The Mitigation Paradox: How AI-Disinformation Experts Converge on Threats but Polarize on Solutions

Alexander Loth[1], Mina Schütz[2], Martin Kappes[1], and Marc-Oliver Pahl[3]

[1]Frankfurt University of Applied Sciences, Germany
[2]AIT Austrian Institute of Technology GmbH, Austria
[3]IMT Atlantique, UMR IRISA, Chaire Cyber CNI, France
{alexander.loth@stud.fra-uas.de, mina.schuetz@ait.ac.at, kappes@fra-uas.de, marc-oliver.pahl@imt-atlantique.fr}

Preprint. Accepted at the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2026), Malmö, Sweden, October 12–14, 2026.
Survey data: https://doi.org/10.5281/zenodo.18703600

Abstract

Policy on AI-generated disinformation is being shaped before the expert community advising it has converged on what works. We report a structured survey of 54 experts across 21 countries (Jul 2025–Mar 2026) designed to map disagreement, not just opinion. Threat assessments converge sharply: 78% rate election-related video deepfakes as the single most urgent risk, and every modality–domain pair is rated above 5/7. Mitigation assessments do not: media literacy is the only strategy whose distribution is genuinely bimodal (BC=0.559), and government regulation is the most-chosen "most effective" (30%) yet is rated least effective by 15% of the same panel. We term this gap between converged threat perceptions and polarized mitigation preferences the *Mitigation Paradox*. Group comparisons further reveal that non-EU respondents rate financial-domain voice cloning as significantly more threatening than EU respondents (p=0.024), and rate technical detection as a more effective mitigation (p=0.060). Combining disagreement metrics with LLM-assisted thematic coding of open-ended responses, we identify three candidate epistemic orientations—infrastructure-oriented, governance-oriented, and harm-proximity—whose threat models appear to diverge; given the sample size we present these as exploratory. The Mitigation Paradox argues for layered, portfolio-style governance rather than single-strategy commitments.

1 Introduction

Despite growing urgency, experts fundamentally disagree on how to respond to AI-generated disinformation—even as they broadly agree on what the threats are. We call this asymmetry, in which converged threat perceptions coexist with polarized mitigation preferences, the *Mitigation Paradox*. Policy is therefore being formulated in a partial vacuum of expert consensus, even though those same experts are frequently consulted in regulation and platform governance. Both the EU AI Act and the Digital Services Act explicitly oblige providers and very large platforms to conduct systemic risk assessments—an obligation whose practical execution presupposes an expert consensus that, as we show, does not yet exist. The central question is therefore no longer only *what* the threats are, but *who* sees which threats, and why their views diverge.

Recent empirical work confirms that generative AI is already a measurable factor in online information environments. Chen et al. (2025) report that approximately 12% of political images shared on X during the 2024 U.S. Presidential Election were AI-generated, with dissemination concentrated in a small group of highly active accounts. Drolsbach and Pröllochs (2025) further show that AI-generated misleading content can spread more effectively than conventional misinformation, even when originating from smaller accounts, while being perceived as similarly credible and harmful. Prior expert surveys document opinion on misinformation broadly (Altay et al., 2023) or on GenAI disinformation in specific regions (Weikmann et al., 2026), and structured elicitations exist in adjacent domains (Kruger et al., 2024). None, however, take expert *disagreement* on AI-generated disinformation as the primary analytic object.

This paper closes that gap. We elicit responses from 54 experts in 21 countries (Jul 2025–Mar 2026), combining 7-point Likert ratings across four modalities and four threat domains, a forced-ranking and best–worst-scaling task on five mitigation strategies, and open-ended prompts on emerging risks. Quantitatively, we move beyond mean comparisons by reporting dispersion (coefficient of variation, CV) and bimodality coefficient (BC) statistics that distinguish broad uncertainty from genuine polarization. Qualitatively, we use LLM-assisted thematic coding with source-traceable quotes to compare positions across professional role, region, and self-rated technical understanding.

We address the following research questions:

1. *On which aspects of AI-generated disinformation do experts show consensus, and where do significant disagreements emerge?*
2. *Do threat perceptions and mitigation preferences differ systematically across professional expertise groups (e.g., AI researchers, journalists, policymakers)?*
3. *Do threat perceptions and mitigation preferences vary by geographic region (e.g., EU vs. non-EU)?*
4. *What emerging risks do experts identify that are not yet addressed by current policy frameworks?*

Contributions. This paper makes three contributions:

- **Empirical:** the first disagreement-focused map of expert assessments on AI-generated disinformation, quantifying polarization across five mitigation strategies via the bimodality coefficient and identifying three candidate epistemic orientations—infrastructure-oriented, governance-oriented, and harm-proximity—whose threat models appear to diverge. Given the sample size and uneven subgroup coverage, we frame the subgroup and epistemic-orientation results as exploratory.
- **Methodological:** a reproducible workflow that combines dispersion-aware disagreement metrics with LLM-assisted thematic coding of open-ended responses, with every model-generated quote traceable to an individual respondent in the raw data.
- **Policy-relevant:** a translation of disagreement signals into a portfolio-vs.-single-strategy decision rule for AI-disinformation governance, including conditions under which monitoring, pilots, or binding regulation are the appropriate next step.

We focus on AI-generated disinformation and do not attempt to measure the full spectrum of non-GenAI misinformation threats.

The remainder of the paper is structured as follows. Section *Related Work* reviews related work on AI-generated disinformation, expert surveys, and mitigation strategies. Section *Methodology* details our survey design, sample, and analysis methods. Section *Findings* presents the quantitative and qualitative findings. Section *Discussion* discusses governance implications, limitations, and directions for future work. Section *Conclusion* concludes.

2 Related Work

The broader problem of misinformation and disinformation has grown substantially with the scale and speed of online information exchange, making deceptive or manipulated content increasingly difficult to identify and assess reliably Zhou and Zafarani (2018); Ceron et al. (2021); Kozik et al. (2022). False narratives, propaganda, and misleading information have been shown to affect journalism, public trust, and democratic processes, while increasing dependence on digital and social media platforms further amplifies their reach and societal impact Hyzen (2021); Khan et al. (2019). Existing research suggests that, despite growing concerns surrounding generative AI, many real-world misinformation and disinformation ecosystems are still predominantly driven by human actors and established influence networks rather than fully AI-generated campaigns Ng et al. (2024); Steffen (2025). Analyses of recent election-related disinformation further indicate that AI-generated content currently represents only a comparatively small proportion of detected misleading material online, with manipulated or manually produced narratives remaining more common European Digital Media Observatory (2024). Nevertheless, several studies and policy reports identify AI-generated audio and voice-cloning technologies as particularly concerning due to their realism, accessibility, and potential use in fraud, political manipulation, and election interference European Digital Media Observatory (2024). Consequently, generative AI is increasingly understood not as a replacement for existing disinformation structures, but as an enabling and amplifying technology that may significantly increase the scalability and persuasive capabilities of harmful narratives.

2.1 AI-Generated Disinformation: Capabilities and Threats

Generative AI now spans four practically distinct attack surfaces. Deepfake video has been the most publicly discussed vector since Chesney and Citron (2019) mapped its legal and political ramifications, with subsequent surveys cataloguing rapidly improving generation and partially keeping detection methods (Rana et al., 2022). Large language models extend the threat to long-form text, where machine-generated political and scientific content is increasingly indistinguishable from human-authored material (De Angelis et al., 2023). Synthetic media at scale propagates through social platforms in patterns that differ from human-authored misinformation in both reach and timing (Corsi et al., 2024), and multimodal threats (audio cloning, synthetic imagery) introduce attack surfaces that prior text-centric work does not address.

