

DOI: https://doi.org/10.48009/1_iis_2026_120

Accuracy in Identifying Deepfake Videos: Exploring Factors That Enhance and Impair Detection

Alan Franklin, *Eastern Kentucky University*, alan.franklin@eku.edu

Blessing Wallin, *Southern Illinois University*, blessing.smithley@siu.edu

Belle Woodward, *Southern Illinois University*, bellew@siu.edu

Sara Incera, *Eastern Kentucky University*, sara.incera@eku.edu

Abstract

The present study examined how people judge the authenticity of deepfake media. Participants ($N = 171$) completed an online task requiring authenticity judgments of short audiovisual clips (five authentic, five deepfake). Overall accuracy was modest ($M = 5.70/10$), and signal detection analyses indicated weak but above-chance sensitivity ($d' = 0.38$) alongside a conservative bias toward classifying content as real ($c = 0.09$). Multiple regression analyses showed that analytical reasoning and skepticism predicted higher detection accuracy, and confidence was positively associated with performance. Emotional measures and deepfake technology familiarity did not reliably predict accuracy. These findings suggest that unaided user judgment remains limited when evaluating synthetic media and provide baseline behavioral evidence that may inform future platform supports, verification aids, and user-facing interventions.

Keywords: deepfakes; authenticity judgments; signal detection theory; analytical reasoning; skepticism; confidence

Introduction

Advances in artificial intelligence have increased the accessibility and realism of synthetic media, creating new challenges for digital platforms and organizational information environments where users must rapidly evaluate the authenticity of audiovisual content (Salman et al., 2023; Bikku et al., 2023). At the same time, users frequently encounter algorithmically curated content streams without fully understanding how recommendation and ranking systems shape what they see, which can complicate judgments of credibility and trust (Eslami et al., 2015; Kizilcec, 2016). Deepfakes—AI-generated audio and video that can appear convincingly real—have already been used to spread misinformation and damage reputations, making authenticity judgments an increasingly important applied research problem.

Prior research indicates that human deepfake detection performance is generally weak. Accuracy rates are often only slightly above chance, with Köbis et al. (2021) reporting modest sensitivity ($d' = 0.484$) and recent meta-analytic evidence suggesting that pooled sensitivity estimates are often not reliably above chance (Diel et al., 2024). Prior work also points to a conservative truth bias, indicating that individuals tend to assume media content is authentic rather than manipulated (Köbis et al., 2021; Diel et al., 2024).

Authenticity judgments in digital environments are shaped not only by what individuals know or how carefully they think, but also by the information system that delivers the content—its interface

affordances, credibility cues, and the broader platform context that determines what information is visible, verifiable, and socially endorsed. In IS research, trust can be conceptualized as trust in the technology artifact itself, not only trust in people or institutions, which is especially relevant when users rely on platform-mediated information under uncertainty (McKnight et al., 2011).

Online credibility judgments are often guided by cognitive heuristics such as reputation, endorsement, consistency, and expectancy-based cues rather than exhaustive verification (Metzger & Flanagin, 2013). These judgments also occur within a broader platform-governance context in which labeling, moderation, and transparency practices shape how authenticity is interpreted (Gorwa, 2019). Consequently, interface-level transparency and explanation design become important considerations, as they influence how users interpret, trust, and respond to system outputs (Kizilcec, 2016). Prior research indicates that individuals frequently rely on source and provenance cues when assessing potentially manipulated media, making source-linked indicators and explanatory signals particularly relevant interface design considerations (Sherman et al., 2021). Accordingly, platforms may benefit from providing clearer provenance-style indicators that prompt users to question potentially deceptive media. It is important that these cues are designed carefully because users may overgeneralize or misinterpret what these indicators actually guarantee (Sherman et al., 2021; Feng et al., 2023).

From an applied IS perspective, this study contributes baseline behavioral evidence about which user-level factors are associated with more accurate authenticity judgments in AI-mediated environments. It also examines whether a familiarity-based measure of deepfake-related technological knowledge was associated with performance in practice. More broadly, the study is intended as a behavioral foundation for later work on user-facing interventions, platform supports, and other system-level mechanisms that may shape authenticity judgments under uncertainty. This line of work has implications for future user-facing interventions, digital literacy strategies, and platform-level supports for evaluating synthetic media. For example, it may motivate later work on authenticity supports such as provenance indicators, verification cues, and user-facing decision aids.

