Discerning Truth: A Qualitative Comparative Analysis of Reliance on AI Advice in Deepfake Detection

Research Paper

Christiane Ernst[1]

[1] University of Innsbruck, Department of Information Systems, Production and Logistics Management, Innsbruck, Austria
christiane.ernst@uibk.ac.at

Abstract. Recent advancements in artificial intelligence (AI) have enabled the creation of deepfakes - highly realistic, manipulated multimedia that challenge the ability to discern authenticity. This paper investigates reliance on AI advice within deepfake detection using a judge-advisor system, where participants first provide an initial judgment and then could revise it after viewing an algorithm's evaluation. Data were analyzed using Qualitative Comparative Analysis to assess how AI literacy, trust in a deepfake detection tools recommendation, and algorithm aversion interact with authenticity assessment in shaping reliance on AI advice. Findings reveal that participants only revised their decisions when the tool indicated that the video was genuine. Multiple sufficient configurations indicate that combinations of high aversion, low trust, or high AI literacy, alongside authenticity assessment, drive reliance on AI advice. These results advance understanding of human-AI interaction in deepfake detection and offer practical insights for designing more resilient and transparent deepfake detection systems.

Keywords: Deepfake, Reliance on AI Advice, Qualitative Comparative Analysis (QCA), Human-AI Collaboration

1 Introduction

Recent advancements in artificial intelligence (AI) have unlocked the capability to create deepfakes, highly realistic, AI-generated multimedia in which videos, images, or audio are synthetically manipulated to appear authentic (Altuncu et al., 2024). While deepfakes hold promise for innovative applications in entertainment and creative industries (Altuncu et al., 2024), they simultaneously blur the line between real and manipulated media, making it increasingly difficult for audiences to distinguish genuine from manipulated content (Vasist & Krishnan, 2022). This growing ambiguity poses serious threats, including the spread of misinformation, reputational damage, and the destabilization of political and social systems. A notable example is a 2022 deepfake video of President Zelenskyy falsely calling for military surrender, illustrating the technology's destabilizing potential (Telegraph, 2022; Vasist & Krishnan, 2022).

20th International Conference on Wirtschaftsinformatik,
September 2025, Münster, Germany

Effectively addressing these challenges requires both enhanced media literacy, enabling individuals to critically assess multimedia content (Hoes et al., 2024) and technological tools, such as deepfake detection tools (Rana et al., 2022). Those tools effectiveness, as seen with TikTok's AI-generated content flag, still depends significantly on user trust and engagement (Alexander et al., 2018). Hence, reliance on AI advice becomes crucial; calibrated trust and reliance on accurate algorithmic recommendations may reduce harmful deepfake impacts (Wischnewski et al., 2023).

Despite extensive research on deepfake detection (Altuncu et al., 2024; Heidari et al., 2023) and media literacy interventions (Hoes et al., 2024), theoretical insights remain sparce regarding cognitive and social processes driving reliance on AI advice. Few studies explicitly examine how users integrate automated recommendations into their judgment, especially concerning trust calibration (Wischnewski et al., 2023), cognitive biases (Dietvorst et al., 2015), and prior AI experience. Addressing this theoretical gap is crucial for better understanding decision making behavior in the context of deepfake detection.

To address this, the present research investigates the research question: *What factors influence participants' reliance on AI advice in the context of deepfake detection?*

Understanding users' decision-making processes regarding reliance on or rejection of algorithmic recommendations has profound implications for optimizing human-AI collaboration in combating misinformation (Kaur et al., 2024). The study employs a judge-advisor system in an online experiment, analyzing collected data through Qualitative Comparative Analysis (QCA) to systematically assess how user trust, AI literacy, and other psychological factors influence reliance on AI advices.

By linking the technological evolution of deepfakes with human cognitive decision-making processes, this paper aims to shed light on both the promise and limitations of current deepfake detection systems. These findings offer insights crucial for designing more effective detection systems that encourage critical engagement and balanced reliance on algorithmic recommendations in an era of pervasive AI-driven media manipulation.

2 Theoretical Background

2.1 Deep Fake Technology and Detection Tools

Deepfakes refer to hyper-realistic media, including videos, images, and audio, generated or altered using sophisticated deep learning techniques (Altuncu et al., 2024; Fagni et al., 2021; Heidari et al., 2023). Initially emerged as a method for face-swapping in videos, deepfake technology has rapidly evolved, now enabling realistic synthetic voices and entirely fabricated personas primarily through Generative Adversarial Networks (Rana et al., 2022).

