Proceedings of the 5th International Conference on International Law and Legal Policy
DOI: 10.54254/2753-7048/2026.39335

Calibrating Trust: Confidence–Accuracy Dissociation in Multimodal Deepfake Detection

Xiaorui Xu

Basis International School Park Lane Harbor, Huizhou, China
xiaorui.xu14458-biph@basischina.com

Abstract. Generative systems allow the same claim to be expressed in text, in photorealistic images, or in deepfake video. The implicit assumption is that the richer media will be more difficult to judge and thus more dangerous. This review brings together the experimental data published about human judgment of authenticity across media modalities from a meta-analysis of 56 papers (86,155 participants), five pre-registered experiments (2,215 participants), and four additional experimental studies. Pooled detection accuracy is close to chance across media modalities, and the confidence intervals of text, image, audio, and video overlap. To the extent that there is an effect of media modality on performance, richer media help with discernment instead of hampering it. Across all media formats, what remains constant is the decoupling of confidence and accuracy: people are highly confident in their judgments irrespective of correctness, biased toward acceptance of synthetic content, and unlikely to question authenticity unless prompted to do so. It appears that the danger of synthetic media is metacognitive and not perceptual, and that interfaces for verification should be assessed on confidence calibration in addition to detection accuracy.

Keywords: Deepfakes, Media modality, Authenticity judgment, Confidence calibration, Media literacy

1. Introduction

A falsehood can be communicated via one of three media modalities: a textual post, a photorealistic image generated by a generative model, or a video featuring a synthesized speaker delivering the message. Generative systems have brought down the marginal cost of rich media to zero, and their output can no longer be distinguished from authentic media through inspection. Synthesized faces are considered just as real and even more trustworthy than photographs [1].

The widely held narrative is that video is the most dangerous media modality because it is the hardest to recognize. That assumption leads people in the wrong direction. According to a meta-analysis of 56 papers including 86,155 participants, human deepfake detection stands at 55.5 percent on average, while video (57.3 percent) outperforms still images (53.2 percent) and text (52.0 percent) [2]. In a study using matched political speeches, Groh et al. discovered that audio and video increased recognition accuracy compared to transcripts due to extra delivery cues available to listeners [3]. The persuasive value of political video compared to text is also modest [4]. Detection accuracy does not decrease as media becomes richer.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.39335

What shifts is the confidence with which people trust their judgment. People who lack the ability to discriminate between deepfakes and actual videos are still confident in their ability to do so: modality impacts metacognitive processes, not discrimination [5]. It increases confidence, reduces the threshold of acceptance of information, and accelerates decision-making. Designing interfaces for the purpose of increasing the ability of people to discriminate between authentic and fake content, therefore, is a mistake.

Is modality able to impact the ability of people to discriminate between authentic and generated content, or does it only increase confidence and readiness to accept information?

This review analyzes the empirical findings addressing this question. It brings together a meta-analysis of 56 papers, five pre-registered experiments, and four more experiments. They are compared based on four variables: detection accuracy, response bias, confidence, and training effect. The accuracy of fakes does not decline with modality richness, and in cases where content is controlled, it increases. What is common for all types of content is a dissociation of confidence and accuracy. The recommendation for the designers of authenticity interfaces is to evaluate their ability to calibrate people rather than to detect fakes.

2. Related work

2.1. Modality and the realism heuristic

The truthiness effect occurs when the nonprobative picture boosts the belief in the claim it accompanies regardless of its informational value [6]. The mechanism of the truthiness effect is based on fluency: the picture makes a claim easy to visualize, and this ease is perceived as truth. The same applies to generated pictures: the higher the photorealism of a picture, the more likely people will be deceived by misinformation [7]. Therefore, the realism heuristic implies that the effect of text, images, and video on beliefs will be monotonic, yet it has nothing to say about accuracy. An increase in belief in a statement made in a certain format does not necessarily mean that people's accuracy changes.

2.2. Human detection of synthetic media

The evidence regarding the discrimination of the generated content is overwhelmingly negative and surprisingly homogeneous regardless of format. Untrained participants discriminated synthesized faces with 48.2 percent accuracy, and the feedback was not able to increase it beyond 59.0 percent. A representative study in the United States, Germany, and China found that people were unable to reliably distinguish between authentic and generated audio, images, and text [8]. For videos, 27 to 50 percent of participants across five populations were not able to distinguish authentic climate science videos from deepfakes [9]. This is not different for text: the repetition of a statement increases its believability even if it is highly unlikely [10].

