Harvard Kennedy School Misinformation Review[1]
July 2026, Volume 7, Issue 4

DOI: https://doi.org/10.37016/mr-2020-205
Website: https://misinforeview.hks.harvard.edu/

Research Note

Experts disagree on how to fight AI disinformation, but agree that health and politics need different solutions

When 54 international experts assessed AI-generated disinformation threats, they revealed a surprising pattern: while video deepfakes received the highest average threat ratings in the political domain (M = 6.31/7), the pattern differed in the health domain, where AI-generated text received the highest average rating (M = 5.80). Experts also fail to converge on what to do: government regulation drew the most "most effective" votes (30%) and substantial "least effective" selections (15%), exposing deep uncertainty about how to respond. These findings offer an initial expert map of an AI-disinformation landscape that is still rapidly forming.

Authors: Alexander Loth (1), Martin Kappes (1), Marc-Oliver Pahl (2)
Affiliations: (1) Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, Germany, (2) Cybersecurity in Critical Networked Infrastructures (Cyber CNI), IMT Atlantique, France
How to cite: Loth, A., Kappes, M., & Pahl, M.-O. (2026). Experts disagree on how to fight AI disinformation, but agree that health and politics need different solutions. *Harvard Kennedy School (HKS) Misinformation Review*, *7*(4).
Received: March 30th, 2026. Accepted: July 22nd, 2026. Published: July 29th, 2026.

Research questions

- How do experts perceive AI-generated disinformation threats across four modalities (text, images, audio, video) and four domains (political, health, financial, social)?
- Which modality-domain combinations do experts view as most dangerous, and do threat profiles differ systematically across domains?
- How do experts evaluate the effectiveness of five mitigation strategies and which do they prioritize?
- What do experts identify as the most urgent near-term risks from generative AI?

Research note summary

- We surveyed 54 experts on AI-driven disinformation (online, July 2025–March 2026) reached via targeted recruitment with snowball extension; 454 invited, 11.9% response rate.
- Video deepfakes received the highest average threat ratings overall (averaging 6.19 on a 7-point scale), but threat patterns varied by domain: in health, AI-generated text had the highest (*M* = 5.80) and video had the lowest (*M* = 5.13); in finance, audio deepfakes had the highest threat ratings (*M* = 5.65). Views of mitigation strategies were contested but not polarized: government regulation drew both the most "most effective" (30%) and, notably, the most "least effective" (15%) votes; media literacy was split 26% "most effective" to 24% "least effective". Effectiveness

ratings were right-skewed and unimodal, suggesting disagreement is about priority and not about whether strategies work. Election interference via deepfake video was rated as the top urgent risk (78%).

- These preliminary findings suggest that respondents perceived domain-specific policy approaches as more appropriate than uniform ones. Experts consistently highlighted voice-cloning fraud as an area warranting particular regulatory attention. Because no single intervention commanded consensus, the findings suggest that layered mitigation approaches may be preferable across provenance, literacy, regulation, and platform enforcement, weighted to each domain's threat profile (see Appendix Table C4).

Implications

Domain-specific threat patterns suggest tailored interventions

Our respondents perceived distinct threat profiles in each domain: text dominated in health, audio in finance, and video in politics. Current governance instruments are not organized along those lines. The EU AI Act (European Parliament, 2024) and the Digital Services Act impose horizontal obligations - transparency labeling, risk assessment, systemic-risk audits - that apply identically whether the content is a fabricated medical claim or a cloned voice in a payment fraud. In the United States, oversight is fragmented across sectoral agencies, with the FDA covering health claims and the FTC covering deceptive commercial practices, and no federal AI statute in place (Bommasani et al., 2024). Neither the horizontal EU approach nor the sectoral U.S. patchwork tracks the domain-by-modality threat structure our respondents described.

In the political domain, where video deepfakes received the highest average threat ratings, respondents frequently identified detection and provenance standards as priority areas for future investment. As Giovanni Spitale (University of Zurich) stated, "My biggest concern is electoral interference and more in general interference in democratic processes, because once you break that you can break each and every other element of democratic societies." The Coalition for Content Provenance and Authenticity (C2PA) standard responds to exactly this concern: it attaches cryptographically signed capture and edit history to a media file, so a viewer can check whether a video originated from a real camera or from a generative model. Its effectiveness depends on broad adoption by camera manufacturers and platforms, and it can be defeated by re-recording or by stripping metadata (Corsi et al., 2024).

In the health domain, where AI-generated text received the highest threat ratings of the four modalities (M = 5.80), respondents suggested greater regulatory attention toward labeling AI-generated health content and improving medical fact-checking infrastructure (De Angelis et al., 2023). The World Health Organization's infodemic-management framework (World Health Organization, 2020) offers an operational template: it pairs active listening for circulating health claims with rapid authoritative rebuttal and amplification through trusted local messengers. It was developed before current large language models, but because it targets the circulation of false health claims rather than their means of production, it transfers to AI-generated text with little modification. Experts gave audio deepfakes the highest average threat ratings in the financial domain (*M* = 5.65), aligning with growing evidence on voice-cloning fraud. In 2024, a Hong Kong finance worker was deceived into transferring $25 million through a deepfake video call (Chen & Magramo, 2024). The U.S. Treasury's Financial Crimes Enforcement Network has since issued a sector-wide alert on deepfake-enabled fraud against financial institutions (FinCEN, 2024)—an early example of the type of domain-specific guidance highlighted by many respondents. The expert responses suggest that voice-authentication standards for high-value transactions deserve increased attention, and

they raise questions about the adequacy of existing know-your-customer (KYC) protocols in an environment where a live voice or video is no longer reliable evidence of identity (Chesney & Citron, 2019).