On the beneficial side, deepfakes have promising applications in film, advertising, and digital arts, enabling innovative storytelling and creative expression (Kaur et al., 2024). However, their negative potential is equally significant, facilitating misinformation, reputational damage, and even political or financial destabilization (Köbis et

al., 2021; Vasist & Krishnan, 2022). These dual effects underscore the complexity of deepfakes and the critical need for effective countermeasures.

To mitigate risks, researchers advocate a two-fold strategy: enhancing media literacy to better equip individuals to discern manipulated content (Hoes et al., 2024; Kietzmann et al., 2020), and developing advanced AI-based tools to automatically detect synthetic media (Al-Khazraji et al., 2023). Current detection methods predominantly employ Convolutional Neural Networks and occasionally Recurrent Neural Networks, to identify slight inconsistencies introduced during the synthesis process (Heidari et al., 2023; Kaur et al., 2024).

These tools identify subtle visual and acoustic inconsistencies, such as unnatural facial movements or spectral noise, that arise during media synthesis (Heidari et al., 2023; Verdoliva, 2020). However, they remain limited in generalizability and are vulnerable to adversarial attacks, fueling a technological arms race between creators and detectors (Gowrisankar & Thing, 2024; Kaur et al., 2024).

Deepfakes, while related to fake news, pose unique detection challenges due to their dynamic and multimedia nature (Kumari et al., 2022; Pennycook & Rand, 2019). Fake news refers to deliberately misleading information, amplified by social media and algorithmic curation. Information Systems (IS) research has examined its characteristics (Budak et al., 2024; Khan et al., 2022), trust-building mechanisms (Rochlin, 2017), and countermeasures (Chen et al., 2023; Kießling et al., 2021). Recently, scholars have also started investigating deepfakes explicitly as tools for fake news dissemination, studying their unique attributes and how audiences interact with such synthetic misinformation (Feuerriegel et al., 2023; Vaccari & Chadwick, 2020).

Overall, while deepfake detection tools constitute an essential defensive layer, their existing limitations underscore the need for continued advancements and complementary strategies. Effective solutions should integrate cross-verification with trusted sources and enfoster individuals' capabilities to critically evaluate digital content.

2.2 Reliance on AI advice

Modern decision-making is increasingly augmented by algorithmic recommendations. Algorithms now guide a wide range of decisions, from interpreting medical images to recommending movies and optimizing routes (e.g., Ochmann et al., 2020; Reich et al., 2023). They serve as crucial advisors in both everyday choices and high-stakes domains such as hiring (Dargnies et al., 2024), lending (Hessler et al., 2022), and medical diagnostics (Jussupow et al., 2021). This widespread use has spurred growing interest in understanding how individuals perceive and integrate algorithmic advice into their decision-making processes.

Despite these advances, algorithm aversion, the tendency to prematurely reject algorithmic advice, especially after errors, is a bias driven by users' disproportionate reactions to mistakes, even when performance remains high (Dietvorst et al., 2015; Jussupow et al., 2024). Studies show that people weigh human input more heavily than algorithmic input when combining advice (Lu & Zhang, 2024) and judge professionals more harshly for following algorithmic recommendations rather than humans advice (Bauer & Gill, 2024). This bias often leads to the premature abandonment of automated

systems, a pattern less evident when human advice is involved (Bonezzi et al., 2022; Renier et al., 2021).

Error sensitivity is crucial in high-stakes environments like deepfake detection. While some experimental paradigms portray algorithms as highly accurate, real-world detection tools often show variable performance influenced by manipulation techniques, input modalities, and domain-specific generalizability (Jussupow et al., 2024). Users may anticipate potential errors, even if not directly observable, resulting in cautious or inconsistent reliance on AI. This underscores the significance of perceived reliability in shaping user trust and reliance on AI advice.

Several factors shape reliance on algorithmic advice, including user-level characteristics and perceptions of system reliability (Klingbeil et al., 2024). Among the most commonly referenced are trust, AI literacy, and algorithm aversion. Trust in algorithmic systems is a central determinant in advice-taking behavior. Users must have confidence in an algorithm's competence and intentions to accept its recommendations (Lee & See, 2004). Prior research has identified several factors that shape trust in algorithmic advice, including perceived performance and accuracy (Kim & Song, 2023), transparency and explainability (Wanner et al., 2022), and the stakes of the decision context (Saragih & Morrison, 2022).