2.3. Literacy, labels and verification support

Interventions against misinformation have been assessed for effectiveness almost exclusively using textual stimuli. Media literacy training lowers susceptibility to disinformation, with an effect size of $d = 0.76$ on discernment in 49 studies, and inoculation enhances knowledge of manipulation tactics [11, 12]. It is unknown whether this carries over to stimuli on which little can be done to the content. People are also prone to rely excessively on automated judgment in a way that looks like a cost-

DOI: 10.54254/2753-7048/2026.39335

benefit decision: excessive reliance on automation decreases when the cost of processing an explanation decreases relative to the cost of accepting the answer [13]. Thus, verification interfaces should be expected to influence users' reasoning rather than judgment accuracy.

2.4. Open problems in cross-modal detection research

This literature suffers from three consistent issues. First, across-modality comparisons are typically confounded by variation in content, as well as format, source, and plausibility of the stimuli. Second, accuracy and believability are measured, but not always in conjunction with confidence and response bias, which makes it difficult to distinguish between a perceptual and a metacognitive deficit. Third, interventions are evaluated almost exclusively on text and images, leaving open whether they are effective on synthetic videos.

3. Evidence basis

3.1. Selection of studies

Studies included in the synthesis satisfy four criteria. First, each provides an empirical assessment of human judgment of authenticity, not the performance of an automated detector. Second, each uses stimuli that are described as high quality by the authors, thus ruling out the possibility that failure to detect forgeries can be attributed to obvious generation artifacts. Third, each reports at least one of four statistics: detection accuracy, sensitivity or response bias, confidence, and the effect of an intervention. Fourth, all the papers are peer-reviewed. The resulting body of evidence includes one meta-analysis, five pre-registered experiments reported in a single paper, and four other experimental studies; it covers text, images, audio, and video stimuli. Table 1 lists the included studies.

Table 1. Studies included in the synthesis

Study	Participants	Focus
Diel et al. [2]	86,155 (56 papers)	Accuracy by modality
Groh et al. [3]	2,215	Modality, suspicion
Nightingale, Farid [1]	757	Images, trust
Köbis et al. [5]	210	Bias, confidence
Doss et al. [9]	Five populations	Repeated exposure
Frank et al. [8]	Three countries	Cross-modal accuracy
Huang et al. [11]	81,155 (49 studies)	Training effect

3.2. Measures compared

There are four common measures used in the above studies. Detection accuracy is the percentage of stimuli correctly identified. It can be compared when the prevalence of forgeries is known. Sensitivity and response bias distinguish the ability to discriminate between fabricated and authentic stimuli from the criterion of acceptance of the stimulus as authentic [14]. Confidence is measured on specific scales in different studies, and hence only the relation to accuracy can be compared. Finally, the effect of an intervention is defined as the difference from the control condition.

DOI: 10.54254/2753-7048/2026.39335

3.3. Limits of comparison

There are three restrictions common to all studies. Different stimuli, rating scales, and base rates of fabrication mean that numbers from different papers are not comparable. Only meta-analysis combines effect sizes numerically; all other comparisons made here are qualitative. No single paper presents measures of accuracy, bias, confidence, and latency for one group of participants judging one kind of content in three modalities; that is exactly what the field lacks.

4. Findings

4.1. Modality-invariant detection accuracy

Pooled across 56 papers and 137 effect sizes, human ability to detect fabrication stands at 55.54 percent with a Confidence Interval (CI) of 48.87 to 62.10, including chance [2]. By modality, it is summarized in Table 2 and Figure 1. Text proves to be the least reliable at 52.00 percent and audio — the most, at 62.08 percent. But all CIs include 50 percent and overlap heavily. Video, traditionally thought of as the most dangerous modality, stands at 57.31 percent, being higher than both text and images. Pooled sensitivity values do not differ significantly from chance either. Whatever advantages richer modalities have over text, they do not affect discrimination negatively.

Table 2. Pooled detection accuracy by modality

Modality	Accuracy (%)	95% CI	k
Text	52.00	[37.42, 65.88]	15
Image	53.16	[42.12, 64.64]	18
Video	57.31	[47.80, 66.57]	26
Audio	62.08	[38.23, 83.18]	8
Text + image + video + audio	55.54	[48.87, 62.10]	67

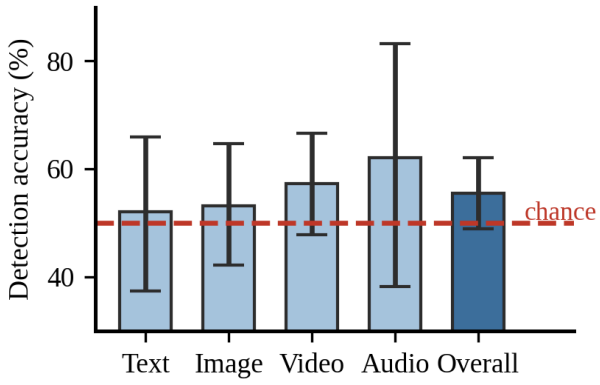


Figure 1. Pooled detection accuracy by modality, with 95% confidence intervals [2]