2.2 Expert Surveys on Misinformation

Two prior survey efforts directly inform our framing. Altay et al. (2023) elicited definitions and proposed solutions from roughly 150 researchers and practitioners, surfacing wide variation in how the field operationalizes "misinformation" itself. Weikmann et al. (2026) narrows the lens to European experts on GenAI disinformation specifically. Kruger et al. (2024) use a Delphi-style structured elicitation to rank strategies, barriers, and research priorities for disinformation, misinformation, and malign influence—demonstrating that expert panels themselves can be used as instruments for surfacing structured disagreement. None of these prior studies, however, take disagreement itself

as the primary analytic object for AI-generated disinformation. That gap motivates the present work.

2.3 Mitigation Approaches

Prior work emphasizes multiple mitigation pathways, including regulatory frameworks (e.g., the EU AI Act (European Parliament, 2024)), technical detection and its adversarial arms race (Rana et al., 2022), and psychological inoculation or prebunking strategies (van der Linden, 2023). Media literacy interventions remain a prominent policy proposal, with mixed evidence on sustained effectiveness (Hoq et al., 2025; Murphy et al., 2023). Platform governance and content moderation are frequently cited as operational levers but remain contentious due to trade-offs between scale, accuracy, and free expression. In applied settings, much harmful disinformation still appears to be human-authored—particularly in extremism and conspiracy communities—which suggests that AI-specific mitigation should be considered alongside broader disinformation countermeasures. Our contribution is to evaluate these strategies through expert-rated effectiveness and to characterize the degree of polarization surrounding each approach.

3 Methodology

3.1 Survey Design

We administered an online survey consisting of 54 items grouped into five sections: demographics, threat perceptions (4 modalities $\times$ 4 domains), urgent risks, mitigation effectiveness, and qualitative open-ended prompts. Threat perceptions and mitigation ratings were captured on 7-point Likert scales, complemented by a forced-ranking question to surface perceived urgency among threat categories. The qualitative items invited respondents to describe emerging risks and to justify their ratings in their own words. The survey instrument is provided in the supplementary materials and was piloted with three domain experts to check clarity and wording.

3.2 Recruitment and Sample

We used purposive sampling to target individuals with demonstrable expertise in AI, disinformation, or policy. Invitations were drawn from authors of relevant publications (Semantic Scholar, Google Scholar), professional networks, and public professional profiles. The final sample includes 54 respondents collected between July 2025 and March 2026 across 21 countries. The geographic distribution is EU (43%), North America (35%), non-EU Europe (13%), and Asia-Pacific/other regions (9%). Professional roles include AI researchers (30%), disinformation professionals (9%), journalists (4%), policymakers (4%), ethicists (4%), and other AI professionals (49%). Average professional experience was approximately 20 years, and self-reported AI familiarity was high (LLMs M=4.89/7; deepfakes M=4.54/7). Table 1 summarizes the sample composition.

3.3 Analysis

We use a mixed-methods analysis. Quantitatively, we report descriptive statistics (means, standard deviations, and dispersion measures) and compare groups using non-parametric tests where appropriate (e.g., Kruskal–Wallis) given the small sample sizes, reporting effect sizes where feasible. Disagreement is characterized using dispersion metrics such as the coefficient of variation (CV) and the bimodality coefficient (BC) to capture polarization beyond mean differences. CV

Table 1: Expert sample overview (N=54).

Category	Share
EU	43%
North America	35%
Non-EU Europe	13%
Other regions	9%
AI researchers	30%
Other AI professionals	49%
Disinformation professionals	9%
Journalists	4%
Policymakers	4%
Ethicists	4%

captures relative dispersion around the mean, while BC—following the SAS formulation that combines skewness and excess kurtosis—flags multimodal response distributions; BC values above 5/9 (≈ 0.555) are conventionally interpreted as evidence of bimodality. Together these metrics allow us to distinguish broad uncertainty from true polarization. Qualitative responses were analyzed using inductive thematic coding, iteratively refined until thematic saturation, and inter-coder agreement is reported in the supplementary materials.

3.4 LLM-Assisted Analysis of Open-Ended Responses

We use GPT-4o-mini OpenAI et al. (2024) at temperature 0.1 to systematically compare open-ended responses across expert groups. Responses are grouped along three dimensions: (a) professional role (Technical, Policy/Regulatory, Journalism/Disinformation, Academic/Research), (b) region (EU/EEA, North America, Asia-Pacific, Other), and (c) self-rated LLM technical understanding (High 6–7, Medium 4–5, Low 1–3). For each open question with at least two groups providing two or more responses, the model returns a structured JSON object identifying shared views (*agreements*), substantive differences with group positions and verbatim or near-verbatim quotes (*disagreements*), and concerns raised exclusively within one group (*group-unique concerns*). A parallel overall-synthesis prompt yields cross-question summaries, group profiles, and key tensions. Each response in the prompt is prefixed with a source identifier (the original CSV row number) so every model-generated quote can be traced back to the raw data. The structured output serves as an interpretive aid; all reported claims are verified against the original responses.

Human validation of the LLM-assisted coding. The model output was treated as a first-pass interpretive aid, not as ground truth, and every claim we report was checked by the authors against the raw responses. Two steps were applied. First, *quote verification*: each of the model-surfaced quotes reproduced in this paper was matched back, via its `[Rn]` row reference, to the exact free-text answer in the source CSV, and any quote that was not a verbatim or near-verbatim match (minor whitespace or truncation only) was discarded or corrected; quotes that could not be traced to a single consenting respondent were removed. Second, *theme validation*: two authors independently reviewed the model-proposed agreements, disagreements, and group-unique concerns against the underlying responses and retained only themes that both coders judged to be supported by at least two source responses; disagreements between coders were resolved by discussion, and model-proposed themes without sufficient textual support were dropped rather than reconciled. Inter-

Table 2: Headline findings by domain, with key challenges for governance (N=54).

Domain	Finding	Key challenge
Cross-domain	Video deepfakes rated most threatening modality overall (M=6.19, SD=1.20); all modalities rated above 5/7	Severity consensus, but no agreement on which lever to pull
Political	Video deepfakes dominate (M=6.31); election interference is the #1 urgent risk (78% of respondents)	Last-minute "deepfake dump" tactics around election dates
Health	Threat hierarchy inverts: text leads (M=5.80), audio lowest (M=5.02)	Long-tail medical claims at LLM scale evade fact-check capacity
Financial	Audio dominates (M=5.65), driven by voice-cloning fraud scenarios	Fraud-at-scale via personalized one-to-one attacks
Social	Video again leads (M=6.00) with text and audio close behind	Harassment / non-consensual deepfake porn cited by 46% of respondents
Mitigation	Government regulation (M=5.24) and watermarking (M=5.20) rated highest mean effectiveness; technical detection lowest (M=4.57)	Media literacy is polarizing (BC=0.559, above the 5/9 bimodality threshold)

coder agreement on the retained thematic structure is reported in the supplementary materials. Because the sample and the number of open-ended responses per subgroup are small, we present the qualitative groupings as *exploratory* rather than as validated typologies.

3.5 Limitations

This study relies on a purposive sample with a Western and technical skew, which may overrepresent certain epistemic communities and underrepresent others. Self-selection bias is possible, and the results capture perceived threats rather than observed impacts. The rapidly evolving nature of generative AI means that expert perceptions may shift faster than survey cycles. The present scope is restricted to AI-generated disinformation specifically and does not attempt to measure the full spectrum of non-GenAI misinformation threats. We therefore interpret the findings as a snapshot of expert disagreement during a specific period rather than a definitive account of long-term consensus.

4 Findings

The findings are presented in two stages. First, we analyze the quantitative survey results derived from closed-ended questions, focusing on distributions, group differences, and overall trends in expert assessments of AI-generated disinformation risks, impacts, and mitigation strategies. Second, we examine the qualitative responses from the open-ended questions to explore how experts justify, contextualize, and interpret these assessments in greater depth. This combined approach allows us to compare measurable patterns in the survey data with the underlying reasoning, concerns, and disagreement structures expressed by participants. Table 2 summarizes the headline results by domain.