A contested, not polarized, mitigation landscape

Respondents rated five mitigation strategies: government regulation, digital watermarking, media literacy, platform enforcement, and technical detection. Likert effectiveness ratings for all five strategies are right-skewed and unimodal: the share of respondents rating a strategy 5 or higher on the 7-point scale ranged from 63% (technical detection) to 74% (government regulation), with no bimodal distribution. The two question formats point in different directions. On the Likert ratings, every strategy is broadly endorsed. On the forced-choice ranking, no strategy commands a majority and the same strategy can appear at both ends: government regulation drew 30% of "most effective" votes but also 15% of "least effective" votes, and media literacy 26% against 24%. Experts agree that each strategy can contribute; they disagree on which should come first. The landscape is contested rather than polarized, mirroring a recent *HKS Misinformation Review* survey on generative AI in Europe (Weikmann et al., 2026).

Psychological inoculation deserves a more prominent place in this priority debate than our respondents gave it; it was not among the five strategies we asked about, and no respondent volunteered it. Pre-exposing audiences to weakened forms of misinformation techniques builds resistance that persists across topics and time horizons (van der Linden, 2023), and game-based inoculation has shown durable real-world effects at scale (Roozenbeek & van der Linden, 2024). Detection tools face a perpetual arms race with generative AI models (Hoq et al., 2025; Murphy et al., 2023; Rana et al., 2022). Direct quotations below are attributed by name only to the 13 respondents who explicitly consented to such attribution; all other respondents are described generically. One respondent, the researcher and digital artist publishing as Merzmensch, captured the urgency: "Literacy! Literacy! Literacy! From the first classes in school." Another respondent, technology professional Alberto Lobato Diogo, underscored the layered character of any realistic response: "GenAI based on LLMs are mathematically impossible to have full-proof mitigation systems, so we need a combination of all of the above."

Public-awareness tools reach the wrong audiences

Experts rated public-awareness tools as moderately effective ($M = 4.43$), and one third identified the same central limitation: these tools reach mainly audiences that are already technically confident. Guo et al. (2025) report the same asymmetry from a different angle, finding that the populations most vulnerable to AI-generated disinformation are also the least likely to adopt technical countermeasures. Our respondents' assessment therefore suggests that effective mitigation may need to extend beyond tool development to distribution, accessibility, and integration into platforms where vulnerable users already consume information. One respondent, Katerina Sedova of the Atlantic Council, warned: "As chatbots become ubiquitous, get intertwined with human lives, and even invite real emotional connection and dependence from humans, it will be critical to ensure that these mediums are not weaponized."

Toward a layered, domain-weighted response

Because no single strategy commands consensus while each is broadly seen as workable, the responses point toward combination rather than selection. Grouping the five rated strategies by the mechanism through which they act yields four pillars: provenance (digital watermarking and technical detection, which act on the artifact), audience-side literacy (media literacy, which acts on the recipient), statutory regulation (government regulation, which acts on the producer), and platform enforcement (which acts

on distribution). The number four here is coincidental and carries no relation to the four modalities or the four domains.

The pillars are not weighted equally across domains. Our respondents' domain-by-modality ratings suggest different entry points: provenance first for political video, where authenticity of the artifact is the contested question; labeling and medical fact-checking for health text, where the claim rather than the medium carries the harm; voice authentication and KYC modernization for financial audio, where the attack targets an identity check; and accessibility-first literacy for the social domain, where harm is diffuse and no single chokepoint exists. Appendix Table C4 sets out this priority matrix. It is a starting point derived from perceptions, not a validated framework, and we present it as a hypothesis for testing.

Findings

Finding 1: Experts rate video deepfakes as the most threatening modality overall.

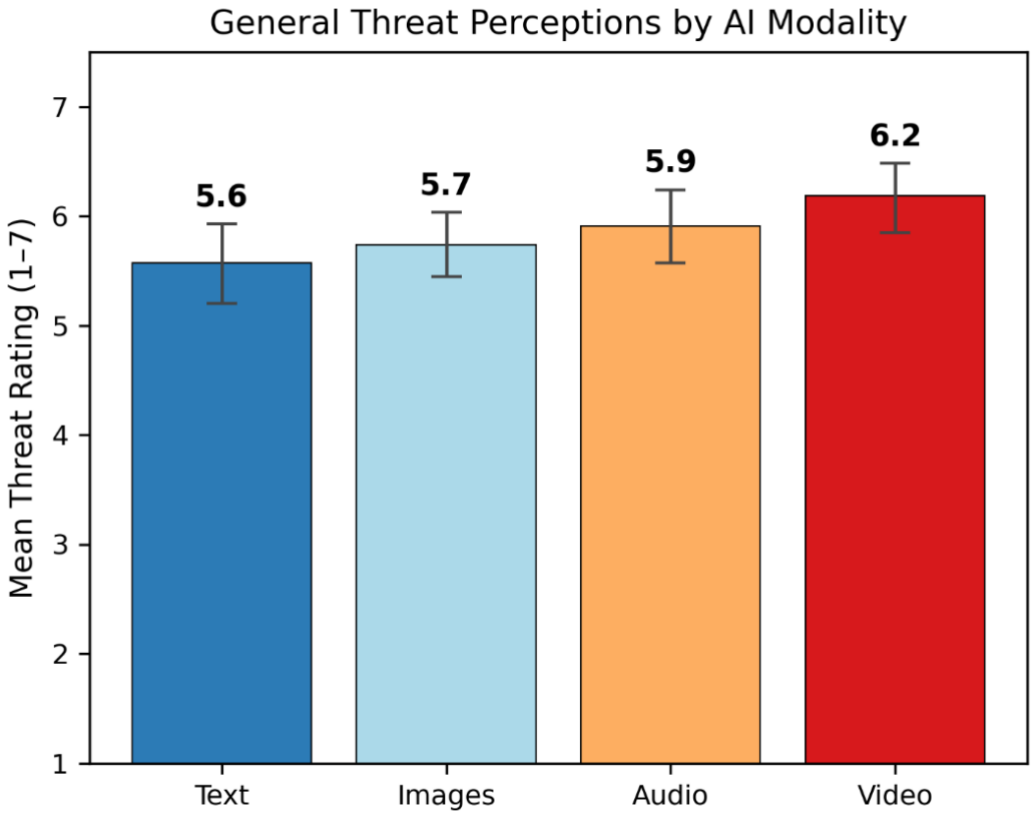

Figure 1. Modality threat perceptions across all domains (7-point scale,* N *= 54). *Error bars represent 95% bootstrap confidence intervals (10,000 iterations).*