4.2. Discernment gains from richer modalities

Meta-analysis includes only studies conducted in one modality, so it cannot control for the possibility that identical content might prove to be easier or more difficult to judge in a different

DOI: 10.54254/2753-7048/2026.39335

form. There is an example of experiments checking this possibility. Groh et al. used the same 32 political speeches (half of them were authentic and half – fabricated) as texts, silent videos, audio, and video with audio in five pre-registered experiments involving 2,215 participants [3]. Accuracy increased with each modality added: in the first experiment, transcripts allowed 57 percent correct decisions, silent videos – 64 percent, audio – 81 percent, and video with audio – 86 percent. In another experiment, where participants were not informed about the base rate of fabrication, these figures proved to be lower, but preserved their order: from 58 percent for transcripts to 74 percent for video with audio. Figure 2 demonstrates both.

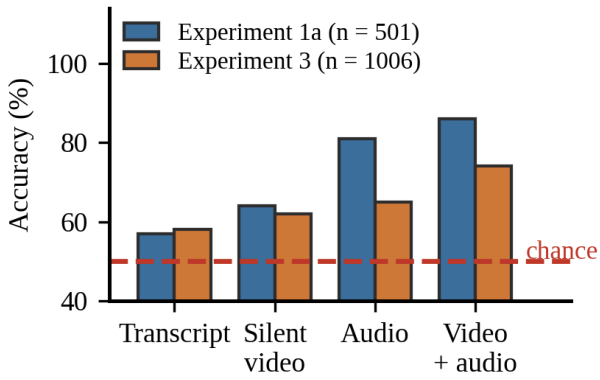


Figure 2. Accuracy by modality in two pre-registered experiments [3]

This turns the logic of much interface design inside out. Video is not the modality in which people are easiest to fool. By the evidence here, it is the modality in which they have the most going for them, because it contains information stripped by the transcript.

4.3. Modality's effect on the acceptance threshold

It is the change in willingness to accept content as authentic that modality affects. Köbis et al. discovered a systematic bias towards misclassifying deepfakes as authentic videos, and neither increased awareness nor monetary incentives improved accuracy [5]. There is a similar asymmetry with modality manipulation. In one experiment, participants were asked to differentiate between a person's authentic voice and that of a voice actor. In 55 percent of audio-only trials, participants thought the stimulus was an actor's voice, but in only 45 percent of trials when the same audio clip was accompanied by video, although the base rate was identical. Authentic-looking video made even audio clips, real and fabricated, look more real. Base rates affected the threshold in the same way: where fabrication was common rather than rare, accuracy on authentic items increased by 7.2 percentage points while accuracy on fabricated items decreased by 5.8 percentage points [3].

This is the point that should change the risk assessment. Deepfakes are not a problem because they are difficult to perceive. They are a problem because they reduce the cost of believing.

4.4. The calibration gap

Confidence does not track any of this. Confidence was very high and unrelated to accuracy in Köbis et al. [5]. Nightingale and Farid found a similar disconnect in their research on images: accuracy was 48.2 percent, statistically equal to chance, and a training session with feedback on each trial did not improve accuracy over time, remaining at 59.3 percent and 58.8 percent in two blocks of 64 faces

DOI: 10.54254/2753-7048/2026.39335

[1]. The same research also showed synthetic faces being perceived as more trustworthy than the real ones. The interface, therefore, can increase accuracy only slightly while leaving confidence completely uninformative, which is precisely the situation in which people follow wrong answers without second thoughts.

In one experiment involving 60 video stimuli, accuracy and confidence correlated positively, meaning that videos for which the participants gave correct assessments were also the videos they were most confident about. This is a correlation between stimuli as opposed to between participants, and it complements the findings described above. It suggests that calibration is dependent on the stimulus to some degree as much as on the individual, giving another reason to consider confidence as an interface variable rather than a fixed property of the user.

4.5. Unprompted suspicion of fabricated content

Each of the above results is based on asking the participants directly whether they thought that something was fabricated. One experiment got rid of this step and showed the participants the stimuli and asked whatever came to mind. Suspicion occurred relatively rarely and even more rarely for texts – 7.3 percent of observations of fabricated transcripts were suspicious, as compared to 4.8 percent for authentic ones. For text-to-speech audio, these numbers are 13.0 and 8.6 percent; for deepfake videos with audio, they were 32.4 percent, which is 6 percentage points higher than for authentic videos. Figure 3 presents the statistics for the fabricated stimuli.