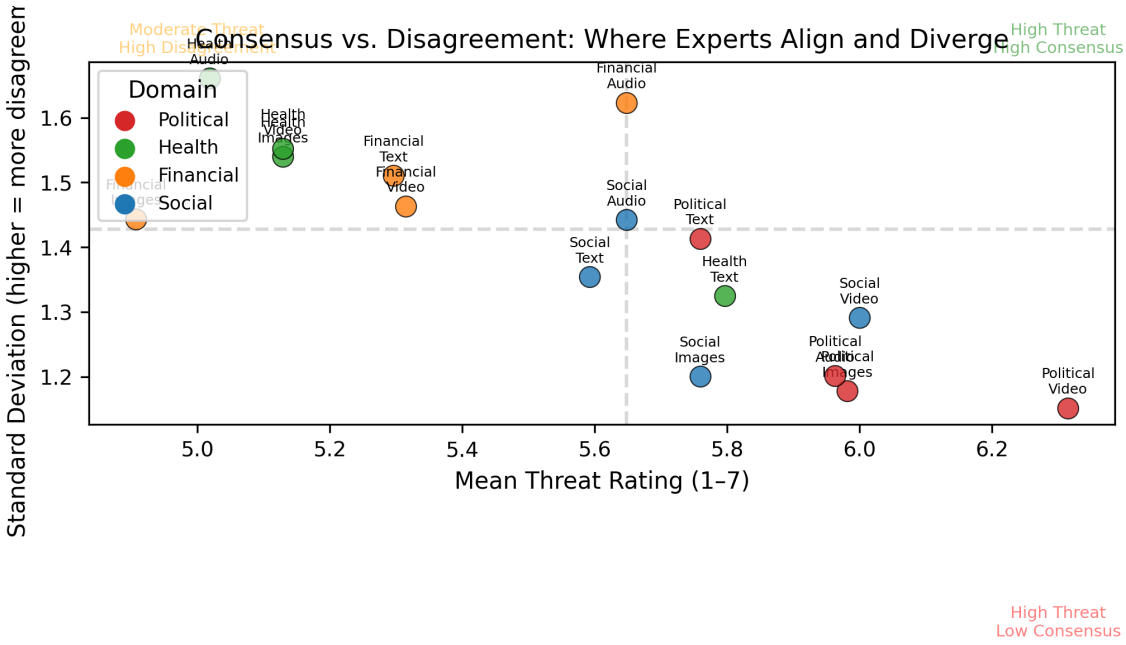


Figure 1: Consensus-disagreement map of expert threat perceptions. Each point represents a domain-modality combination. High mean + low SD (bottom-right) indicates consensus on high threat; high SD (top) indicates disagreement.

4.1 Consensus: Where Experts Agree

Across the four domains and four modalities surveyed, three patterns of agreement stand out. First, experts converge on video deepfakes as the most threatening modality overall (M=6.19 on the 1–7 scale, SD=1.20), with the political domain pushing this value to M=6.31. Second, on the question of which single risk is most urgent, election interference via deepfake video was selected by 78% of respondents, far ahead of any other category—a level of agreement rarely seen elsewhere in the survey. Third, every modality–domain cell in our matrix received a mean rating above 5/7, indicating that no respondent group dismissed any modality as low-threat. The consensus map in Figure 1 visualizes this: the bottom-right cluster, indicating high mean and low standard deviation, is densely populated by political modalities, while the upper part of the plot—where standard deviations exceed 1.5—is occupied by mitigation-relevant items rather than threat-severity items.

4.2 Disagreement: Where Experts Diverge

4.2.1 Mitigation Strategy Polarization

While experts agree that threats are severe, they polarize sharply on how to respond (Figure 2). Government regulation receives the highest mean effectiveness rating (M=5.24, SD=1.65) and is also the most-frequently chosen *most effective* strategy in the best–worst scaling task (30%), yet 15% of respondents simultaneously rate it the *least* effective. Media literacy is even more polarized: 26% rate it most effective and 24% rate it least effective, and it is the only strategy in our set whose bimodality coefficient (BC=0.559) crosses the conventional threshold of 5/9, indicating a genuinely two-peaked distribution rather than broad uncertainty. Technical detection produces the opposite pattern: it has the lowest mean (M=4.57) but also one of the lower BC values, meaning experts agree it is limited rather than disagreeing about its effectiveness. Watermarking and platform

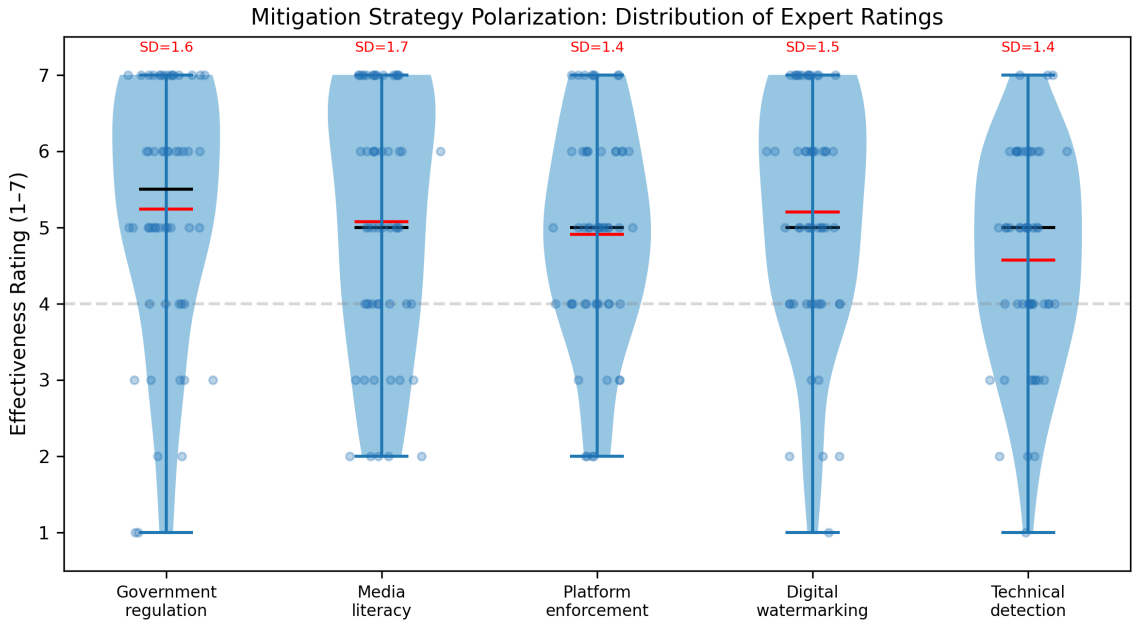


Figure 2: Distribution of expert effectiveness ratings for five mitigation strategies. Violin plots reveal polarization: government regulation and media literacy show bimodal distributions, while technical detection shows tight clustering at lower ratings. Red dots = means, black lines = medians, individual responses shown as jittered points.

enforcement sit between these poles, with means of 5.20 and 4.91 respectively. Figure 3 positions all five strategies on a polarization quadrant: mean effectiveness on the x-axis, BC on the y-axis, marker size and color encoding the coefficient of variation. The quadrant view shows that no strategy yet occupies the "consensus good" position (high effectiveness, low polarization), which is the central practical finding: the field has agreement that something must be done, but no consensus on what.

4.2.2 Group Differences

Table 3: Kruskal–Wallis group differences (p<0.10). 52 tests in total (26 DVs × 2 grouping factors); no multiple-comparison correction (exploratory). Effect size $\varepsilon^2 = H/(N-1)$; Cohen-style conventions: .01 small, .06 medium, .14 large. Group means in text.