Throughout the Findings, ratings use a 7-point scale (1 = not threatening; 7 = extremely threatening) *MSD*. Across all domains, video deepfakes received the highest average threat rating (*M* = 6.19, *SD* = 1.19), followed by audio (*M* = 5.91, *SD* = 1.25), images (*M* = 5.74, *SD* = 1.13), and text (*M* = 5.57, *SD* = 1.36). This pattern is broadly consistent with previous work suggesting that synthetic content in higher-fidelity modalities carries greater persuasive force, and that untrained observers find fabricated video and audio harder to identify as synthetic than fabricated text (Corsi et al., 2024; Guo et al., 2025).

Finding 2: Perceived modality threats vary systematically across domains.

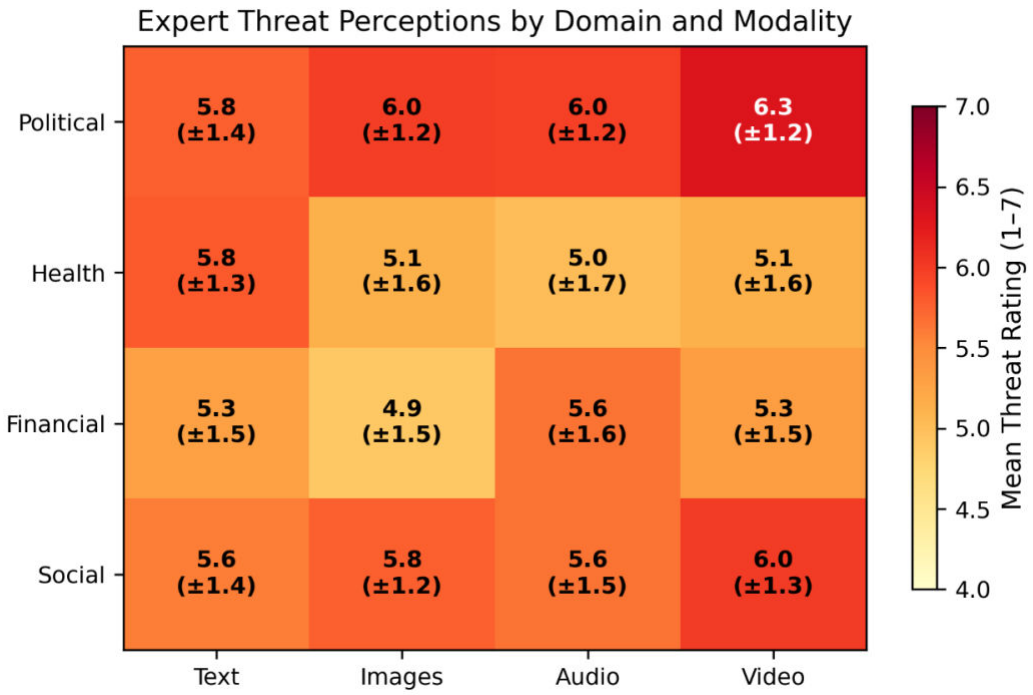

Figure 2. Domain × Modality threat perception heatmap (N = 54). *Cell values show M (±SD). Higher values indicate greater perceived threat.*

Within this sample, the political domain received the highest average threat ratings, with deepfake video peaking at *M* = 6.31 (*SD* = 1.15). The confidence intervals political-domain video and audio overlap, however. In plain terms: with 54 respondents, the difference between these two ratings is small enough that it could plausibly be a product of who happened to answer the survey rather than a real difference in expert opinion. The ordering should be read as suggestive, not established.

The social domain followed a similar pattern (video *M* = 6.00; images *M* = 5.76). The health domain showed a different pattern: text was rated as the most threatening (*M* = 5.80, *SD* = 1.32), while images (*M* = 5.13), audio (*M* = 5.02), and video (*M* = 5.13) clustered lower. However, confidence intervals overlap across all four health-domain modalities, so this pattern should be treated as suggestive rather than definitive. One possible explanation is that health misinformation reaches audiences mainly in written form, from fabricated studies to AI-generated health advice, so experts see text as the more likely route to harm in this domain (De Angelis et al., 2023).