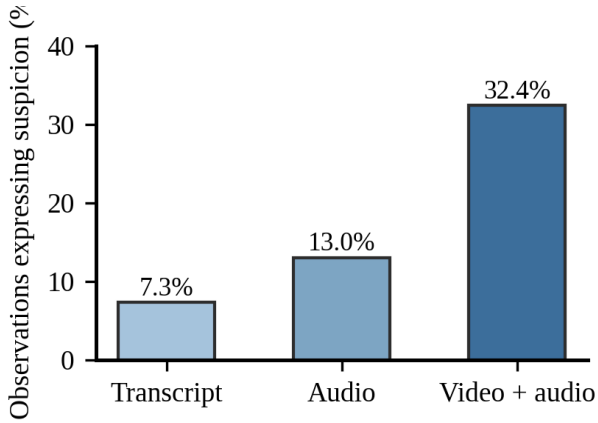


Figure 3. Suspicion of fabrication in absence of any question about authenticity [3]

4.6. Narrow effects of training interventions

Interventions are effective to some extent. Media literacy training increases misinformation discernment with an effect size of d = 0.76 in 49 experiments involving 81,155 participants; however, the experiments in question use texts and images almost exclusively. Video training is also possible – one experiment showed that media literacy training increased discernment accuracy by 24 percentage points compared to the control group [15]. The strategies included in the meta-analysis, such as feedback training and artificial intelligence (AI) assistance, raise pooled accuracy to 65.14 percent, which is well above chance but far from reliable [2]. On the other hand, repetition without training does the opposite – across educators and students, the third and fourth videos were 13 to 47 percentage points less accurately identified in terms of authenticity [9].

DOI: 10.54254/2753-7048/2026.39335

4.7. Design implications

Three conclusions can be made based on this data. First, verification interfaces need to be evaluated in terms of calibration as opposed to accuracy of detection. If an interface increased accuracy of detection by just a couple of percentage points but raised confidence drastically, it would make its users more dangerous, as their mistakes would be made with a much higher degree of certainty. The relationship between confidence and accuracy needs to be at the heart of the evaluation process, not at the periphery.

The second difference is that the intervention point is the threshold, not the artifact. Overlaying generation artifacts trains users' existing surface-based heuristics in evaluating videos, and both their performance and efficacy decay with each generation of models. Signaling provenance, that is, ensuring that every file contains a verified history of capturing and editing, addresses the bias directly and provides a principle-based way to abstain from accepting such content.

Lastly, content reasoning needs to become cheap in rich formats. Presenting a transcript of the video, showing the claim as text alongside the image, and making it necessary to verify the source in order to unlock a response interface all decrease the cost of the reasoning. This agrees with the finding that the tendency to rely on content is decreased when content reasoning becomes cheaper than accepting the conclusion [13].

5. Conclusion

In the published research, users were unable to distinguish between authentic and fabricated content of any modality, and their ability approached chance when pooled. Moreover, richer modalities did not decrease that ability, and when content was held constant, they even increased it. What did not increase was the calibration of users. They remained highly confident regardless of accuracy; the thresholds for accepting the content as authentic were decreased when the format seemed real, and the authenticity was not questioned at all if not asked explicitly.

The practical implication of that finding is that the problem of deepfakes has been framed as the problem of detection while the evidence shows that it is a problem of calibration. Training users to recognize artifacts buys a short-lived skill that repeated exposure erodes. User interfaces that make it more expensive to have false confidence and cheaper to verify provenance and plausibility should outperform detection aids, and should be evaluated based on that hypothesis.

There are three limitations in this literature review. First of all, the research base is uneven – videos and images have been studied intensively, while synthetic texts – much less. Second, comparative analysis of the results of research using different stimuli and different sample sizes is done qualitatively, not quantitatively. Lastly, the design that would answer the question conclusively by comparing the performance of users holding the content constant across three modalities has not been implemented.

References

[1] Nightingale, S. J., Farid, H., & (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America, 119*(8), e2120481119.

[2] Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, p. 100538.

DOI: 10.54254/2753-7048/2026.39335

[3] Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications, 15*, p. 7629.
[4] Wittenberg, C., Tappin, B. M., Berinsky, A. J., & Rand, D. G. (2021). The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences of the United States of America, 118*(47), e2114388118.
[5] Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience, 24*(11), p. 103364.
[6] Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review, 19*(5), 969-974.
[7] Jang, E., Lee, H. M., Lee, S., Jung, Y., & Sundar, S. S. (2025). Too good to be false: How photorealism promotes susceptibility to misinformation. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems*, Article 531.
[8] Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., & Holz, T. (2024). A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy* (SP), 55-73.
[9] Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports, 13*, 13429.
[10] Udry, J., & Barber, S. J. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology, 56*, 101736.
[11] Huang, G., Jia, W., & Yu, W. (2024). Media literacy interventions improve resilience to misinformation: A meta-analytic investigation of overall effect and moderating factors. *Communication Research*. Advance online publication.
[12] Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34), eabo6254.
[13] Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 129.
[14] Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17*(1), 78–98.
[15] Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).