Our pilot study (Alsobeh et al., 2024) provided exploratory groundwork for this preregistered follow-up study. The pilot suggested that analytical reasoning, trust/skepticism, and affective factors might be associated with deepfake detection. Building on those exploratory patterns, the present study was preregistered with the Open Science Framework (OSF) to distinguish confirmatory tests from exploratory analyses and to evaluate five specific hypotheses:

- H1: Higher analytical reasoning will be associated with greater deepfake detection accuracy.
- H2: Lower emotional reactivity will be associated with greater deepfake detection accuracy.
- H3: Greater familiarity with deepfake technologies will be associated with greater detection accuracy.
- H4: Greater skepticism toward digital media will be associated with greater detection accuracy.
- H5: Higher confidence will be associated with lower detection accuracy.

User-Level Factors in Authenticity Judgments of AI-Generated Media

Multiple psychological and technical factors may influence detection accuracy. In digital environments, users must evaluate the authenticity of audiovisual content under conditions of uncertainty, where traditional cues of “realness” may be less reliable. The present study examines whether analytical reasoning, emotional responses, technological familiarity, skepticism, and confidence systematically predict performance in authenticity judgments.

Analytical Reasoning

Research has shown a link between analytical thinking and identifying misinformation (Pennycook & Rand, 2019, 2021). In psychology, analytical thinking is often measured using cognitive reflection tasks

(Bialek & Pennycook, 2018). Cognitive reflection tasks may be questions that elicit an incorrect but intuitive response unless the individual pauses to reflect and override their initial impulse (Frederick, 2005; Bialek & Pennycook, 2018). This tendency to slow down, question one's instincts, and engage in more deliberate reasoning is key to avoiding deception in various forms. The Cognitive Reflection Test captures a tendency toward reflective reasoning rather than perceptual or visual detection skill and is therefore interpreted here as a marker of judgment style rather than sensory acuity.

In the context of deepfake detection, this reflective, analytical mindset may be similarly useful. As AI-generated media becomes more realistic, surface-level cues alone may no longer reliably indicate authenticity, increasing the importance of deliberate evaluation. Detecting AI-generated distortions may therefore require not only visual attention but also the ability to override intuitive impressions and engage in more careful judgment.

Previous studies suggest that training individuals to look for visual anomalies improves detection accuracy (Groh et al., 2021). Additionally, individuals with an analytical thinking style have been shown to be more resistant to deceptive content (Appel & Prietzel, 2022). From a signal detection theory perspective, analytical reasoning should increase sensitivity (d'), reflecting better discrimination between real and fake content (Green & Swets, 1966; Stanislaw & Todorov, 1999). Supporting this view, Alsobeh et al. (2024) suggested a positive role of analytical thinking in deepfake detection. On that basis, the present study examined whether analytical reasoning would be associated with greater detection accuracy.

Emotional Responses

Emotional responses may also influence detection accuracy when individuals evaluate potentially manipulated media. While slow, analytical thinking can enhance accuracy, fast, emotionally charged reactions could hinder it. This aligns with Kahneman's (2011) distinction between System 1 (fast, intuitive, and emotional) and System 2 (slow, deliberate, and logical) thinking.

Rather than focusing on specific discrete emotions, the present study conceptualizes emotional responses as general affective reactivity that may influence how individuals process uncertain information. Emotional reactivity may increase reliance on intuition, potentially impairing judgment when authenticity cues are ambiguous or misleading.

Prior research suggests that emotional states can bias information processing and decision-making. For example, emotional arousal may increase susceptibility to misleading content, while reduced cognitive engagement may limit careful evaluation. In AI-generated media contexts, these effects may be amplified because highly realistic audiovisual content can trigger intuitive trust, reducing critical scrutiny. Based on this reasoning, we hypothesized that emotional reactivity will impair detection accuracy. Pilot findings hinted at possible effects of affective responses, although the direction and magnitude remain uncertain.

Technological Knowledge

A common assumption in both research and practice is that increasing users' knowledge of deepfakes will improve their ability to evaluate content authenticity. Familiarity with social media platforms may shape how users interpret content, and some studies suggest that prior exposure to deepfakes can improve recognition (Nas & de Kleijn, 2024; Malinka et al., 2024). However, research also indicates that rapid, low-effort content consumption may reduce attention to accuracy and increase vulnerability to misinformation (Pennycook & Rand, 2019, 2021; Lewandowsky et al., 2012).

Importantly, evidence for knowledge-based improvements is mixed. While awareness and training may provide heuristics for identifying manipulated media, these cues may become less reliable as generative quality improves and as platforms reduce contextual and provenance signals. As a result, individuals may understand deepfakes conceptually but still struggle to apply that knowledge effectively in real-world evaluation tasks.

From an applied perspective, this raises a critical concern: awareness and familiarity may not be sufficient for supporting accurate user judgment in practice. Accordingly, the present study tests whether greater knowledge of deepfakes is associated with improved detection accuracy.

Trust and Skepticism

Digital trust refers to an individual's willingness to trust online platforms and organizations, including expectations that systems will safeguard content and operate reliably (Pietrzak & Takala, 2021). Investigating digital trust is important because it may influence how readily individuals accept information or question its authenticity. In digital environments, users often operate from a default assumption that content is genuine, which may increase vulnerability when authenticity cues are unreliable.