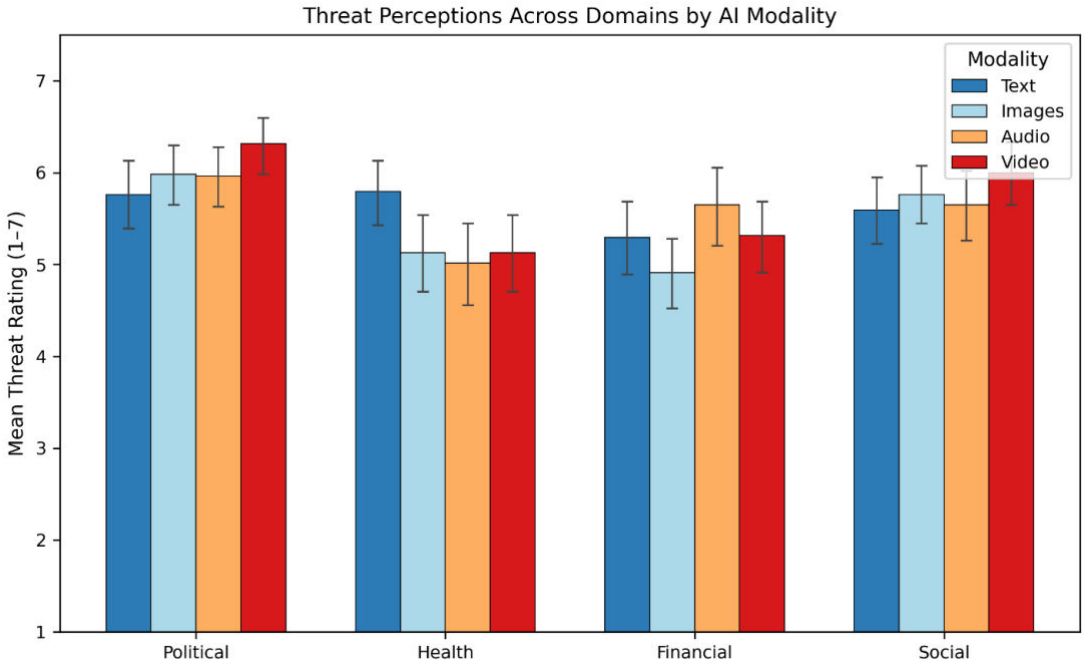

Figure 3. Threat perceptions across domains by AI modality (N = 54). *Error bars represent 95% bootstrap confidence intervals (10,000 iterations). Substantial overlap indicates that most cross-domain and cross-modality differences are not statistically distinguishable at this sample size.*

In the financial domain, audio received the highest mean rating (*M* = 5.65, *SD* = 1.62), though confidence intervals do not clearly separate it from video or text.

Finding 3: Experts identify voice-cloning fraud as the most urgent risk.

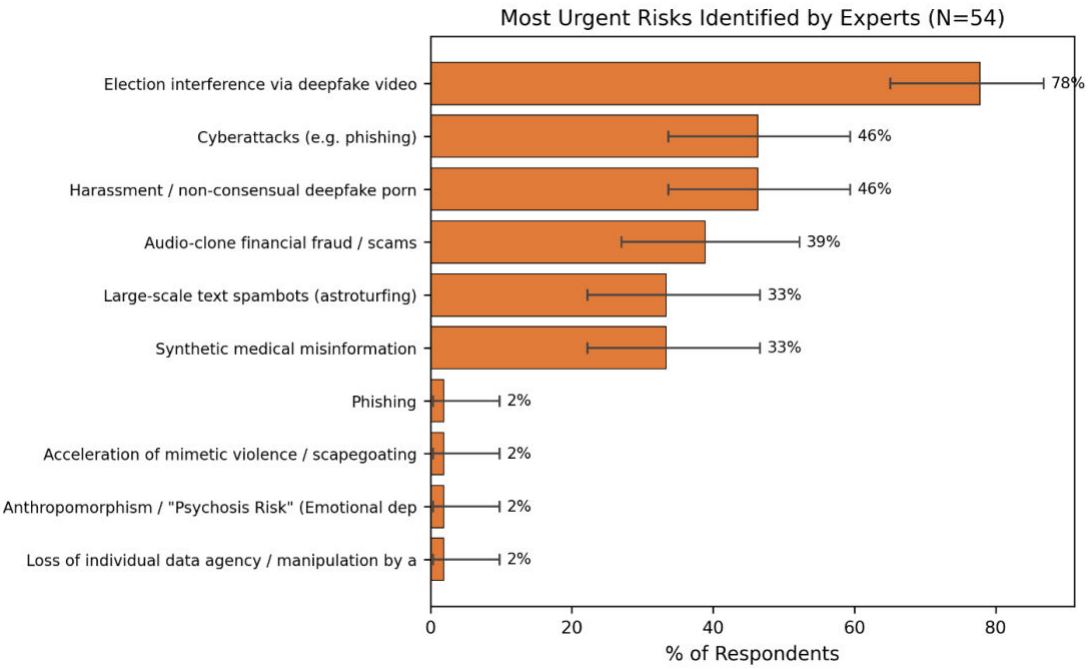

Figure 4. Most urgent risks identified by experts (N = 54). *Error bars represent 95% Wilson score confidence intervals for binomial proportions.*

Election interference via deepfake video was identified as the most urgent risk by 78% of respondents, consistent with widespread public concern documented in recent surveys of U.S. voters (Yan et al., 2025). The next tier comprised cyberattacks (46%) and non-consensual deepfake imagery (46%), though Wilson confidence intervals for these proportions are wide (approximately 33%–60%), indicating substantial uncertainty in the precise ranking below election interference.

Finding 4: Experts view no single mitigation strategy as universally effective.

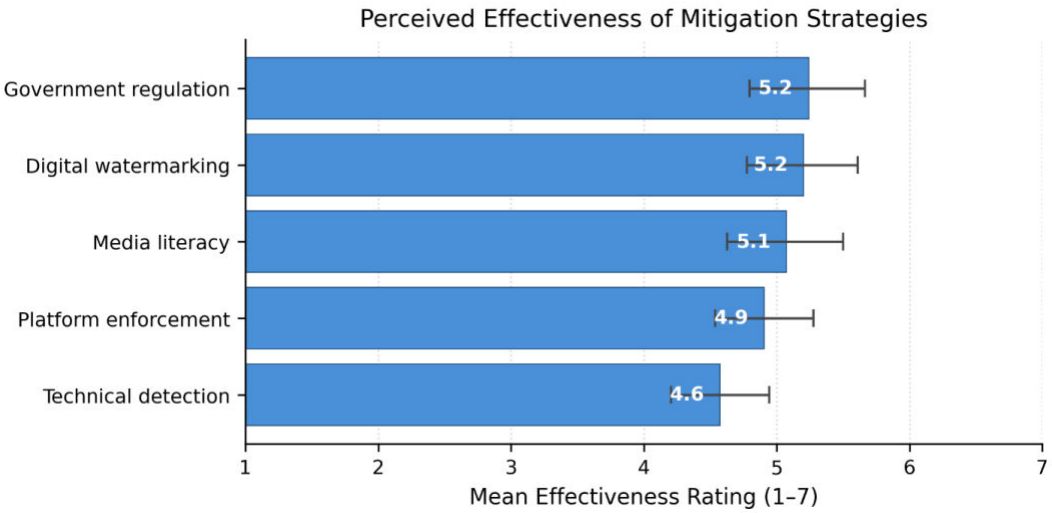

Figure 5. Mitigation strategy effectiveness ratings (7-point scale, N = 54). *Error bars represent 95% bootstrap confidence intervals (10,000 iterations).*