AI literacy, understood as users' knowledge of artificial intelligence systems and their principles, plays an essential role in moderating reliance behavior (Pinski & Benlian, 2024). When users comprehend an algorithm's operational principles, they are more likely to accept its recommendations and use them appropriately (You et al., 2022). Enhanced AI literacy helps unravel the so-called "black box" nature of these systems, fostering more informed and balanced reliance (Cadario et al., 2021).

While substantial research has elucidated factors influencing reliance on AI advice (e.g., Bauer & Gill, 2024), a critical gap remains in understanding how these factors interact in novel, high-stakes contexts—such as deepfake detection. The unique challenges posed by deepfake technology (e.g., the rapid evolution of synthetic media and the high potential for misinformation) call for an examination of whether established drivers of reliance on AI advice apply similarly (Jussupow et al., 2024; Leffrang & Mueller, 2024). In particular, the interplay between user trust, error sensitivity, and AI literacy in the context of deepfake detection tools remains underexplored.

Deepfake detection represents a distinct and underexplored domain of reliance on AI advice for several reasons. Unlike traditional contexts such as navigation or hiring, users must evaluate not just the correctness of the recommendation, but also the authenticity of highly deceptive visual or auditory content, often without clear ground truth or immediate feedback. This adds a layer of epistemic uncertainty that complicates reliance calibration. Moreover, deepfake detection is often emotionally and politically charged, involving content with reputational, legal, or social consequences. These contextual features may intensify cognitive biases such as algorithm aversion or overreliance, making it imperative to examine whether established insights from other domains generalize to this setting.

3 Method

3.1 Experimental Design and Procedure

An online experiment was conducted to examine the influence of different factors on reliance on AI advice. Upon providing informed consent, participants received a detailed description of the study. They were then randomly presented with four videos (two genuine and two deepfake). For each video, participants first watched the clip and then rated whether they believed it to be genuine or a deepfake. Subsequently, participants were shown the outcome produced by the deepfake detection tool of deepware.ai (see Figure 1). The tool was introduced to them in the following way: "To counter and detect Deepfakes, software and hardware tools have been developed specifically to detect them. These tools analyse visual and acoustic anomalies, such as errors in facial movements, unnatural voices, and examine file metadata to identify possible manipulations." Although participants were unaware, the tool's recommendations presented in the experiment were 100% accurate. After viewing the tool's output, participants were given the opportunity to revise their initial judgment regarding the video's authenticity. This procedure is known as judge-advisor system paradigm (Sniezek & Buckley, 1989). Following the four video-rating rounds, participants completed a post-experiment questionnaire and were subsequently debriefed regarding which of the videos watched were deepfakes. Debriefing was essential to ensure participants understood the deceptive nature of deepfake content, to correct any potential misconceptions arising from exposure to manipulated media, and to mitigate potential distress or confusion caused by viewing realistic but falsified videos (Greenspan & Loftus, 2022).

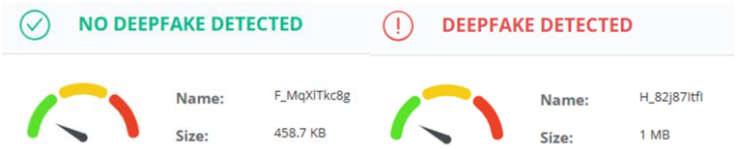

Figure 1. Stimuli for the Outcome of the Deepfake Detection Tool (Genuine vs. Deepfake)

3.2 Sample Description

A total of 62 participants were initially recruited from the University of Innsbruck's mailing lists. 10 participants were excluded because they failed at least one of the two administered attention checks (e.g., "This is an attention check. If you are reading this, please select "strongly agree"), resulting in a final sample of 52 participants. The final sample comprised 27 females and 25 males, with a mean age of 24.86 years (M = 24.86, SD = 4.74). Regarding their highest level of education, 33 participants reported having a high school diploma as their highest qualification, 14 held a bachelor's degree, and 5 had attained a master's degree. Looking at the participants initial judgement, they achieved a detection rate of 39.9% overall.