People with higher levels of digital trust may therefore be more likely to accept deepfakes as genuine, whereas skepticism may reduce susceptibility to misleading content (Lewandowsky et al., 2012). Individuals who exhibit skepticism or question information more readily have been shown to engage in more critical evaluation (Stavrova & Ehlebracht, 2018). This reasoning aligns with signal detection theory's criterion (c): trusting individuals may adopt a conservative bias ($c > 0$; judging clips as 'real'), whereas skeptics may adopt a more liberal response tendency ($c < 0$; judging clips as 'fake') (Stanislaw & Todorov, 1999; Köbis et al., 2021). Importantly, these differences reflect shifts in decision strategy rather than differences in perceptual ability.

Consistent with the preregistered trust-related hypothesis, the present analysis operationalized this expectation primarily through skepticism rather than through a broader multidimensional trust construct. This framing is also consistent with IS research distinguishing trust in technological systems from interpersonal trust, suggesting that authenticity judgments in digital environments may reflect both beliefs about content and beliefs about the system through which the content is encountered (McKnight et al., 2011). This framing helps explain how a group-level truth bias can coexist with individual differences in evaluation strategy, particularly in environments where users must make judgments under uncertainty.

Confidence

Research on participant confidence in deepfake detection consistently shows that human accuracy is only slightly above chance (Bray et al., 2023; Köbis et al., 2021). Despite this, individuals often report relatively high confidence in their judgments, even when performance is limited (Hashmi et al., 2024).

Participants make both false-positive and false-negative errors, and prior research frequently finds weak or inconsistent relationships between confidence and accuracy. This suggests that confidence may not always reliably correspond to performance. When individuals are highly confident but incorrect, they may be more likely to act on false beliefs or share misleading content.

From an applied perspective, confidence is especially important because it may influence whether users trust their judgments and act on them. High confidence may increase reliance on one's decision, regardless of whether that decision is correct. This creates potential risk in digital environments where users must evaluate uncertain or manipulated content.

Accordingly, based on prior research emphasizing overconfidence and inconsistent confidence–accuracy relationships, we hypothesized that higher confidence would be associated with lower detection accuracy. By identifying which user-level characteristics are associated with better performance, this study clarifies how individuals evaluate authenticity under uncertainty.

Taken together, these hypotheses examine user-level factors that may shape authenticity judgments in digital environments where traditional cues of authenticity are increasingly unreliable. In that sense, the study is intended as a behavioral foundation for later work on user-facing interventions and support mechanisms rather than as a direct test of a system feature or platform design.

Method

Participants

A total of 171 undergraduate students participated in the study. Participants were recruited from Eastern Kentucky University (EKU) and Southern Illinois University Carbondale (SIU). EKU students participated through the SONA research management system for course or extra credit. SIU students volunteered through classroom announcements and email invitations without compensation. All participants were 18 years of age or older and completed the online survey voluntarily. The final sample included students from diverse majors, primarily computer science/information technology (CS/IT) and psychology. The mean age was 26.6 years ($SD = 8.2$). All procedures were approved by the institutional review boards at both universities. Participation was anonymous and conducted in accordance with APA ethical guidelines.

Preregistration and Power Analysis

Hypotheses were preregistered as directional predictions about relationships among the study variables. A priori power analysis using G*Power ($\alpha = .05$, $f^2 = .10$, nine predictors) indicated that at least 166 participants were needed to achieve 80% power to detect a small-to-medium overall effect in the planned regression model. The final sample ($N = 171$) exceeded this threshold. However, power for individual coefficients and exploratory analyses may still have been limited. Five confirmatory hypotheses (H1–H5) examined analytical reasoning, emotion, knowledge, skepticism, and confidence as predictors of detection accuracy.

Open Science Statement

All materials, data, and analytic code are archived on the Open Science Framework (OSF; <https://osf.io/zef2c/>) and will be made publicly available upon publication.

Materials and Procedure

The procedure replicated and extended a prior exploratory pilot study (Alsobeh et al., 2024) by incorporating preregistered hypotheses and Signal Detection Theory (SDT) analyses. Participants completed the 20–30-minute online survey via Qualtrics. Each participant viewed 10 short video clips (five authentic, five deepfake) presented in randomized order. Each clip was approximately 20–30 seconds in length and drawn from publicly available online sources, reflecting common social and informational media contexts (e.g., public statements, news-style segments). After each clip, participants indicated whether it was real or fake and rated their confidence on a 5-point Likert scale (1 = not at all confident to 5 = extremely confident).