On a 7-point effectiveness scale, government regulation received the highest mean rating (M = 5.24, SD = 1.65), followed by digital watermarking (M = 5.20), media literacy (M = 5.07), platform enforcement (M = 4.91), and technical detection (M = 4.57). Bootstrap confidence intervals overlap for the top four strategies, indicating that the differences among them are not statistically distinguishable; only technical detection falls somewhat below the others. All five distributions are right-skewed and unimodal (63–74% of respondents at ≥ 5; see Appendix Table C2). In a separate item, respondents ranked the same five strategies from 1 (most effective) to 5 (least effective). Mean ranks fell in a narrow band from 2.98 to 3.28, close to the midpoint of 3 that would result from no agreement at all. This aligns with prior expert survey research showing persistent disagreement about optimal responses to misinformation (Altay et al., 2023).

Methods

This study is exploratory. We did not pre-register hypotheses; we sought a broad descriptive map of expert threat assessments across modalities, domains, and mitigation strategies. The present manuscript reports preliminary findings from this dataset.

Going in, we held three loose expectations grounded in the existing literature: (a) higher-fidelity modalities (video, audio) would dominate threat ratings, with video especially salient in the political domain (Corsi et al., 2024); (b) audio threats would cluster in the financial domain, given the rise of voice-cloning fraud (Chen & Magramo, 2024); and (c) experts would split into recognizable camps along a tech-fix-versus-regulation axis (Altay et al., 2023). The data confirmed (a) and (b) but contradicted (c): rather than camp-formation, we observed broad agreement that each strategy can work alongside disagreement

on which deserves priority. We treat this contrast between expectation and finding as a hypothesis source for confirmatory follow-up rather than as a tested claim.

We conducted a structured online survey between July 2025 and March 2026. Eligible participants were professionals or scholars with demonstrated recent activity on AI-driven disinformation; eligibility was operationalized by requiring respondents to provide a link to a recent publication, project, or public engagement on the topic. The link served as a soft validation; no formal post-response screening or exclusion was applied, and all 54 valid responses are retained in the analyses.

Recruitment combined targeted expert recruitment with snowball extension. We identified 516 candidates through systematic Google Scholar, LinkedIn, Mastodon, and Bluesky searches, supplemented by authors of recent publications cited in our broader research program. We directly contacted 454 candidates via individualized messages (email 47%, LinkedIn 21%, Mastodon DM 8%, Bluesky 4%, mixed/other 20%); recipients were invited to forward the survey to relevant colleagues. The questionnaire (54 items, mostly 1-to-7 rating scales) is reproduced in Appendix A; the recoding scheme for free-text role descriptions is documented in Appendix B.

We received 54 valid responses (response rate 11.9% from over 454 directly contacted; the absolute number of forwards via snowball is unknown). The equal number of items (54) and respondents (54) is coincidental. Recoding the open-text role field yielded seven coherent categories: AI/ML researchers and developers (39%); disinformation, fact-checking, journalism, or media research (26%); other practitioners and academics (13%); cybersecurity, defense, and threat-intelligence (6%); policymakers and regulators (6%); provenance, standards, and infrastructure professionals (6%); and ethics, legal, and governance (6%). Respondents had a median of 17 years of experience and high self-rated familiarity with both large language models (M = 4.89/7) and deepfake technology (M = 4.54/7). Respondents were distributed across five geographic areas: European Union (n = 23, 43%), North America (n = 19, 35%), non-EU Europe including the United Kingdom (n = 7, 13%), Asia-Pacific (n = 4, 7%, comprising East and Southeast Asia, South and Central Asia, and Oceania), and one respondent (2%) who selected “Global,” an option offered for experts whose work is not tied to a single region. No respondent was based in Africa or in Latin America, although both were available options. Sample characteristics by category appear in Appendix Table C1.

Participants chose one of three publication-handling preferences at the end of the survey: full anonymity (n = 27, 50.0%), name and affiliation in Acknowledgments (n = 14, 25.9%), or attribution of specific quotes by name in addition to acknowledgment (n = 13, 24.1%). All four respondents quoted by name in this manuscript fall in the third group; “Merzmensch” is a long-standing artistic identity under which the respondent has published creative and scholarly work and was supplied as the preferred attribution.

Quantitative data were analyzed descriptively (means, standard deviations, frequencies). Consequently, the reported differences should be interpreted as descriptive patterns rather than evidence of statistically significant differences between modalities or domains. Given the exploratory design and sample size, we report descriptive statistics rather than inferential tests. To convey estimation uncertainty, figures display 95% bootstrap confidence intervals (10,000 iterations with a fixed random seed for reproducibility) for mean ratings, and 95% Wilson score confidence intervals for binomial proportions.

Limitations

The sample (N = 54) is purposive, not representative, and skews toward Western, technically proficient experts: 91% of respondents are based in the European Union, North America, or non-EU Europe, with the Global South substantially underrepresented. Self-selection bias may inflate threat perceptions among respondents already concerned about AI-driven disinformation. The survey measures perceptions

rather than observed impacts, and the rapidly evolving capabilities of generative AI may shift expert assessments substantially within months. Sub-domain comparisons are descriptive only and underpowered for inferential testing.