3.3 Qualitatative Comparative Analysis

QCA is employed as a robust method that is gaining attention in IS research (Lee et al., 2019; Mattke et al., 2021). This approach is particularly appropriate when the sample size is too small for general linear regression models yet too extensive for traditional cross-case analysis (Mattke et al., 2022; Oana et al., 2021). Unlike regression or structural equation modeling, which focus on net effects of individual variables (Mueller & Hancock, 2018), QCA identifies equifinal and conjunctural causal configurations, making it well suited for exploring complex, condition-based outcomes like reliance on AI advice. Rooted in set theory, the mathematical discipline that examines the relationships among well-determined collections of objects (Schneider & Wagemann, 2012), QCA allows researchers to systematically evaluate configurations of causal conditions (Oana et al., 2021). In this study, a combination of fuzzy and crisp sets is utilized to accommodate the mixed nature of the data, where Likert-scale measurements are treated as fuzzy sets and other binary variables are modeled as crisp sets (Schneider & Wagemann, 2012). The analytical procedure adheres to the established guidelines for executing QCA in IS by Mattke et al. (2021) ensuring a rigorous and context-sensitive investigation of complex causal relationships.

3.4 Measurement and Calibration

Causal Conditions. Four conditions were included in the analysis: *AI literacy*, a*version*, *trust*, and *authenticity assessment*. The calibration process utilized both fuzzy and crisp set methods to reflect the nature of each variable and to capture nuanced differences within the data. *Authenticity assessment* was measured as a binary variable and therefore calibrated as a crisp set, clearly distinguishing whether the tool classified the video as genuine (unaltered; coded as 0) or deepfake (manipulated; coded as 1).

For the conditions assessed as fuzzy sets, established scales were utilized. *AI literacy* was measured using the 16-item scale developed by Weber et al. (2023), which includes questions designed to evaluate both the social and technical dimensions of participants' AI literacy. *Trust* was operationalized using the 25-item scale by Madsen & Gregor (2000), comprising five items for each of the following dimensions: perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment. This scale was selected for its ability to capture both cognitive and affective components of trust, providing a multidimensional perspective that surpasses narrower instruments focused solely on constructs like reliability or competence. It is widely cited in IS, HCI, and technology trust research, and strikes a balance between conceptual comprehensiveness and participant manageability (e.g., Fügener et al., 2021; Vössing et al., 2022). In the case of *aversion*, a new measure was constructed in line with recent literature (Jussupow et al., 2024; Wang et al., 2024), consisting of three items that capture participants' willingness to use a deepfake detection tool, perceived decision-making support, perceived comparative performance. This measure was developed to assess algorithm aversion as a cognitive bias (Dietvorst et al., 2015), distinct from behavioral reliance measures such as weight on advice. While reliance could be inferred from decisions in the binary task, the self-report scale provides insight into

participants' subjective resistance to algorithmic assistance, an essential aspect of algorithm aversion. Responses for *trust* and *aversion* were captured on a 7-point Likert scale ranging from "completely disagree" to "completely agree".

In accordance with best practices in QCA, calibration thresholds were set to align both with theoretical distinctions and the empirical distribution of the data (Mattke et al., 2022). *AI literacy* conceptualized as a continuum reflecting different levels of familiarity with AI where the maximum achievebale score is 12; accordingly, it was calibrated as a fuzzy set with thresholds of 3.5, 8, and 11.5, allowing for the detection of subtle gradations in literacy. *Aversion* was calibrated as a fuzzy set with thresholds at 12, 15, and 21, reflecting the first quartile, median, and maximum values observed in the data, and drawing on established theories of risk and aversion (Kahneman & Tversky, 1979; Wang et al., 2024). Similarly, *trust* was operationalized as an aggregate construct, calculated as the sum of scores across the five dimensions. It was calibrated as a fuzzy set using thresholds of 15, 20, and 25, informed by both the distribution of trust scores and previous research on trust in technological systems (Lee & See, 2004).

Outcome Condition. The outcome variable, *reliance on AI advice*, was operationalized as a crisp set using a threshold of 0.5. A case was coded as "1" (indicating reliance on AI advice) if the participant's final decision changed to match the recommendation provided by the deepfake detection tool. Notably, if a participant's initial guess was correct, no reliance on AI advice was recorded, ensuring that only changes in decision reflecting tool influence were captured.