The task was designed to approximate short-form audiovisual judgments commonly encountered in digital media environments, although it did not reproduce full platform context such as comments, account metadata, provenance indicators, sharing cues, or other authenticity indicators that may shape judgments in real digital environments (Sherman et al., 2021; Feng et al., 2023; Gamage et al., 2025). Participants also completed validated measures of individual differences:

- Analytical reasoning: Seven-item Cognitive Reflection Test (Frederick, 2005).
- Emotional states: Depression, Anxiety, and Stress Scale-21 (DASS-21) and the Positive and Negative Affect Schedule (PANAS).
- Knowledge of deepfakes: A familiarity-based Technology Recognition Test adapted from author-recognition paradigms, assessing recognition of 26 software tools (e.g., DeepFaceLab, Zao, DALL-E).
- Skepticism: Items assessing skepticism toward digital media content. Consistent with the preregistration, trust-related expectations were framed in terms of both digital trust and skepticism; in the present analysis, this construct was operationalized primarily through the skepticism measure.
- Confidence: Mean of trial-level confidence ratings across all judgments.

Participants also provided demographic information (age, gender, major, education level, race/ethnicity).

Signal Detection Theory Approach

Because raw accuracy conflates discrimination ability and response bias, results were analyzed using Signal Detection Theory (SDT). Sensitivity (d') measured the ability to discriminate real from fake videos, while criterion (c) indexed response bias (positive = conservative “real” tendency; negative = liberal “fake” tendency). The Hautus (1995) correction ($+5 / N$) was applied to avoid infinite z-scores. This approach distinguishes true discrimination ability from response tendencies.

Following Düvel and Kopiez (2022), Botella and Suero (2020), Stanislaw and Todorov (1999), and Hautus (1995), we used established signal detection methods appropriate for this design. Although each participant judged only 10 clips (5 real, 5 deepfake), this structure is consistent with some applied SDT studies, although it necessarily involves tradeoffs in measurement precision. Consistent with prior findings (Köbis et al., 2021; Diel et al., 2024), a small conservative “realness” bias ($c > 0$) was expected.

Data Analysis

Analyses were conducted in Jamovi (v2.6.44). Internal consistency estimates were calculated for all multi-item scales used in the analyses. Multicollinearity diagnostics were examined for the regression model using variance inflation factors (VIFs) and tolerance statistics.

The preregistered workflow included:

- One-sample t-tests comparing accuracy scores to chance levels (5/10, 2.5/5).
- One-sample t-tests on d' and c against zero.
- Pearson correlations and multiple regressions testing preregistered predictors (H1–H5).
- Exploratory regressions for non-preregistered factors (e.g., academic major, demographics).
- All tests were two-tailed ($\alpha = .05$) and results were reported per APA 7th standards (R^2 , F , B or β as appropriate, df , p).

Results

Descriptive Statistics

Participants performed slightly above chance in identifying deepfakes. The mean total accuracy was 5.70 ($SD = 1.63$) out of 10, which was slightly above the chance level of 5, $t(170) = 5.58$, $p < .001$, Cohen’s $d = .43$.

The mean accuracy for real videos was 3.01 (SD = .99) out of 5, which was also slightly above chance (2.5/5), $t(170) = 6.71$, $p < .001$, Cohen's $d = .51$. The mean accuracy for deepfake videos was 2.69 (SD = 1.20) out of 5, which was slightly but significantly above chance (2.5/5), $t(170) = 2.06$, $p = .041$, Cohen's $d = .16$. The multi-item scales showed acceptable to excellent internal consistency: Analytical Reasoning ($\alpha = .69$), Confidence ($\alpha = .83$), Knowledge ($\alpha = .91$), DASS Anxiety ($\alpha = .79$), DASS Depression ($\alpha = .90$), DASS Stress ($\alpha = .81$), Positive Affect ($\alpha = .86$), and Negative Affect ($\alpha = .89$).

Signal Detection Theory Analysis

Signal detection metrics are summarized in Table 1. The mean d' was .380 (SE = .068), which was significantly greater than zero, $t(170) = 5.61$, $p < .001$. This indicates that, on average, participants were able to discriminate between real and deepfake videos at a level better than chance. The mean c was .088 (SE = .031), which was significantly different from zero, $t(170) = 2.87$, $p = .005$. This indicates a conservative bias, meaning that participants were more likely to classify videos as real rather than as deepfakes. These findings align with prior research showing weak sensitivity ($d' \approx .48$; Köbis et al., 2021) and a conservative “realness” bias, while being somewhat stronger than the near-chance pooled results of the recent meta-analysis (Diel et al., 2024).

Table 1. Signal Detection Theory Results.