Because the sample contains a comparatively large proportion of AI researchers and disinformation specialists, the observed priorities likely reflect expert perspectives on AI-related risks rather than broader societal perceptions. This composition may have influenced which domains were viewed as most concerning.

All respondents were 18 or older and provided informed consent prior to participation. The consent form, displayed at the start of the survey, described the study purpose, voluntary nature of participation, anticipated time commitment, data-handling procedures, and the three publication-handling options described above. The survey did not collect special-category personal data and was conducted in accordance with the EU General Data Protection Regulation. The study protocol was reviewed against the data-minimization and lawful-basis requirements of the EU General Data Protection Regulation (consent under Art. 6(1)(a) GDPR). Direct quotations are attributed by name only for the 13 respondents who explicitly opted in to such attribution.

Bibliography

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School (HKS) Misinformation Review*, *4*(4). https://doi.org/10.37016/mr-2020-119

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2024). *The 2024 Foundation Model Transparency Index*. arXiv. https://arxiv.org/abs/2407.12929

Chen, H., & Magramo, K. (2024, February 4). *Finance worker pays out $25 million after video call with deepfake 'chief financial officer.'* CNN. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753–1820. https://doi.org/10.2139/ssrn.3213954

Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. *Harvard Kennedy School (HKS) Misinformation Review*, *5*(3). https://doi.org/10.37016/mr-2020-140

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, Article 1166120. https://doi.org/10.3389/fpubh.2023.1166120

European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* EUR-lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Financial Crimes Enforcement Network. (2024, November 13). *FinCEN alert on fraud schemes involving deepfake media targeting financial institutions (FIN-2024-Alert004)*. U.S. Department of the Treasury. https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf

Guo, S., Zhong, Y., & Hu, X. (2025). People are more susceptible to misinformation with realistic AI-synthesized images that provide strong evidence to headlines. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(6). https://doi.org/10.37016/mr-2020-189

Hoq, A., Facciani, M. J., & Weninger, T. (2025). Feedback and education improve human detection of image manipulation on social media. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(2). https://doi.org/10.37016/mr-2020-175

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016–2022). *Harvard Kennedy School (HKS) Misinformation Review*, *4*(6). https://doi.org/10.37016/mr-2020-130

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

Roozenbeek, J., & van der Linden, S. (2024). *The psychology of misinformation.* Cambridge University Press. https://doi.org/10.1017/9781009214414

van der Linden, S. (2023). *Foolproof: Why misinformation infects our minds and how to build immunity.* W. W. Norton.

Weikmann, T., Wouters, F., Tulin, M., Hameleers, M., de Vreese, C. H., Zarouali, B., & Opgenhaffen, M. (2026). On the same page? Experts are mostly, but not always aligned about disinformation in times of generative AI. *Harvard Kennedy School (HKS) Misinformation Review, 7*(2). https://doi.org/10.37016/mr-2020-196

World Health Organization. (2020, September 23). *Managing the COVID-19 infodemic: Framework for managing the COVID-19 infodemic*. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

Yan, H. Y., Morrow, G., Yang, K.-C., & Wihbey, J. (2025). The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(1). https://doi.org/10.37016/mr-2020-171

Acknowledgements
We gratefully acknowledge the 54 experts who contributed their time and expertise to this survey. We especially thank the following respondents who consented to be named: Marc-Oliver Pahl (IMT Atlantique), Gürkan Solmaz (ACM Europe TPC), Gerhard Schimpf, Volker Klaeren, Merzmensch, Arne Stenmanns, Melanie Siegel (Darmstadt University of Applied Sciences), Filippo Menczer (Observatory on Social Media, Indiana University), Jeff Jarvis (CUNY), Stephan Lewandowsky (University of Bristol), Emilio Ferrara (University of Southern California), Timothy R. Levine (University of Oklahoma), Jesper Strömbäck (University of Gothenburg), Giovanni Spitale (University of Zurich), Eelco Herder (Utrecht University), Alberto Lobato Diogo, Sam Stockwell (The Alan Turing Institute), Evan M. Williams (Carnegie Mellon University), Peter Carragher (Carnegie Mellon University), Manju Rose Mathews (Christ Nagar College), Katerina Sedova (Atlantic Council), William Kempster (ar.io network), Scott Perry (Digital Governance Institute), Doug Finke, Roberto V. Zicari (Z-Inspection® Initiative), Jason Potel (Goldsmiths, University of London), and Muhammad Zubair (Fortex Solutions).

Authorship
A. L. conceived and designed the study, developed the survey instrument, conducted data collection, performed the analysis, and wrote the manuscript. M. K. and M.-O. P. provided supervision and critical feedback on the study design and manuscript. All authors reviewed and approved the final version.

Funding
No funding has been received to conduct this research.

Competing interests
The authors declare no competing interests.

Ethics
This study was conducted in accordance with ethical guidelines for human subjects research. All participants provided informed consent prior to completing the survey, confirming they were 18 years or older and understood the purpose, voluntary nature, and data handling procedures of the study. Participants could withdraw at any time without consequences. The survey collected professional opinions on AI-generated disinformation threats and mitigation strategies; no sensitive personal data, medical information, or vulnerable populations were involved. Respondents self-selected their attribution preference, choosing whether to remain anonymous, be acknowledged by name, or allow direct quotation. Ethnicity and gender were not collected, as they were not relevant to the research questions. The study was reviewed and approved by the supervising institution. Direct quotations are attributed by name only for the 13 respondents who explicitly opted in to such attribution. One of these attributions uses a pseudonymous professional identity ("Merzmensch"), an established public byline that we verified in direct correspondence with the respondent rather than through institutional affiliation.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/BXO2QA. The deposit includes: (i) the verbatim survey instrument as administered (also reproduced in Appendix A below), (ii) the aggregated, de-identified response tables underlying all figures and supplementary tables, (iii) the analysis code used to produce them, and (iv) the recoding scheme for the open-text role field (Appendix B). Individual-level free-text responses are withheld within IRB restrictions and in keeping with respondents' anonymity preferences (50.0% of participants opted for full anonymity); only aggregated counts and consented quotations are released.

