., Shane D. Johnson, and Bennett Kleinberg. "Testing Human Ability to Detect 'Deepfake' Images of Human Faces." *Journal of Cybersecurity* 9, no. 1 (2023): tyad011. https://doi.org/10.1093/cybsec/tyad011.

Groh, Matthew, Ziv Epstein, Chaz Firestone, and Rosalind Picard. "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds." *Proceedings of the National Academy of Sciences* 119, no. 1 (2021): e2110013119. https://doi.org/10.1073/pnas.2110013119.

Hausken, Liv. "Photorealism Versus Photography: AI-Generated Depiction in the Age of Visual Disinformation." *Journal of Aesthetics & Culture* 16, no. 1 (2024): 2340787. https://doi.org/10.1080/20004214.2024.2340787.

Kietzmann, Jan, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. "Deepfakes: Trick or Treat?" *Business Horizons* 63, no. 2 (2020): 135–146. https://doi.org/10.1016/j.bushor.2019.11.006.

Kuźnicka-Błaszkowska, Dominika, and Nadiya Kostyuk. "Emerging Need to Regulate Deepfakes in International Law: The Russo-Ukrainian War as an Example." *Journal of Cybersecurity* 11, no. 1 (2025): tyaf008. https://doi.org/10.1093/cybsec/tyaf008.

Lewis, Andrew, Patrick Vu, Raymond M. Duch, and Areeq Chowdhury. "Deepfake Detection with and Without Content Warnings." *Royal Society Open Science* 10, no. 11 (2023): 231214. https://doi.org/10.1098/rsos.231214.

Lu, Zeyu, Di Huang, Lei Bai, et al. "Seeing Is Not Always Believing: Benchmarking Human and Model Perception of AI-Generated Images," 2023. https://doi.org/10.48550/ARXIV.2304.13023.

Łabuz, Mateusz. "Deep Fakes and the Artificial Intelligence Act—an Important Signal or a Missed Opportunity?" *Policy & Internet* 16, no. 4 (2024): 783–800. https://doi.org/10.1002/poi3.406.

Łabuz, Mateusz. "Regulating Deep Fakes in the Artificial Intelligence Act." *Applied Cybersecurity & Internet Governance* 2, no. 1 (2023): 252–291. https://doi.org/10.60097/acig/162856.

Łabuz, Mateusz, and Christopher Nehring. "Information Apocalypse or Overblown Fears—What AI Mis- And Disinformation Is All About? Shifting Away from Technology Toward Human Reactions." *Politics & Policy* 52, no. 4 (2024): 874–891. https://doi.org/10.1111/polp.12617.

Mirsky, Yisroel, and Wenke Lee. "The Creation and Detection of Deepfakes: A Survey." *ACM Computing Surveys* 54, no. 1 (2021): article 7. https://doi.org/10.1145/3425780.

Mitręga, Adrian. "Wojna poznawcza we współczesnym środowisku bezpieczeństwa." *Annales Universitatis Paedagogicae Cracoviensis: Studia de Securitate* 13, no. 2 (2023): 121–136. https://doi.org/10.24917/26578549.13.2.7.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce, January 2023. https://doi.org/10.6028/nist.ai.100-1.

Nightingale, Sophie J., and Hany Farid. "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy." *Proceedings of the National Academy of Sciences* 119, no. 8 (2022): e2120481119. https://doi.org/10.1073/pnas.2120481119.

Pawelec, Maria. "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions." *Digital Society* 1 (2022): article number 19. https://doi.org/10.1007/s44206-022-00010-6.

Pocol, Andreea, Lesley Istead, Sherman Siu, Sabrina Mokhtari, and Sara Kodeiri. "Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media." In *Advances in Computer Graphics*, 427–440. Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-50072-5_34.

Puczyńska, Julia, and Youcef Djenouri. "AI in Disinformation Detection." *Applied Cybersecurity & Internet Governance* 3, no. 2 (2024): 211–232. https://doi.org/10.60097/acig/200200.

Ternovski, John, Joshua Kalla, and Peter Aronow. "Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments." *Journal of Online Trust and Safety* 1, no. 2 (2022): 1–16. https://doi.org/10.54501/jots.v1i2.28.

Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6, no. 1 (2020): 1–13. https://doi.org/10.1177/2056305120903408.

Web publications

Coalition for Content Provenance and Authenticity. *C2PA Technical Specification*. V. 2.1, 2024. https://spec.c2pa.org/specifications/specifications/2.1/index.html.