Measure	M	SE	t(170)	p	Interpretation
d' (sensitivity)	.38	.07	5.61	< .001	Weak but above-chance discrimination
c (bias)	.09	.03	2.87	.005	Conservative bias (tendency to classify as real)

Note. M = mean; SE = standard error; t = t statistic

Correlation Analysis

A correlation matrix of key variables is presented in Table 2. Detection accuracy was positively correlated with analytical reasoning ($r = .16$, $p < .05$), skepticism ($r = .19$, $p < .05$), and confidence ($r = .26$, $p < .001$). A small negative correlation was observed between response bias (c) and accuracy ($r = -.19$, $p < .05$), indicating that a more liberal response tendency (i.e., greater willingness to classify content as fake) was associated with higher accuracy. No significant associations were observed between emotional variables or knowledge and detection accuracy.

Table 2. Correlations among key study variables (N = 171).

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Accuracy	—											
2. Analytical	.16*	—										
3. Knowledge	.00	.15*	—									
4. Confidence	.26***	.02	.19*	—								
5. Skepticism	.19*	-.02	-.04	.08	—							
6. DASS Anxiety	.05	-.01	.02	.06	.01	—						
7. DASS Depression	.11	.02	.00	.05	-.03	.64***	—					
8. DASS Stress	.07	-.08	-.01	.12	.07	.67***	.72***	—				
9. Positive Affect	-.05	-.10	.12	.14	-.09	-.09	-.27***	-.14	—			
10. Negative Affect	.13	-.07	.07	.02	.01	.65***	.70***	.65***	-.01	—		
11. d' (sensitivity)	.99***	.15*	.01	.26***	.19*	.04	.11	.07	-.05	.13	—	
12. c (bias)	-.19*	-.06	-.02	.10	.09	-.05	-.02	-.09	-.12	-.11	-.18*	—

Note. Values are Pearson's *r*. * $p < .05$, ** $p < .01$, *** $p < .001$.

Predictors of Detection Accuracy

A multiple regression analysis examined predictors of detection accuracy. The overall model was significant, $F(9, 161) = 3.26$, $p = .001$, with $R^2 = .154$. Multicollinearity diagnostics indicated acceptable model stability, with VIF values ranging from 1.04 to 3.06 and tolerance values ranging from .33 to .96. Analytical reasoning positively predicted accuracy, $B = .1508$, $t(161) = 2.253$, $p = .026$. Skepticism also positively predicted accuracy, $B = .2613$, $t(161) = 2.319$, $p = .022$. Confidence was positively associated with accuracy, $B = .0630$, $t(161) = 3.521$, $p < .001$. Emotional measures and deepfake technology familiarity did not significantly predict detection accuracy in the multivariable model. Because the present study did not assess trial-level confidence calibration directly, this finding should be interpreted as a person-level association between mean confidence and mean performance, not as evidence that participants were well calibrated on individual judgments.

Hypothesis Tests

H1: Supported. Analytical reasoning positively predicted detection accuracy, $B = .1508$, $t(161) = 2.253$, $p = .026$.

H2: Not supported. Emotional reactivity measures did not significantly predict detection accuracy in the multivariable model.

H3: Not supported. Familiarity with deepfake technologies did not significantly predict detection accuracy, $B = -.0220$, $t(161) = -1.001$, $p = .318$.

H4: Supported. Greater skepticism positively predicted detection accuracy, $B = .2613$, $t(161) = 2.319$, $p = .022$.

H5: Not supported. Contrary to the preregistered prediction, confidence positively predicted detection accuracy, $B = .0630$, $t(161) = 3.521$, $p < .001$.

Table 3. Multiple Regression to Examine Predictors of Accuracy.

	<i>B</i>	<i>t</i>	<i>p</i>
Intercept	2.7488	3.661	< .001
Analytical	.1508	2.253	.026
Positive Affect	-.0108	-.622	.535
Negative Affect	.0427	1.958	.052
DASS Stress	-.0260	-.635	.526
DASS Anxiety	-.0337	-.858	.392
DASS Depression	.0123	.307	.759
Skepticism	.2613	2.319	.022
Knowledge	-.0220	-1.001	.318
Confidence	.0630	3.521	< .001
$R^2 = .154$ $F_{(9, 161)} = 3.26$ $p = .001$			

Discussion

This study examined human performance in detecting deepfake videos and identified which user-level factors are associated with more accurate authenticity judgments in AI-mediated environments. The findings are best understood as baseline behavioral evidence rather than as a direct test of platform features or intervention effects.

Summary of Findings

Participants performed only slightly above chance in distinguishing real from manipulated content, consistent with prior research (Diel et al., 2024; Köbis et al., 2021). Signal detection analyses revealed weak but reliable sensitivity alongside a conservative bias toward classifying content as authentic, reflecting a persistent “truth bias” in user judgments (Levine, 2022; Köbis et al., 2021).