Factor	DV	H	df	N	ε^2	p
Region	Threat: Audio (financial)	5.10	1	54	.096	.024 *
Role	Threat: Text (general)	7.74	3	52	.152	.052 .
Region	Mitigation: Tech detection	3.53	1	54	.067	.060 .

* $p<.05$; . $p<.10$.

Comparing AI researchers, journalists, and policymakers reveals that professional background shapes both threat perception and mitigation preference. AI researchers and other AI professionals show similar threat profiles overall, but diverge sharply from media-affiliated respondents on text-based threats, where the latter assign markedly higher severity (Figure 4). Non-parametric Kruskal–Wallis tests across the four role groups and the EU/non-EU split surface three significant

or marginal group differences (Table 3): disinformation/ethics respondents rate general text threats highest (M=6.20) and journalism respondents lowest (M=4.67); non-EU respondents rate financial-domain voice cloning more threatening than EU respondents (M=6.13 vs. 5.00) and judge technical detection as a marginally more effective mitigation (M=4.87 vs. 4.17). All other DV × factor combinations produced $p>0.10$; we report the three signals without multiple-comparison correction and treat them as exploratory. The associated effect sizes (ε^2) place the role × text-threat split in the *large* range and the two region splits in the *medium* range, indicating that the disagreement is substantively meaningful rather than a knife-edge p-value artifact. Mitigation preferences track these professional lenses: AI researchers cluster around technical detection and provenance standards, journalism and disinformation professionals lean toward literacy interventions, and policymakers prioritize binding regulation (Figure 5). The journalist and policymaker subgroups are small (each $N\approx2$–4) and group-level statistical tests for those cells are reported as exploratory; nevertheless, the pattern is internally consistent across the rating scale, the ranking task, and the open-ended responses, suggesting it reflects a stable epistemic-community effect rather than sampling noise.

4.2.3 Domain-Specific Disagreements

The threat hierarchy is not constant across domains. In health, the political-domain ordering inverts: text-based threats lead (M=5.80) while video drops to M=5.13, consistent with concerns that LLM-generated medical claims can spread faster than fact-checkers can respond. In the financial domain, audio becomes the dominant modality (M=5.65), driven by voice-cloning fraud scenarios that figure prominently in respondents' open-ended comments. The social domain returns to a video-led profile (M=6.00), with harassment and non-consensual deepfake imagery cited by 46% of respondents as an urgent risk. These domain-specific reorderings reinforce the finding that "video deepfakes are the biggest threat" is a political-domain effect, not a universal one.

4.3 Qualitative Findings: Emerging Risks and Expert Reasoning

Thematic coding of the open-ended responses surfaced five emerging-risk themes that are largely absent from current policy frameworks (Figure 6). *Weaponized intimacy*: AI companions used as personalized disinformation vectors, capable of building affective trust before introducing manipulated content. Arne Stenmanns frames this directly: "While we currently focus on deepfakes (fake content), the next frontier is deep agents (fake relationships)" [R16]. *Synthetic consensus and covert influence*: large-scale poisoning of widely used LLMs so that downstream answers reflect attacker-aligned narratives. Gerhard Schimpf describes the mechanism as "flooding the systems with Russian friendly articles that creep into the models. Questions asked may result in synthetic consensus" [R04]. *Chatbots as long-lived influence channels*: as conversational agents become embedded in daily routines, Katerina Sedova warns that "it will be critical to ensure that these mediums are not manipulatable by threat actors to influence political, economic, medical, and social decision making" [R45]. *Information ecosystem degradation*: Peter Carragher names the dynamic as "information degradation via tragedy of the commons effects" [R43], with current platform incentives rewarding low-quality content production; one respondent simply asks "Will be in confusion what is truth?" [R44]. *Research-integrity and demographic-specific harms*: AI-generated fake studies and citation graphs that contaminate the scholarly record ("AI disinformation in research" [R*23*, Siegel]), and targeted exploitation of children and other users with limited critical-media skills (Alberto Lobato Diogo, R38). These themes converge on a single structural observation: the emerging risks treat AI not as a content generator but as an actor inside the information environment, a framing that current regulation has only begun to address.

The LLM analysis returned, for each open-ended question and each grouping dimension, a list of agreements, named disagreement topics with per-group positions, and representative quotes keyed to source respondents. Figure 7 shows a representative excerpt.

Three agreements hold consistently across all dimensions: AI-generated disinformation is treated as an immediate threat rather than a future one; scale and accessibility are seen as its core danger; and all groups agree that no single intervention is sufficient. Below, we report four recurring *disagreement patterns* that appear across multiple dimensions and questions.

Finding 1: Which content modality is most dangerous. The sharpest and most persistent disagreement concerns whether text/audio or audiovisual content poses the greater risk. Technical respondents and North American respondents converge on text and audio cloning as the more dangerous vectors: personalized one-to-one persuasion operates in private communication channels that resist public debunking, whereas deepfake video, once detected, can be corrected publicly. Academic/Research respondents and medium-understanding respondents (LLM self-rating 4–5) take the opposing position, arguing that audiovisual threats have intensified since 2019 and that even visibly AI-generated content is frequently misread as factual. Policy/Regulatory respondents and Asia-Pacific respondents focus on a third scenario—*not* the long-run epistemic question but concrete near-term misuse: last-minute deepfake drops before elections and voice-cloning used for payment fraud. The same three-way split (text > video; video > text; both matter in specific attack scenarios) appears in the political domain, financial domain, and why-risks-worry-you questions across all three dimensions, suggesting it reflects a genuine divergence in threat models rather than question-specific framing effects.

Finding 2: What kind of harm AI disinformation is. Groups disagree not about severity but about the *unit* of harm. One cluster—technical profession, the Other group, and high-understanding respondents—frames the core harm as structural-epistemic: the erosion of shared reality, *"Reality Apathy"* [R14], and the collapse of truth infrastructure that democracy and science depend on. A second cluster (Asia-Pacific and low-understanding respondents) frames it as direct individual victimization: severe personal harm cases including suicide threats linked to deepfake video and life savings lost to voice scams. Policy/Regulatory respondents and EU/EEA respondents occupy a middle position, treating both systemic democratic harm and direct fraud as central, without strongly prioritizing one. This split has methodological implications: the two clusters emphasize different survey constructs (epistemic fragmentation vs. measurable individual harm), which affects how impact metrics should be designed.

Finding 3: Three competing mitigation philosophies. On mitigation, the analysis identifies three distinct primary orientations that do not overlap neatly with any single grouping dimension. *Technical provenance* is the dominant preference among Technical profession respondents and North American respondents: browser- and OS-level verification, C2PA content credentials, and AI-assisted detection, often combined with skepticism that regulation does more than entrench incumbents. *Binding regulation* is the dominant preference among Policy/Regulatory respondents and EU/EEA respondents, who treat mandatory transparency requirements and age-based platform restrictions as the foundational fix. *Literacy education* is the dominant preference among Academic/Research respondents, Asia-Pacific respondents, and medium-understanding respondents, who see literacy as both the most democratically safe and the most scalable intervention. Journalism/Disinfo respondents occupy a structurally distinct fourth position: they are skeptical of all three orientations in isolation, and are the only group to cite empirical evidence that literacy

interventions can increase generalized distrust while reducing real-news detection—a backfire effect that neither the technical nor the regulatory cluster addresses.