4 Analysis and Results

Following the guidelines of Mattke et al. (2021), a QCA was performed to examine the combinations of conditions leading to reliance on AI advice. We conducted both an analysis of necessary conditions and an analysis of sufficient configurations in line with established QCA procedures (Mattke et al., 2022). Four causal conditions were included in the analysis: AI literacy, aversion, trust, and authenticity assessment. In theory, these four conditions yield 16 (2^4) possible configurations. With a total of 208 observations (52 participants each rating 4 videos), our sample size is sufficient according to the criteria proposed by Oana et al. (2021).

A truth table (see Table 1) was constructed that displays all possible configurations along with the number of cases per configuration, raw consistency (incl), and the proportional reduction in inconsistency (PRI). In our data, 14 unique configurations were observed in the data. Overall, these configurations cover 87.5% of the possible cases, which is in line with or above the coverage levels typically reported in IS QCA studies (Mattke et al., 2021). The number of cases per configuration ranged from 6 to 32, ensuring a robust empirical basis for the analysis.

Notably, the two missing configurations are characterized by high aversion, low trust, and low AI literacy, regardless of whether authenticity assessment was given as 0 (deepfake) or 1 (genuine). The absence of these configurations suggests that the combination of high aversion with low trust and low AI literacy does not occur in the data,

irrespective of the authenticity assessment. Overall, the distribution of the outcome variable was balanced, with 110 cases characterized by the rejection of AI Advice (Reliance on Advice = 0) and 98 cases exhibiting reliance on AI advice (Reliance on Advice = 1). This balanced distribution enhances the validity of the subsequent analyses.

Table 1. Truth Table of All Possible Configurations

Aversion	Trust	AI Literacy	Auth Assmt	Reliance on Advice	n	incl	PRI
0	0	0	0	0	8	0.000	0.000
0	0	0	1	1	8	1.000	1.000
0	0	1	0	0	16	0.000	0.000
0	0	1	1	1	16	0.938	0.938
0	1	0	0	0	6	0.000	0.000
0	1	0	1	0	6	0.667	0.667
0	1	1	0	0	8	0.000	0.000
0	1	1	1	1	8	1.000	1.000
1	0	0	0	?	0	-	-
1	0	0	1	?	0	-	-
1	0	1	0	0	6	0.000	0.000
1	0	1	1	1	6	1.000	1.000
1	1	0	0	0	32	0.000	0.000
1	1	0	1	1	32	0.875	0.875
1	1	1	0	0	28	0.000	0.000
1	1	1	1	1	28	1.000	1.000

AuthAssmt: Authentiticity Assessment

4.1 Analysis of Necessary Conditions

The analysis of necessary conditions uses the measure of proportional reduction in inconsistency to determine whether a condition is required for advice-taking to occur (Mattke et al., 2022). In our analysis (see Table 2), the condition authenticity assessment achieves a perfect consistency score (inclN = 1.000), a coverage (covN) of 0.933, and a Rate of Necessity (RoN) of 0.937, indicating that advice-taking occurs only when the video is authentic. In contrast, none of the other conditions (or their negations) reach the recommended consistency threshold of 0.90 required to be considered necessary (Ragin, 2009) underscoring the central role of authenticity assessment in this model.

Table 2. Results of Analysis of Necessity for Advice taking (Reliance on AI Advice = high/1) and Advice Rejection (Reliance on AI Advice = low/0)

Conditions	Advice Taking			Advice Rejection		
	inclN	covN	RoN	inclN	covN	RoN
AI Literacy	0.588	0.491	0.609	0.532	0.509	0.617
~AI Literacy	0.412	0.435	0.690	0.468	0.565	0.744
Aversion	0.639	0.470	0.521	0.631	0.530	0.551
~Aversion	0.361	0.461	0.763	0.369	0.539	0.790
Trust	0.701	0.459	0.429	0.721	0.541	0.469
~Trust	0.299	0.483	0.827	0.279	0.517	0.836
AuthAssmt	**1.000**	**0.933**	**0.937**	0.063	0.067	0.517
~AuthAssmt	0.000	0.000	0.500	**0.937**	**1.000**	**1.000**

AuthAssmt: Authentiticity Assessment

In summary, the necessity analysis reveals that authenticity assessment is the only condition that is necessary for reliance on AI advice, whereas the other conditions do not exhibit the necessary consistency levels.