From an applied IS perspective, this study contributes behavioral baseline evidence about where unaided authenticity judgments may be insufficient in AI-mediated information environments. Rather than treating synthetic-media evaluation solely as an individual cognitive problem, the findings also point to a broader design and governance challenge for digital platforms and information systems. In particular, the results suggest that familiarity alone may be insufficient, whereas evaluative processes such as skepticism and analytical reasoning were more closely tied to performance, helping identify where user-facing supports may be most needed.

Predictors of Detection

Analytical reasoning emerged as a significant positive predictor of detection accuracy, supporting prior work linking reflective thinking to improved misinformation detection (Martire et al., 2023). Skepticism was also positively associated with accuracy, consistent with recent work highlighting its adaptive role in evaluating digital content (Pennycook et al., 2021). Although the preregistration framed this hypothesis in terms of both digital trust and skepticism, the reported model operationalized this expectation through skepticism, which emerged as a reliable predictor of detection accuracy.

Contrary to the preregistered hypothesis, confidence was positively associated with detection accuracy in this sample. This indicates that participants who were more confident on average also tended to perform somewhat better at the person level. One possible explanation is that mean confidence may partly reflect broader judgment-related characteristics, such as perceived competence, decisiveness, or self-efficacy, rather than trial-level calibration alone. It is also possible that participants who performed better were somewhat aware of their stronger performance, producing a positive person-level confidence–accuracy association. However, because the present study did not directly assess trial-level calibration, this finding should not be interpreted as evidence of well-calibrated confidence or as evidence that overconfidence was absent. Conclusions about confidence–accuracy alignment therefore remain tentative.

Our study examined emotional factors (e.g., anxiety, stress, positive and negative affect) as predictors of deepfake detection accuracy. Based on prior work linking emotional reactivity to heuristic processing and reduced critical thinking (Forgas, 1995; Groh et al., 2021), we expected that emotions would impair detection accuracy. However, emotional reactivity was not a significant predictor of detection performance. This suggests that affective states may play a more limited role in deepfake detection than expected, or that their influence is more context-dependent than captured in the present design. Consistent with this interpretation, no significant associations were observed between emotional variables and sensitivity (d') (Hautus, 1995).

We also expected knowledge of deepfakes to be associated with detection accuracy; however, no significant relationship was observed. This measure was intended as a proxy for familiarity with deepfake-related technologies rather than as a direct assessment of detection skill, digital literacy, or media verification expertise. This null result should be interpreted cautiously because the measure captured familiarity with deepfake-related technologies rather than applied evaluative expertise. Although familiarity and visual literacy are often assumed to support deception detection (Botella & Suero, 2020), the present findings suggest that this familiarity-based operationalization was not associated with better performance in the present task. One possible explanation is that the technology recognition measure did not capture the contextual, visual, and strategic expertise required for real-world evaluation.

Taken together, these results suggest that analytical reasoning, skepticism, and mean confidence showed the clearest associations with detection accuracy in this dataset, whereas emotional reactivity and the familiarity-based knowledge measure did not emerge as reliable unique predictors. These patterns should nonetheless be interpreted in light of the study's correlational design and limited task context.

This study builds on prior exploratory work (Alsobeh et al., 2024). The preregistered follow-up provided further evidence that analytical reasoning and skepticism were reliable predictors of deepfake detection. It did not support effects for emotional reactivity or the familiarity-based knowledge measure, but it did clarify the role of confidence in detection performance. Together, these findings refine our understanding of individual differences in deepfake detection and highlight the value of combining accuracy measures with signal detection theory metrics (d' and c).

Reporting both accuracy and SDT metrics aligns the study with current best practices in deception detection and synthetic media evaluation research (Stanislaw & Todorov, 1999; Diel et al., 2024). Although each participant judged only 10 clips (5 real, 5 deepfake), standard corrections (Hautus, 1995) were applied to stabilize estimates. Even so, the small fixed stimulus set remains an important limitation on measurement precision and generalizability. This approach allows clearer separation of discrimination ability (d') and response bias (c), thereby strengthening interpretation of the results.

Practical Implications for Digital Platforms and Information Systems

The present findings have practical implications for digital platforms and information systems that increasingly mediate authenticity judgments under uncertainty. Because the present task intentionally omitted feed-level context, metadata, provenance indicators, and social signals, the following implications should be understood as a rationale for testing user-facing supports rather than as evidence that any particular platform feature will improve performance in practice.

First, the present findings provide a rationale for future work targeting confidence calibration rather than detection alone. Although confidence was positively associated with performance at the person level, the present study did not assess trial-level calibration directly. It is also important to note that confidence is operationalized differently across the literature. Much prior work examines item-level calibration or overconfidence relative to trial-level accuracy, whereas the present study assessed confidence at the person level; accordingly, the positive association observed here should not be interpreted as evidence of well-calibrated confidence in the stricter metacognitive sense.