Finding 4: Abstraction level tracks technical self-assessment. The LLM-understanding dimension reveals a systematic gradient in how experts frame both threats and solutions. High-understanding respondents produce the most analytically developed accounts: they frame disinformation as a systems-design and incentive problem, raise second-order risks (e.g., that AI literacy tools may be architecturally unable to achieve full-proof mitigation, or that over-correction toward centralized AI control may itself be dangerous), and show the most internal disagreement on regulation. Low-understanding respondents characterize the same threats through concrete operational harm cases without theorizing mechanisms or examining intervention trade-offs. Medium-understanding respondents are intermediate and lean toward literacy optimism. This is consistent with the profession and region patterns: Technical and Journalism/Disinfo respondents cluster at the high end, Asia-Pacific and Other respondents are more distributed. The implication for interpreting the survey is that group differences in risk framing may partly reflect differences in the conceptual frameworks experts bring rather than differences in underlying concern levels.

Group profiles. Across all three dimensions, three candidate epistemic orientations appear. *Infrastructure-oriented* respondents (Technical profession; North America; high LLM understanding) frame disinformation as a systems-design problem requiring automated, scalable defenses and treat governance as secondary. *Governance-oriented* respondents (Policy/Regulatory; EU/EEA; medium understanding) treat it as a regulatory compliance problem and favor mandatory rules, with literacy as a complementary layer. *Harm-proximity* respondents (Other; Asia-Pacific; low understanding) center concrete individual harm cases and favor practical resilience measures without strong preferences on systemic architecture. Journalism/Disinfo respondents cut across this typology: they share the harm-proximity group's skepticism of techno-optimism and the governance group's emphasis on platform accountability, but uniquely combine these with the strongest insistence on measurement validity and empirically validated intervention design.

5 Discussion

5.1 The Mitigation Paradox and What It Means for Policy

Disagreement among experts has direct consequences for policy design. The asymmetry we observe—broad convergence on threats, sharp polarization on mitigation—defines what we term the *Mitigation Paradox*: a setting in which a policy community agrees a problem must be addressed urgently, yet provides no coherent signal about which lever to pull. If experts cannot agree on the relative effectiveness of mitigation strategies, policymakers are unlikely to find a single dominant solution. Instead, the evidence supports layered or portfolio approaches that combine regulation, technical detection, and public resilience initiatives. This mirrors early climate-policy debates, where consensus on severity emerged long before agreement on the most effective interventions. The present results suggest that policy should prioritize flexibility and staged evaluation rather than irreversible single-strategy commitments. Table 4 translates each disagreement signature into a recommended policy action.

Table 4: From disagreement signature to policy action. BC = bimodality coefficient (5/9 threshold); CV = coefficient of variation.

Disagreement signature	Recommended action
High mean, low BC, low CV (e.g., severity of video deepfakes)	**Bind:** ready for regulation, evaluation, and resourcing
High mean, BC above 5/9 (e.g., media literacy effectiveness)	**Pilot:** RCT-style deployment with sunset clauses; do not legislate as primary lever
Mixed mean, BC near 5/9 with cross-jurisdictional split (e.g., government regulation, technical detection)	**Differentiate:** allow jurisdiction-specific instruments; require multi-stakeholder governance
Low-to-medium mean, high CV (e.g., emerging-risk categories)	**Monitor:** fund horizon-scanning and re-elicitation, not codification

5.2 Professional Lenses Shape Threat Perception

Preliminary group comparisons suggest that professional background may shape which threats are considered most salient and which mitigation strategies appear credible. Because several subgroups are small and the comparisons are exploratory, we read these as tentative signals rather than established effects. To the extent the pattern holds, it implies that multi-stakeholder governance is not only a normative requirement but also plausibly an epistemic one: no single professional community appears to see the full risk landscape. On this reading, disagreement reflects the partial viewpoints of different expert communities at least as much as simple noise.

5.3 Emerging Risks Require Proactive Governance

Experts consistently flagged emerging risks that are not yet addressed by current regulation, including AI companions used as disinformation vectors, research integrity threats from synthetic studies, and broader information ecosystem collapse. These risks extend beyond present policy categories and suggest the need for dedicated horizon-scanning mechanisms. A proactive governance agenda should therefore include periodic expert elicitation, scenario analysis, and rapid update pathways for regulatory guidance.

5.4 Limitations of Expert Elicitation

Experts are not oracles. Their judgments may reflect anchoring, availability effects, or professional confirmation bias. The observed polarization could therefore reflect genuine uncertainty, systematic group bias, or both. We treat expert disagreement as a signal rather than a defect, but note that future work should triangulate expert perceptions with empirical evidence and behavioral outcomes.

5.5 Future Work

Two extensions follow directly from the present findings. First, a dedicated decision-maker stratum (e.g., executive practitioners versus formal policymakers) should be sampled with sufficient cell sizes to test whether perceived regulatory versus technical solutions diverge predictably by professional position; the present sample is too thin in these specific subgroups to draw conclusions.

Second, longitudinal re-elicitation post EU AI Act implementation, combined with replications in the United States and Asia, would let the disagreement map evolve into a moving picture rather than a snapshot. We also recommend triangulating expert perception data with platform-level measurement studies, so that the questions experts disagree on can be progressively resolved with behavioral evidence.

6 Conclusion

Expert consensus on AI-generated disinformation collapses where it matters most for policy: at the choice of mitigation. We call this asymmetry the *Mitigation Paradox*. Severity is broadly agreed (every modality–domain pair above 5/7; 78% select election deepfakes as the top urgent risk), but the same panel splits on every intervention except technical detection, which is consensually viewed as limited. The polarization is not noise: media literacy and government regulation receive both the highest "most effective" endorsements and substantial "least effective" votes from the same respondents. The disagreement also has a geographic signature: non-EU experts rate voice-cloning fraud as significantly more threatening and technical detection as more promising than their EU peers, suggesting that the same governance instruments may be received very differently across jurisdictions. We further observe that three candidate epistemic orientations—infrastructure, governance, and harm-proximity—appear to account for much of this split, with journalism and disinformation experts uniquely flagging backfire risks that the other clusters do not address; given the sample size these orientations are offered as an exploratory lens rather than a validated typology. The practical consequence of the Mitigation Paradox is that AI-disinformation policy should be designed as a portfolio with iterative evaluation, not as a search for a single dominant fix; and that expert elicitation in this field should be repeated routinely, with explicit attention to the question of who is disagreeing with whom.

Data Availability

The anonymized survey instrument, aggregated results, and the analysis scripts used for the descriptive statistics and disagreement metrics are available in a public repository (https://github.com/aloth/verification-crisis) and archived on Zenodo (https://doi.org/10.5281/zenodo.18703600). Qualitative responses are shared in coded form where consent permits; raw personal data are not released, consistent with the consent terms and GDPR obligations described in the Ethics statement and supplement.

7 Acknowledgments

We thank the survey respondents for their time and the pilot reviewers for feedback on the instrument. Among those who consented to be named, we are grateful to (in alphabetical order): Peter Carragher (Carnegie Mellon University), Doug Finke, Emilio Ferrara (University of Southern California), Eelco Herder (Utrecht University), Jeff Jarvis (CUNY), William Kempster (ar.io network), Volker Klaeren, Stephan Lewandowsky (University of Bristol), Timothy R. Levine (University of Oklahoma), Alberto Lobato Diogo, Manju Rose Mathews (Christ Nagar College Maranalloor), Filippo Menczer (Observatory on Social Media, Indiana University), Merzmensch, Scott Perry (Digital Governance Institute), Gerhard Schimpf (ACM Europe TPC), Katerina Sedova (Atlantic Council), Melanie Siegel (Darmstadt University of Applied Sciences), Gürkan Solmaz (ACM Europe TPC), Giovanni Spitale (Institute of Biomedical Ethics and History of Medicine), Arne Stenmanns, Sam

Stockwell (The Alan Turing Institute), Jesper Strömbäck (University of Gothenburg), Evan M. Williams (Carnegie Mellon University), Roberto V. Zicari (Z-Inspection® Initiative), Muhammad Zubair (Fortex Solutions), and Jason Potel (Goldsmiths, University of London).