4.2 Analysis of Sufficient Conditions

To identify sufficient configurations for reliance on AI advice, we first constructed a truth table using an inclusion cutoff of 0.8. Next, we applied both complex and parsimonious minimization procedures, which yielded an identical solution, designated as equation 1, with the solution parameters summarized in Table 3.

$$Aversion{*}AuthAssmt + {\sim}Trust{*}AuthAssmt + AI\ Literacy{*}AuthAssmt \rightarrow RelyAdvice \quad (1)$$

Table 3. Solution Parameters of the Analysis of Sufficiency

Configuration	inclS	PRI	covS	covU
Aversion * AuthAssmt	0.939	0.939	0.639	0.289
~Trust * AuthAssmt	0.967	0.967	0.299	0.082
AILiteracy*AuthAssmt	0.983	0.983	0.588	0.082
	0.949	0.949	0.959	

Equation (1) indicates that reliance on AI advice occurs when a video is deemed authentic, combined with either high aversion, low trust, or high AI literacy. The high inclusion scores (inclS) and proportional reduction in inconsistency (PRI) values indicate that these pathways are highly consistent across cases, while the raw coverage (covS) and unique coverage (covU) values demonstrate that these configurations explain a substantial portion of the observed outcome. These findings highlight the centrality of authenticity assessment in following deepfake detection advice, illustrating the QCA principle of equifinality, different condition combinations can yield the same outcome (Mattke et al., 2021).

4.3 Robustness Checks

To evaluate the stability of our QCA results, we conducted a series of robustness checks following the guidelines of Oana & Schneider (2024). We reported the complete truth table (Table 1) as recommended by Mattke et al. (2022) to demonstrate threshold robustness. Additionally, we also tested robustness to the chosen calibration anchors (Oana & Schneider, 2024). In addition, we tested alternative calibration anchors, e.g., adjusting the lower bound for AI literacy from [3.5, 8, 11.5] to [3.0, 8, 11.5], which produced consistent configuration patterns and an unchanged minimized solution.

Moreover, both the complex and parsimonious minimization procedures yielded identical solutions, indicating a stable core unaffected by the treatment of logical remainders. Together, these tests confirm that our results are not sensitive to calibration choices and reflect genuine empirical patterns.

5 Discussion and Implications

The objective of this study was to understand which factors influence participants' reliance on AI advice in the context of deepfake detection. Our findings reveal, that authenticity assessment emerges as necessary condition for reliance on AI advice: participants only accepted the deepfake detection tool's recommendation when it confirmed that a video was genuine. This behavior suggests an asymmetry in how participants interpret the tool's recommendations and highlights the influence of cognitive biases such as confirmation bias. One possible explanation, though not directly tested in this study, is that confirming a video as genuine is perceived as more straightforward or trustworthy than identifying deepfakes. Prior work suggests that verifying authenticity through known references may reinforce confidence (Hameleers et al., 2024). Additionally, psychological factors, such as discomfort with acknowledging manipulation, could make "genuine" judgments more acceptable (Martel et al., 2020). Further research would be needed to investigate these potential mechanisms.

This pattern persisted regardless of variations in trust, AI literacy, or aversion, suggesting that the tool's validation of authenticity is the critical driver of decision revisions. This pattern raises concerns about the unreflective acceptance of recommendations (Inuwa-Dutse et al., 2023), especially considering the current limitations of deepfake detection technology (Heidari et al., 2023). These results suggest that, in our study, participants were more inclined to accept the tool's advice when it confirmed a video as genuine. This raises the possibility that users currently treat such tools more as confirmatory aids for authenticity than as reliable detectors of manipulation. However, this interpretation should be treated with caution, as the underlying reasons for this asymmetry were not directly measured. On the other hand, this can also results from how well partcipants already performed in recognizing deepfakes as the outcome variable, reliance on AI advice, was only coded as a taken advice if the particpants initial decision was wrong (recalling the participants' initial jugement with an overall detection rate of 39.9%). However, if we consider the current technological possibilities in the context of generating and detecting deepfakes it is predicted that these tools will improve in the next few years a lot (Verdoliva, 2020). If those technologies improve it might get nearly

impossible for individuals to detect such deepfake generated content just by examining visual cues. This future increase in detection difficulty could further heighten reliance on algorithmic tools, especially if users develop calibrated reliance over time.