Second, the findings suggest that platforms may benefit from evaluating lightweight verification prompts or friction before users reshare potentially manipulated media. Because analytical reasoning and skepticism were associated with better detection performance, interface designs that encourage slower, more deliberate scrutiny of authenticity cues may be worth testing in high-risk contexts. Examples include “check source before sharing” prompts, brief verification reminders, or added friction when users attempt to reshare unverified content.

Third, the results provide a rationale for testing provenance and source-context cues as external supports for authenticity judgments. In the present study, awareness-based approaches alone appeared insufficient for improving detection accuracy, at least when knowledge was operationalized as familiarity with deepfake-related technologies. In practice, this suggests that platforms and organizations may need to evaluate external supports—such as provenance labels, source-verification panels, or authenticity indicators—rather than relying on users’ prior knowledge alone.

From an IS/HCI perspective, this pattern of modest discrimination ability alongside a tendency to classify content as real highlights a core design question for platforms: how to support appropriate reliance when verification is costly and ground truth is uncertain. Prior HCI work suggests that user-facing supports such as provenance displays and AI-generated-content labels can shift perceived trust and accuracy, but may also be misunderstood or overgeneralized if cues are poorly designed or inconsistently applied. Taken together, the present findings provide a behavioral baseline for future interface and governance experiments examining which cues—and which designs—improve calibration without producing unintended reductions in trust for authentic content.

Although the study does not directly test a platform feature or interface intervention, it identifies user-level patterns that can inform later work on platform supports, verification cues, and other system-level mechanisms.

More broadly, these findings suggest that synthetic-media evaluation should be treated as a shared platform and governance responsibility rather than as an individual user burden alone. They also provide a rationale for future evaluation of structured, calibration-oriented interventions designed to improve confidence–accuracy alignment in digital environments without encouraging blanket skepticism. Such interventions may be especially valuable if they help users reflect more proportionately on uncertainty while preserving agency rather than simply increasing caution.

Limitations and Future Directions

Several limitations should be considered when interpreting the present findings. The relatively small fixed stimulus set (10 clips per participant) may have introduced measurement noise, made performance partly dependent on the selected clips, and limited generalizability. A larger and more diverse stimulus pool could yield more stable estimates of both accuracy and response bias and reduce the influence of clip-specific characteristics. Because the task intentionally omitted feed-level context, metadata, provenance indicators, and social signals, the present study should be understood as a baseline assessment of unaided judgment rather than as a simulation of authentic platform use. Future studies should incorporate larger and more diverse video sets and examine authenticity judgments in more realistic interface contexts that include source, social, and provenance cues.

The sample consisted of undergraduate students from two universities, which may not fully represent broader populations. Future research with more diverse samples is needed to assess generalizability across user groups.

The study employed a correlational design, limiting causal inference and preventing direct conclusions about whether targeted interventions or interface features would improve performance. Future research should examine how user-level factors interact with system-level features such as verification prompts, interface cues, provenance displays, AI-generated-content labels, and confidence calibration tools, ideally in more realistic platform or organizational contexts.

Conclusion

Overall, the findings indicate that detection performance was shaped more by how participants evaluated uncertain information than by familiarity with deepfake-related technologies alone. Analytical reasoning, skepticism, and confidence emerged as the clearest predictors of detection accuracy, whereas emotional measures and technology familiarity did not reliably predict performance.

These findings replicate and extend prior work (Köbis et al., 2021; Diel et al., 2024) while highlighting a key applied insight: effective evaluation of AI-generated media may depend less on familiarity with deepfake-related technologies alone and more on how users process and interpret uncertain information. More broadly, the study provides baseline behavioral evidence that can inform future research on platform supports, user-facing interventions, and other tools designed to assist authenticity judgments when traditional cues become unreliable.

Acknowledgements

Disclosure of Interests

The authors declare that they have no competing interests.

Disclosure of AI Use

The authors declare they have used AI services, specifically Grammarly and ChatGPT, for grammar correction and minor style refinements. They carefully reviewed all suggestions from these services to ensure the original meaning and factual accuracy were preserved.