8 Ethics, Positionality, and Adverse-Impact Statement

Research ethics and consent. Participation was voluntary, uncompensated, and preceded by an informed-consent screen that disclosed the study purpose, data handling, expected duration, and the option to withdraw at any point. Respondents chose one of three identifiability levels for the final publication: (i) fully anonymous (no name, no quotes; n=27), (ii) name and affiliation in the acknowledgments only (no attributed quotes; n=14), or (iii) name, affiliation, and attribution of specific quotes or ideas (n=13). We treat these tiers as binding: respondents in tiers (i) and (ii) are never quoted with attribution, even where their open-ended responses would have made compelling material. Quotes in the main text are drawn exclusively from tier-(iii) respondents and are identified by a stable respondent ID (Rnn). The instrument and procedure were reviewed against the data-minimization and lawful-basis requirements of the GDPR (consent under Art. 6(1)(a) GDPR); raw responses are held by the lead author's institution and will not be released in identifiable form. One co-author participated in the pilot phase and is excluded from the quotable pool to avoid self-citation.

Positionality. The four authors are based in Germany (2), Austria (1), and France (1), all within the EU regulatory perimeter. Our institutional homes span a public university (Frankfurt UAS), a national applied-research institute (AIT), and a Grande École research chair on cyber critical-infrastructure (IMT Atlantique). Disciplinary backgrounds include AI/ML, cybersecurity, and information-systems governance; none of the authors holds a primary appointment in journalism, fact-checking, or platform-trust-and-safety practice. These positionings plausibly bias the framing toward infrastructure- and governance-oriented epistemic communities (two of the three orientations we identify), and away from harm-proximity perspectives. We have tried to compensate by oversampling journalism, fact-checking, and policy respondents in recruitment, but the EU/non-EU and AI-researcher skew of the final sample reflects, in part, our own networks. Readers from non-Western contexts in particular should treat the absolute disagreement levels we report as a lower bound; the true global heterogeneity of expert opinion is almost certainly larger than what 54 mostly Western respondents can capture.

Adverse impact and dual use. A disagreement map of expert opinion can in principle be misused. Three risks deserve explicit naming. *Divide-and-conquer framing*: state or commercial actors could cite intra-expert polarization as a delay tactic against AI-disinformation regulation, in the same rhetorical shape that fossil-fuel actors used residual scientific uncertainty against climate policy. *Targeted persuasion against epistemic minorities*: knowing which professional communities under-rate which threats could be exploited to design influence campaigns that preferentially attack those blind spots. *Re-identification risk on small subgroups*: with only 2–4 policymakers and 3 journalists in the sample, group-level rating disclosures are close to single-respondent disclosures. We mitigate these risks by (a) reporting subgroup statistics only where $n \geq 3$ and never publishing per-cell raw rating tables, (b) framing the polarization findings as an argument *for* portfolio governance and earlier evaluation rather than against action, and (c) recommending in Section *Discussion* that follow-up elicitations release disagreement signatures at aggregate, not per-respondent, granularity. We see no realistic path by which the analyses or scripts released alongside this paper could

materially improve the operational capacity of a disinformation adversary; the upside for governance researchers, policymakers, and the expert communities themselves substantially outweighs the residual risk.

References

Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4), 2023. doi: 10.37016/mr-2020-119.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226*, 2023.

Wilson Ceron, Mathias-Felipe de Lima-Santos, and Marcos G. Quiles. Fake news agenda in the era of covid-19: Identifying trends through fact-checking content. *Online Social Networks and Media*, 21:100116, 2021. ISSN 2468-6964. doi: https://doi.org/10.1016/j.osnem.2020.100116. URL https://www.sciencedirect.com/science/article/pii/S2468696420300562.

Heather Chen. Finance worker pays out $25 million after video call with deepfake "chief financial officer". CNN, 2024. URL https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

Zhiyi Chen, Jinyi Ye, Beverlyn Tsai, Emilio Ferrara, and Luca Luceri. Synthetic politics: Prevalence, spreaders, and emotional reception of ai-generated political images on x, 2025. URL https://arxiv.org/abs/2502.11248.

Robert Chesney and Danielle Keats Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6):1753–1820, 2019.

Giulio Corsi, Bill Marino, and Willow Wong. The spread of synthetic media on X. *Harvard Kennedy School Misinformation Review*, 5(3), 2024. doi: 10.37016/mr-2020-140.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11: 1166120, 2023.

Chiara Drolsbach and Nicolas Pröllochs. Characterizing ai-generated misinformation on social media, 2025. URL https://arxiv.org/abs/2505.10266.

European Digital Media Observatory. Prebunking ai-generated disinformation ahead of eu elections, 2024. URL https://edmo.eu/publications/prebunking-ai-generated-disinformation-ahead-of-eu-elections/.

European Parliament. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI act). Official Journal of the European Union, 2024.

Sean Guo, Yiwen Zhong, and Xiaoqing Hu. People are more susceptible to misinformation with realistic AI-synthesized images that provide strong evidence to headlines. *Harvard Kennedy School Misinformation Review*, 6(6), 2025. doi: 10.37016/mr-2020-189.

Adnan Hoq, Matthew Facciani, and Tim Weninger. Feedback and education improve human detection of image manipulation on social media. *Harvard Kennedy School Misinformation Review*, 6(2), 2025. doi: 10.37016/mr-2020-175.

Aaron Hyzen. Revisiting the theoretical foundations of propaganda. *International Journal of Communication*, 15(0), 2021. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/15769.

Sayeed Ahsan Khan, Mohammed Hazim Alkawaz, and Hewa Majeed Zangana. The use and abuse of social media for spreading fake news. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 145–148, 2019.

Rafał Kozik, Sebastian Kula, Michał Choraś, and Michał Woźniak. Technical solution to counter potential crime: Text analysis to detect fake news and disinformation. *Journal of Computational Science*, 60:101576, 2022. ISSN 1877-7503. doi: https://doi.org/10.1016/j.jocs.2022.101576. URL https://www.sciencedirect.com/science/article/pii/S187775032200014X.

Ariel Kruger, Morgan Saletta, Atif Ahmad, and Piers Howe. Structured expert elicitation on disinformation, misinformation, and malign influence: Barriers, strategies, and opportunities. *Harvard Kennedy School Misinformation Review*, 5(6), 2024. doi: 10.37016/mr-2020-169.

Gillian Murphy, Constance de Saint Laurent, Megan Reynolds, Omar Aftab, Karen Hegarty, Yuning Sun, and Ciara M. Greene. What do we study when we study misinformation? a scoping review of experimental research (2016–2022). *Harvard Kennedy School Misinformation Review*, 4(6), 2023. doi: 10.37016/mr-2020-130.

Lynnette Hui Xian Ng, Ian Kloo, Samantha Clark, and Kathleen M. Carley. An exploratory analysis of covid bot vs human disinformation dissemination stemming from the disinformation dozen on telegram. *Journal of Computational Social Science*, 7(1):695–720, 2024. ISSN 2432-2725. doi: 10.1007/s42001-024-00253-y. URL https://doi.org/10.1007/s42001-024-00253-y.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Md. Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 10:25494–25513, 2022.

Elisabeth Steffen. More than memes: A multimodal topic modeling approach to conspiracy theories on telegram. *Proceedings of the International AAAI Conference on Web and Social Media*, 19: 1831–1844, June 2025. ISSN 2162-3449. doi: 10.1609/icwsm.v19i1.35904. URL http://dx.doi.org/10.1609/icwsm.v19i1.35904.

Sander van der Linden. *Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity*. W. W. Norton, 2023.

Teresa Weikmann, Ferre Wouters, Marina Tulin, Michael Hameleers, Claes H. de Vreese, Brahim Zarouali, and Michaël Opgenhaffen. On the same page? experts are mostly, but not always aligned about disinformation in times of generative AI. *Harvard Kennedy School Misinformation Review*, 7(2), 2026. doi: 10.37016/mr-2020-196.