These findings close a notable theoretical gap. While previous research has predominantly focused on the technological performance of deepfake detection tools (Altuncu et al., 2024; Heidari et al., 2023) or on enhancing media literacy (Hoes et al., 2024), few studies have examined how cognitive and social processes, such as trust calibration and error sensitivity, interact with reliance on AI advice in high-stakes contexts like deepfake detection (Dietvorst et al., 2015; Wischnewski et al., 2023). Our results suggest that authenticity assessment is central: participants are unlikely to revise their decisions unless the tool confirms genuineness. This challenges current models of algorithm aversion, which predict general resistance to algorithmic input after perceived risk or error. Instead, it shows that aversion may be selectively overridden when the recommendation supports an existing belief.

Looking at the found sufficient configurations that result in reliance on AI advice it can be noted that authenticity assement in combination with either high aversion, low trust, or high AI literacy led to participants accepting the deepfake detection tools advice. Contradicting in that sense is that even if participants had a high aversion they accepted the tool's advice if their initial guess was deepfake and it suggested genuine. One possible interpretation of this behavior is that participants experienced cognitive dissonance when the tool contradicted their initial judgment. In such cases, they may have adjusted their opinion to reduce the psychological discomfort of holding conflicting beliefs (McGrath, 2017). Cognitive dissonance thus might provide a complementary theoretical lens to algorithm aversion: the discomfort of being wrong may override a general bias against algorithms, particularly when switching one's decision restores internal consistency. This opens up the question whether the reliance on AI advice under different configurations might change when technology evolves and will become then more important. Additionally, participants with low trust still accepted the tool's advice if it suggested to switch to genuine. This is in line with the above mentioned reasons for aversion if considering trust and aversion as closely linked concepts (Jussupow et al., 2024). Finally, partcipants with high AI literacy also accepted the tools advice to categorize the watched video as genuine. Surprisingly, participants with high AI literacy were still likely to rely on the tool when it confirmed authenticity. While AI literacy is often associated with greater awareness of system limitations (Pinski & Benlian, 2024), it may also lead to greater confidence in interpreting algorithmic outputs. This suggests that high AI literacy does not automatically lead to greater skepticism but may, under certain conditions, increase reliance on algorithmic advice, particularly when the recommendation aligns with expectations. Given this knowledge in the context of AI-generated content detection and the current technological possibilities for deepfake detection it is contradicting that they followed advice (Leffrang & Mueller, 2024), however this also only holds for the assessment of the video being genuine.

For practitioners, the findings underscore the necessity for deepfake detection systems to promote not only accurate recommendations but also reflective user engagement. Designers should consider incorporating transparency features that help users understand why a recommendation is given, thereby preventing blind acceptance. As

detection technologies continue to evolve, fostering a balance between trust and healthy skepticism will be crucial to avoid overreliance on automated advice, especially when the cost of erroneous decisions is high (Verdoliva, 2020).

5.1 Limitations and Future Work

While this study contributes novel insights into reliance on AI advice in the context of deepfake detection, several limitations should be acknowledged.

The sample primarily consisted of university students, which is suitable for initial exploration of human-AI interaction due to their higher digital and media literacy. However, this may not represent broader demographics, as reliance on AI advice could vary in older or less tech-savvy populations. Future studies should replicate these findings in more diverse samples to evaluate the generalizability of the observed configurations.

Second, the deepfake detection tool used in the experiment was presented as having 100% accuracy, though participants were unaware of this. This allowed for an isolated examination of user trust and aversion without confounding effects from performance variability. However, real-world systems are rarely perfect. Future research should test how reliance on AI advices change under realistic error rates, where participants are aware of possible system fallibility.

Third, reliance on AI advice was operationalized as a binary outcome, whether a participant changed their initial judgment to match the tool's recommendation. While this allowed for a clear QCA-based configuration analysis, it may oversimplify the cognitive processes behind such decisions. Participants may revise their responses due to uncertainty, cognitive dissonance, or heuristic reasoning rather than trust or aversion alone. Including measures such as confidence ratings or open-ended justifications could help capture these subtleties in future work.

Together, these limitations suggest valuable avenues for future investigation. Rather than detracting from the study's contribution, they underscore the complexity of reliance on AI advice in algorithmic contexts, especially where misinformation and visual deception are involved, and point to the importance of more nuanced, real-world approaches to studying human-AI collaboration.