References

- Alsobeh, A., Franklin, A., Woodward, B., Porche, M., & Siegelman, J. (2024). Unmasking media illusion: Analytical survey of deepfake video detection and emotional insights. *Issues in Information Systems*, 25(2), 96–112. https://doi.org/10.48009/2_iis_2024_108
- Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4). <https://doi.org/10.1093/jcmc/zmac008>
- Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, 50(5), 1953–1959. <https://doi.org/10.3758/s13428-017-0963-x>
- Bikku, T., Bhargavi, K., Bhavitha, J., Lalithya, Y., & Vineetha, T. (2023). Deep residual learning for unmasking deepfake. *Proceedings of the 2023 International Conference on Advanced Electronics, Communication and Information Systems (ICAECIS)*
- Botella, J., & Suero, M. (2020). Recovering the variance of d' from hit and false alarm statistics. *Behavior Research Methods*, 52(1), 1–22. <https://doi.org/10.3758/s13428-018-1181-x>
- Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect ‘deepfake’ images of human faces. *Journal of Cybersecurity*, 9(1). <https://doi.org/10.1093/cybsec/tyad011>
- Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1), 108–117. <https://doi.org/10.1016/j.jarmac.2018.09.005>
- Diel, A., Lalgı, 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, 100538. <https://doi.org/10.1016/j.chbr.2024.100538>
- Düvel, N., & Kopiez, R. (2022). The paired A–Not A design within signal detection theory: Description, differentiation, power analysis and application. *Behavior Research Methods*, 54(5), 2334–2350. <https://doi.org/10.3758/s13428-021-01728-w>
- Feng, K. J. K., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2). <https://doi.org/10.1145/3610061>
- Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, 117(1), 39–66. <https://doi.org/10.1037/0033-2909.117.1.39>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. K. (2025). Labeling synthetic content: User perceptions of label designs for AI-generated content on social media. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*.

- Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2). <https://doi.org/10.14763/2019.2.1407>
- Grahek, I., Shenhay, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neuroscience & Biobehavioral Reviews*, 102, 371–381. <https://doi.org/10.1016/j.neubiorev.2019.04.011>
- Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. John Wiley.
- Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*.
- Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152. <https://doi.org/10.1089/cyber.2021.29208.jth>
- Hashmi, A., Shahzad, S. A., Lin, C.-W., Tsao, Y., & Wang, H.-M. (2024). Unmasking illusions: Understanding human perception of audiovisual deepfakes. *TechRxiv preprint*. <https://doi.org/10.36227/techrxiv.171560542.29368554/v1>
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers*, 27(1), 46–51. <https://doi.org/10.3758/BF03203619>
- Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
- Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), 103364. <https://doi.org/10.1016/j.isci.2021.103364>
- Levine, T. R. (2022). Truth-default theory and the psychology of lying and deception detection. *Current Opinion in Psychology*, 47, 101380. <https://doi.org/10.1016/j.copsyc.2022.101380>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/1529100612451018>
- Malinka, K., Firc, A., Šalko, M., Prudký, D., Radačovská, K., & Hanáček, P. (2024). Comprehensive multiparametric analysis of human deepfake speech recognition. *EURASIP Journal on Image and Video Processing*, 2024(1), Article 24. <https://doi.org/10.1186/s13640-024-00641-4>
- Martire, K. A., Robson, D. G., Drew, M., Nicholls, K., & Faasse, K. (2023). Thinking false and slow. *Psychonomic Bulletin & Review*, 30(5), 1886–1901. <https://doi.org/10.3758/s13423-023-02321-2>
- Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. (2020). Methods of deepfake detection based on machine learning. *Proceedings of the IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*.

- McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2), Article 12. <https://doi.org/10.1145/1985347.1985353>
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>
- Nas, E., & de Kleijn, R. (2024). Conspiracy thinking and social media use are associated with ability to detect deepfakes. *Telematics and Informatics*, 87, 102093. <https://doi.org/10.1016/j.tele.2023.102093>
- Nelson, B. D., Hodges, A., Hajcak, G., & Shankman, S. A. (2015). Anxiety sensitivity and threat anticipation. *Journal of Anxiety Disorders*, 33, 62–71. <https://doi.org/10.1016/j.janxdis.2015.05.003>
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, 590–595. <https://doi.org/10.1038/s41586-021-03344-2>
- Pietrzak, P., & Takala, J. (2021). Digital trust: A systematic literature review. *Futures Studies Research Journal*, 9(3), Article 4. https://doi.org/10.23762/FSO_VOL9_NO3_4
- Salman, S., Shamsi, J., & Qureshi, R. (2023). Deep fake generation and detection. *IT Professional*, 25(1), 52–59. <https://doi.org/10.1109/MITP.2022.3230353>
- Schwarz, N., Jalbert, M., Noah, T., & Zhang, L. (2020). Metacognitive experiences as information. *Consumer Psychology Review*, 4(1), 4–25. <https://doi.org/10.1002/arcp.1067>
- Sherman, I. N., Stokes, J. W., & Redmiles, E. M. (2021). Designing media provenance indicators to combat fake media. *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '21)*.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. <https://doi.org/10.3758/BF03207704>
- Stavrova, O., & Ehlebracht, D. (2018). The cynical genius illusion. *Personality and Social Psychology Bulletin*, 45(2), 254–269. <https://doi.org/10.1177/0146167218783195>