Appendix A: Survey instrument

The instrument was administered in English via Google Forms between July 3, 2025, and March 15, 2026. It is organized in nine sections plus consent, future-outlook, and attribution blocks. Several items use a modality- or strategy-keyed grid (e.g., the four threat-domain matrices each elicit 16 modality × domain ratings), which is how the 34 numbered questions yield the 54 substantive data points per respondent referenced in the main text. Items marked * were required; all others were optional. Likert-type response options are stated once per scale. The full LaTeX source of the instrument is included in the Harvard Dataverse deposit (see Data Availability section); the structured summary below preserves item wording, ordering, and response options.

A.1 Consent & data protection (1 item). GDPR notice naming the data controller (Alexander Loth, Frankfurt UAS), purpose, lawful basis (explicit consent), retention, and participant rights (withdrawal, access, rectification, erasure, complaint to the Hessian Data Protection Commissioner). *Q1** — "I confirm I am 18 years or older and I have read, understood, and agree to the consent terms outlined above." — *Yes / No* (gating).

A.2 Section 1/9 — Screening & expert group (2 items). *Q2** Primary professional role — *single choice:* AI researcher/ML engineer; disinformation/fact-checking professional; journalist covering tech or politics; policymaker/regulator; ethicist/legal scholar; other (free text). Recoded post hoc into seven categories (see Appendix B). *Q3** Link to a recent activity (recent paper, organizational role, profile link) — *free text.*

A.3 Section 2/9 — Demographics & baseline knowledge (4 items). *Q5** Years of professional experience — *numeric. Q6** Region of primary professional activity — *single choice:* European Union; Europe (non-EU, incl. UK); North America (USA & Canada); East & Southeast Asia; South & Central Asia; Latin America & Caribbean; MENA; Sub-Saharan Africa; Oceania; other. *Q7** Self-rated technical understanding of Large Language Models — *1–7 (Novice–Expert). Q8** Self-rated familiarity with deepfake video technology — *1–7 (Novice–Expert).*

A.4 Section 3/9 — Perceived threats, modality only (1 item). *Q9.* "How threatening is each modality for spreading disinformation in general?" — *1–7 grid (Not a threat–Severe threat) for:* AI-generated text, images, audio (voice cloning), video (deepfakes).

A.5 Section 4/9 — Perceived threats, modality × domain matrix (4 rated items + 4 optional comments). Stem (repeated per domain): "Threats in the **{DOMAIN}** Domain (1 = Not a threat, 7 = Severe threat)." Domains: **Political** (Q10), **Health** (Q12), **Financial** (Q14), **Social** (Q16). Each is a 4-row × 7-column grid (text/images/audio/video). Each domain is followed by an *optional* free-text comment item (Q11, Q13, Q15, Q17): "Which specific points concerning the {DOMAIN} Domain would you like to share with us?"

A.6 Section 5/9 — Key risks, top-three (1 item + 1 optional). *Q18** Select up to three risks judged most urgent — *multi-select (max 3):* election interference via deepfake video; large-scale text spambots (astroturfing); audio-clone financial fraud/scams; synthetic medical misinformation; harassment/non-consensual deepfake porn; cyberattacks (e.g., phishing); other (free text). *Q19.* In 1–2 sentences, why these worry you most — *optional, free text.*

A.7 Section 6/9 — Mitigation strategies, ratings (1 item + 1 optional). *Q20.* "Rate the effectiveness of the following mitigation strategies" — *1–7 grid (Not effective–Very effective) for:* technical detection tools; digital watermarking/provenance standards; government regulation (e.g., AI Act, DSA); media/public literacy programs; platform enforcement policies. *Q21.* Additional information on mitigation strategies — *optional, free text.*

A.8 Section 7/9 — Mitigation strategies, forced ranking (1 item). *Q22.* Rank the five strategies above from 1 (most effective) to 5 (least effective) — *full ranking, no ties.*

A.9 Section 8/9 — Best–worst scaling (3 items). *Q23** Which strategy is **most** effective? — *single choice from the same five strategies. Q24** Which strategy is **least** effective? — *single choice from the same five strategies. Q25.* Single biggest obstacle (technical, political, or economic) blocking adoption of your top-ranked strategy — *optional, free text.*

A.10 Section 9/9 — Public-literacy tools (4 items). A short methodology note describes a representative tool, JudgeGPT (https://github.com/aloth/JudgeGPT), in which respondents view five short news fragments and make two slider-based judgments per fragment (origin: human vs. AI; veracity: legitimate vs. fake), with accuracy feedback and badges. *Q26.* Effectiveness of such public-awareness tools for mitigating AI-generated fake news — *1–7 (Not Effective–Very Effective). Q27** Single greatest limitation — *single choice:* easily bypassed by adversarial content; reaches only tech-savvy audiences; impact fades once novelty wears off; resource-intensive to maintain; other (free text). *Q28.* Main strength of such tools — *optional, free text. Q29.* One concrete feature or program that would make public-resilience tools more effective — *optional, free text.*

A.11 Future outlook & final comments (2 items). *Q30.* One emerging or currently underestimated risk at the AI–disinformation frontier on a five-year horizon — *optional, free text. Q31.* Final thoughts or feedback on the survey — *optional, free text.*

A.12 Attribution & final submission (4 items). *Q32** Handling of contribution in the final publication — *single choice:* keep responses fully anonymous; list name and affiliation in the Acknowledgments; in addition, attribute specific quotes or ideas by name. *Q33** Full name — *free text. Q34** Affiliation — *free text. Q35.* Link to preferred professional/academic profile — *optional, free text.*

Appendix B: Recoding scheme for professional roles

The survey collected primary professional role as a free-text field. Thirty-two distinct strings were submitted. We recoded these into seven coherent categories using the rules below, applied conservatively to preserve respondent self-description.