Harry Yaojun Yan, Garrett Morrow, Kai-Cheng Yang, and John Wihbey. The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election. *Harvard Kennedy School Misinformation Review*, 6(1), 2025. doi: 10.37016/mr-2020-171.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 0(ja), 2018. ISSN 0360-0300. doi: 10.1145/3395046. URL https://doi.org/10.1145/3395046.

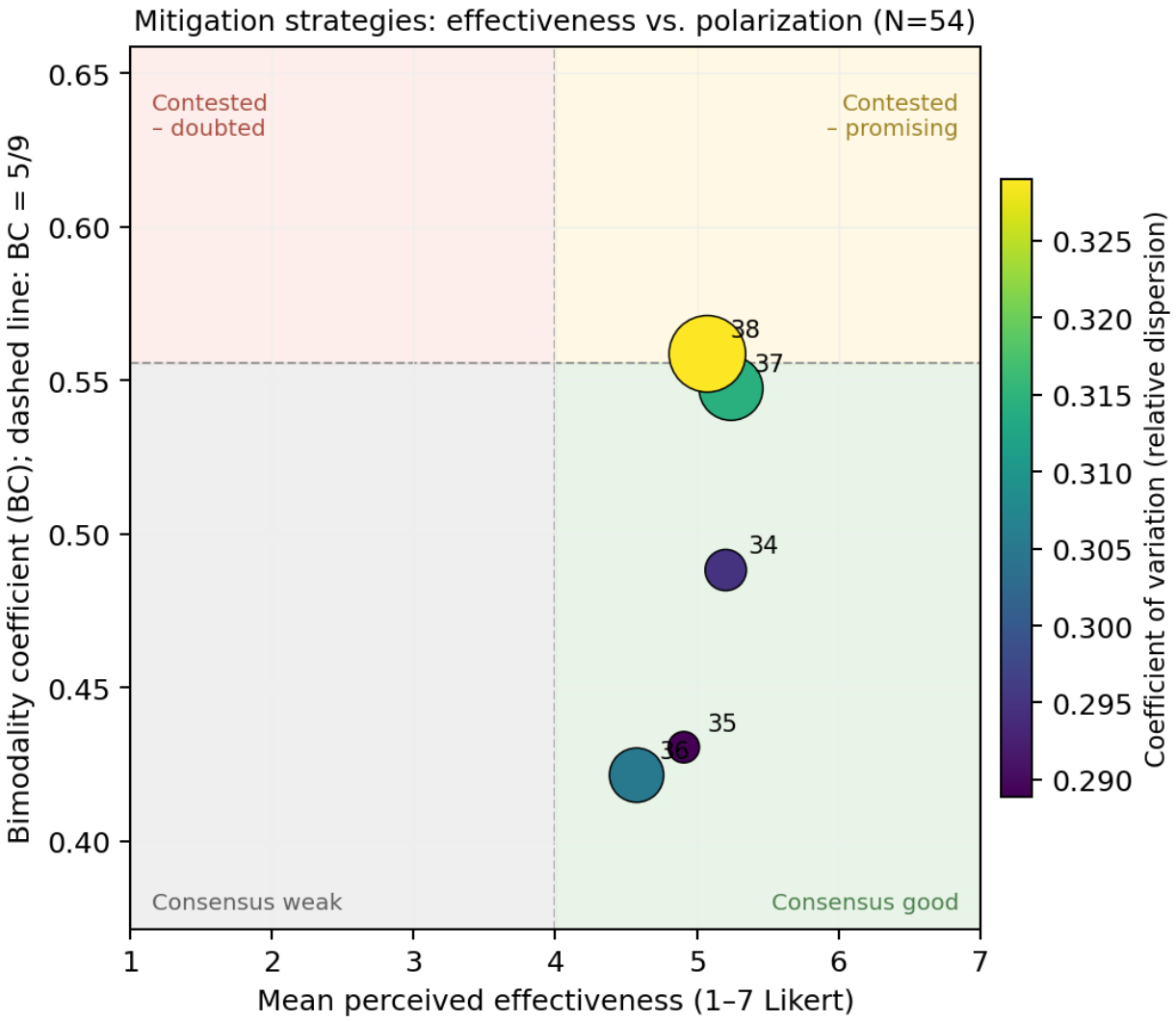


Figure 3: Mitigation strategies positioned by mean effectiveness (x) and bimodality coefficient (y). The dashed horizontal line marks the conventional bimodality threshold (BC=5/9). Marker color/size encodes relative dispersion (CV). Media literacy is the only strategy above the bimodality threshold; government regulation sits at the boundary.

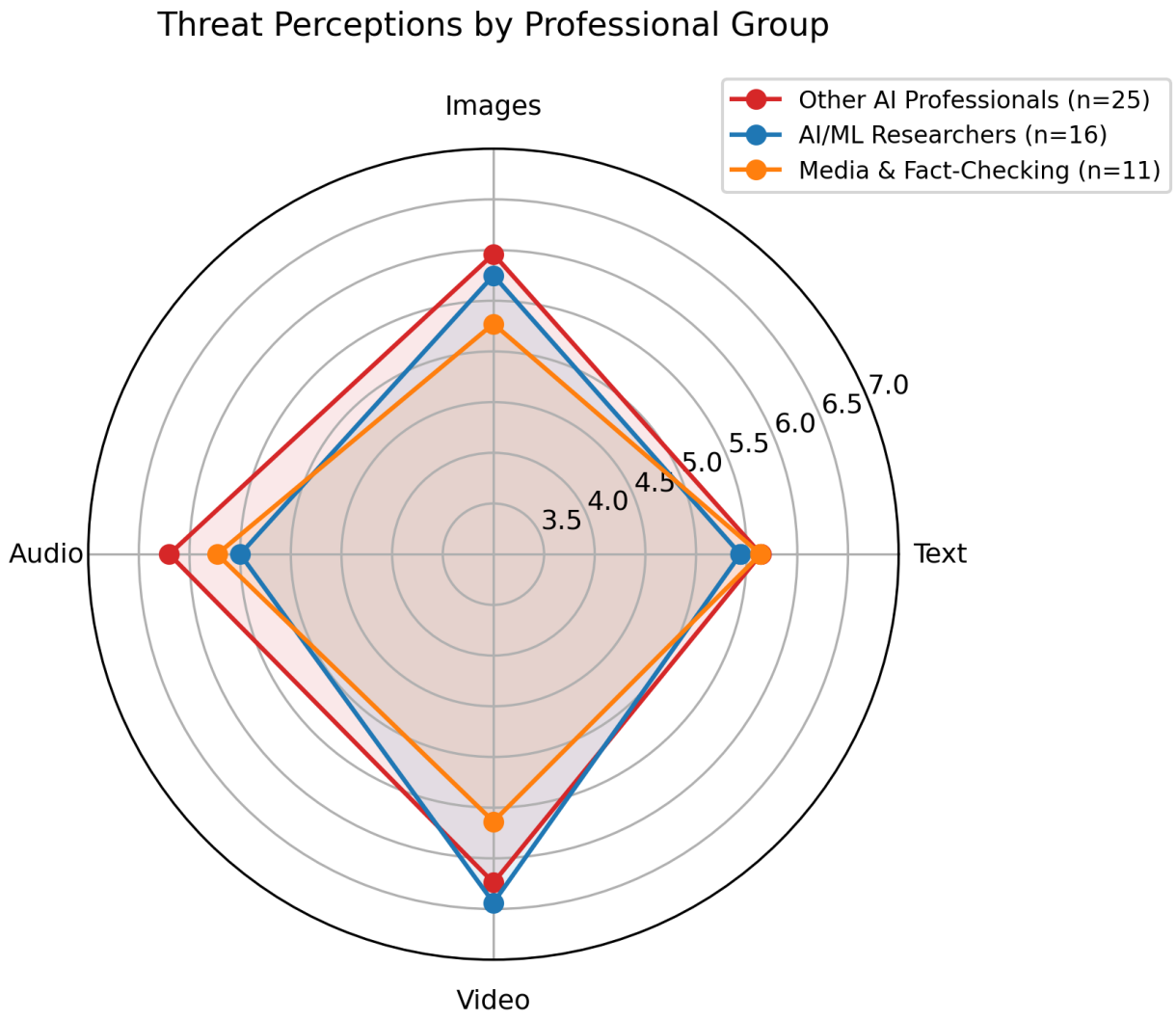


Figure 4: Threat perception profiles by professional group. AI/ML researchers and other AI professionals show similar patterns, while media professionals diverge on text-based threats.

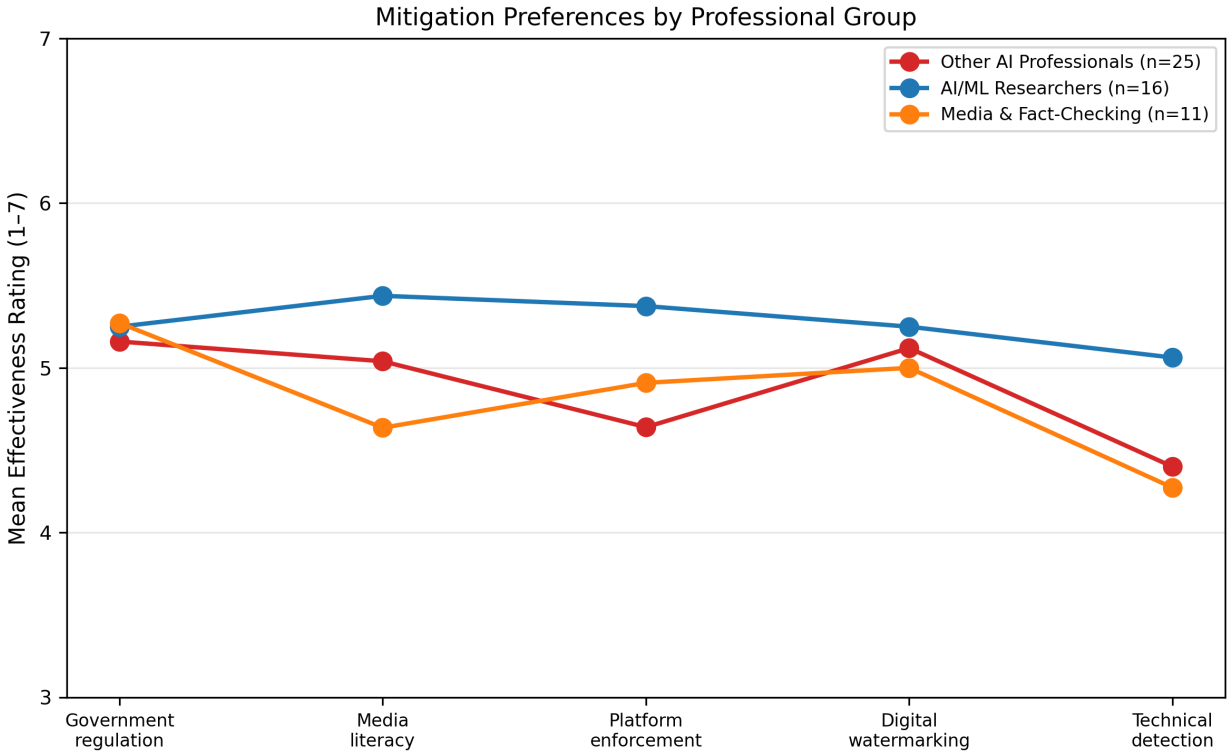


Figure 5: Mitigation strategy preferences by professional group. Different expertise backgrounds lead to divergent strategy preferences, with AI researchers favoring technical detection and media professionals prioritizing literacy.

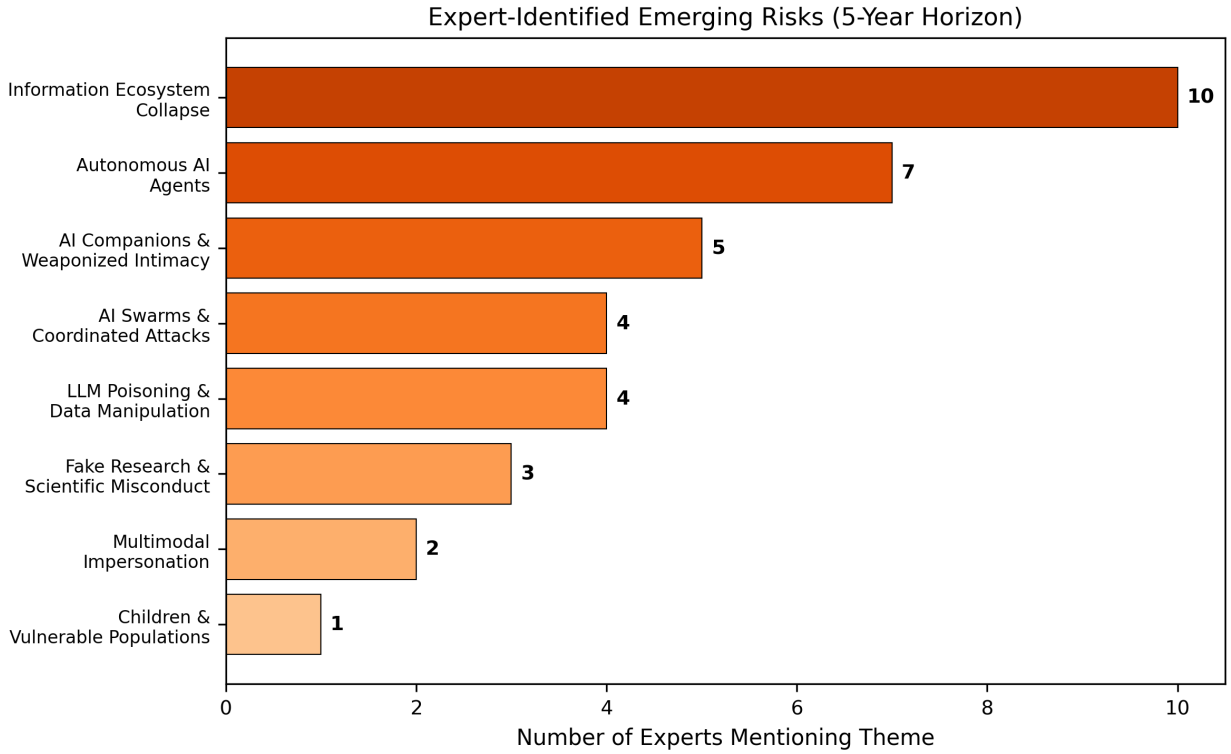


Figure 6: Emerging risks identified by experts for the next five years, coded from open-ended responses. Themes extend beyond current policy frameworks.

[**Political domain**]

Agreements (4):

- AI is seen as a threat to political processes through misinformation, disinformation, and manipulation.
- Groups broadly agree that elections are a key vulnerability.

Disagreements (3):
How dangerous are deepfakes vs. text/audio persuasion

Technical: Argues personalized text persuasion may be a bigger threat than fake video because video can be debunked.
`[R10]` "Personalized 1:1 persuasion via text is a bigger threat to democracy than a fake video which can be debunked."

Policy/Reg: Highlights "deepfake dump" tactics right before voting days. `[R13]`

Academic: Notes audiovisual threats are more serious than textual threats. `[R15]`

Figure 7: Excerpt from the LLM JSON outputs for the *Political domain* question, profession dimension. `[R`*n*`]` identifiers are 1-based CSV row numbers.