Table B1. *Recoding scheme mapping open-text professional role descriptions to seven coherent categories (*N *= 54).*

Category (n)	Includes raw strings such as
AI/ML researcher or developer (21)	"AI researcher/ML engineer," "Software developer, deep in the AI space," "AI trainer," "Professor for semantic technologies and computer science," "Economist working w AI," "Trustworthy AI co-creation and assessment," "Product management," "IT architect"
Disinformation, fact-checking, journalism, or media research (14)	"Disinformation/fact-checking professional," "Journalist covering tech or politics," "Misinformation academic researcher," "Communications researcher," "Journalism scholar researching the impact of AI," "Media psychology researcher," "University researcher of information ecosystems," "Researcher on disinformation, AI, etc.," "Deception researcher and author of TDT"
Other practitioners & academics (7)	"Teacher," "Private person (IT consultant)," "Culture and art history researcher," "Investor/venture capitalist (focus on AI and defense tech)," "Academic," "Technical training"
Cybersecurity/defense/threat intelligence (3)	"Professor cybersecurity," "Subject matter expert and researcher of AI-enabled threats and opportunities in cyber, information environment, and defense technology"
Policymaker/regulator (3)	"Policymaker/regulator involved with AI, media or digital-services policy," "Member of the ACM Europe Technology Policy Committee"
Provenance/standards/infrastructure (3)	"Infrastructure provider supplying verifiable data storage and access," "Advocate for adoption of C2PA provenance technologies," "Technologist/web standards advocate (focus on decentralization and data sovereignty)"
Ethics, legal, or governance (3)	"Ethicist/legal scholar working on AI or deepfakes," "Governance consultant and C2PA conformance program administrator"

Note: Unique role strings: 32. All recoded; none discarded.

Appendix C: Supplementary statistical tables

Table C1. *Sample characteristics of the expert respondents (N = 54).*

Characteristic	Value
Region — European Union	23 (42.6%)
Region — North America	19 (35.2%)
Region — Non-EU Europe (incl. UK)	7 (13.0%)
Region — Asia-Pacific / Global	5 (9.3%)
Years of professional experience — median	17
Years — range	1–65
Self-rated LLM understanding (1–7) — mean	4.89
Self-rated deepfake familiarity (1–7) — mean	4.54
Attribution preference — anonymous	27 (50.0%)
Attribution preference — acknowledgment only	14 (25.9%)
Attribution preference — quote-attribution opt-in	13 (24.1%)

Note: Descriptive statistics, distributional diagnostics, recruitment details, and the domain-level policy priority matrix.

Table C2. *Distribution shape of mitigation effectiveness ratings across strategies (N = 54).*

	Rating											
Strategy	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Mean**	***SD***	**High %**	**Mid %**	**Low %**
Government regulation	2	2	5	5	13	12	15	5.24	1.65	74.1	9.3	16.7
Digital watermarking / provenance	1	3	2	11	11	13	13	5.20	1.53	68.5	20.4	11.1
Media / public literacy	0	5	6	9	9	10	15	5.07	1.67	63.0	16.7	20.4
Platform enforcement	0	4	4	12	15	11	8	4.91	1.42	63.0	22.2	14.8
Technical detection	1	3	8	13	13	13	3	4.57	1.40	53.7	24.1	22.2

Note: Likert votes (count per rating) per strategy, with summary share of high (≥ 5), mid (= 4), and low (≤ 3) ratings. All five distributions are right-skewed and unimodal. None exhibits the bimodal pattern that would indicate genuine effectiveness polarization. Disagreement in the forced-choice rankings therefore reflects priority-setting rather than effectiveness disagreement.

Table C3. *Recruitment funnel from candidate identification to valid responses.*

Stage	*n*
Candidates identified	516
Directly contacted (any channel)	454
Channels — Email	214
Channels — LinkedIn	95
Channels — Mastodon DM	37
Channels — Bluesky / X DM	23
Channels — Mixed / other	85
Survey responses received	54
Response rate (responses / contacted)	11.9%

Note: Snowball forwards from contacted candidates to colleagues are not captured in the denominator and are unknown.

Table C4. *Policy priority matrix showing perceived threat by domain and modality (N = 54).*

Domain	Primary modality threat (*M*)	Priority lever	Supporting strategies
Political	Video deepfake (6.31)	Audiovisual provenance (e.g., C2PA) + platform enforcement	Statutory regulation of election-period synthetic media; rapid-response fact-checking; literacy targeted at electoral periods
Health	AI-generated text (5.80)	Labelling of AI-generated health content + medical fact-checking infrastructure	Clinician-facing literacy on LLM hallucinations; platform demotion of unverified medical claims; regulator coordination (e.g., FDA/EMA guidance)
Financial	Audio deepfake (5.65)	Voice-authentication standards for high-value transactions + KYC modernization	Sector-specific regulation of synthetic-voice fraud; cross-bank threat-intelligence sharing; staff training
Social (cross-cutting)	Mixed (video 6.00 / images 5.76)	Accessibility-first literacy reaching non-technical audiences	Provenance signals embedded in mainstream platforms; transparent platform moderation; community-driven counter-speech

Note: A preliminary, operational mapping from the descriptive threat patterns observed in this study to a prioritized mitigation stack per domain. The matrix is intended as a starting point for policymakers and platform operators rather than a closed framework; cells should be re-weighted as the AI capability landscape evolves. Cell content draws on respondent ratings (Findings) and on existing instruments (Bommasani et al., 2024; Chesney & Citron, 2019; Corsi et al., 2024; De Angelis et al., 2023; European Parliament, 2024). The matrix operationalizes the layered, contested-not-polarized mitigation landscape identified in the Implications section.