6 Conclusion

This study highlights that authenticity assessment is crucial for reliance on AI advice in deepfake detection. Participants adjusted their judgments only when the tool indicated a video was genuine, suggesting a tendency for "blind following" of algorithmic recommendations. This raises concerns about the potential for incorrect judgments influenced by the tool. Overall, our research offers insights into the factors influencing reliance on AI advice in deepfake detection, emphasizing the importance of authenticity assessment alongside trust, aversion, and AI literacy, which presents both opportunities and risks for automated decision support.

References

Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I. & Mishkhal, I. A. (2023) Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics.* 23, 429-441.

Alexander, V., Blinder, C. & Zak, P. J. (2018) Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior.* 89, 279-288.

Altuncu, E., Franqueira, V. N. L. & Li, S. (2024) Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Front Big Data.* 7, 1400024. doi: 10.3389/fdata.2024.1400024.

Bauer, K. & Gill, A. (2024) Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. *Information systems research.* 35 (1), 226-248. doi: 10.1287/isre.2023.1217.

Bonezzi, A., Ostinelli, M. & Melzner, J. (2022) The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General.* 151 (9), 2250.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E. & Watts, D. J. (2024) Misunderstanding the harms of online misinformation. *Nature.* 630 (8015), 45-53. doi: 10.1038/s41586-024-07417-w.

Cadario, R., Longoni, C. & Morewedge, C. K. (2021) Understanding, explaining, and utilizing medical artificial intelligence. *Nature human behaviour.* 5 (12), 1636-1642.

Chen, S., Xiao, L. & Kumar, A. (2023) Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior.* 141, 107643. doi: https://doi.org/10.1016/j.chb.2022.107643.

Dargnies, M.-P., Hakimov, R. & Kübler, D. (2024) Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science.*

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General.* 144 (1), 114.

Fagni, T., Falchi, F., Gambini, M., Martella, A. & Tesconi, M. (2021) TweepFake: About detecting deepfake tweets. *PLOS ONE.* 16 (5), e0251415. doi: 10.1371/journal.pone.0251415.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M. & Pröllochs, N. (2023) Research can help to tackle AI-generated disinformation. *Nature human behaviour.* 7 (11), 1818-1821. doi: 10.1038/s41562-023-01726-2.

Fügener, A., Grahl, J., Gupta, A. & Ketter, W. (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS quarterly.* 45.

Gowrisankar, B. & Thing, V. L. (2024) An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *Computers & Security.* 139, 103684.

Greenspan, R. L. & Loftus, E. F. (2022) What happens after debriefing? The effectiveness and benefits of postexperimental debriefing. *Memory & cognition.* 50 (4), 696-709.

Hameleers, M., van der Meer, T. G. L. A. & Dobber, T. (2024) They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes. *European Journal of Communication.* 39 (1), 56-70. doi: 10.1177/02673231231184703.

Heidari, A., Jafari Navimipour, N., Dag, H. & Unal, M. (2023) Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery.* 14 (2). doi: 10.1002/widm.1520.

Hessler, P. O., Pfeiffer, J. & Hafenbrädl, S. (2022) When Self-Humanization Leads to Algorithm Aversion What Users Want from Decision Support Systems on Prosocial Microlending

Platforms. *Business & Information Systems Engineering*. 64 (3), 275-292. doi: 10.1007/s12599-022-00754-y.

Hoes, E., Aitken, B., Zhang, J., Gackowski, T. & Wojcieszak, M. (2024) Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature human behaviour*. 8 (8), 1545-1553. doi: 10.1038/s41562-024-01884-x.

Inuwa-Dutse, I., Toniolo, A., Weller, A. & Bhatt, U. (2023) Algorithmic loafing and mitigation strategies in Human-AI teams. *Computers in Human Behavior: Artificial Humans*. 1 (2), 100024.

Jussupow, E., Benbasat, I. & Heinzl, A. (2024) An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making. *MIS quarterly*. 48 (4), 1575-1590. doi: https://doi.org/10.25300/MISQ/2024/18512.

Jussupow, E., Spohrer, K., Heinzl, A. & Gawlitza, J. (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information systems research*. 32 (3), 713-735.

Kahneman, D. & Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica*. 47 (2), 363-391.

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S. & Xia, F. (2024) Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*. 57 (6). doi: 10.1007/s10462-024-10810-6.

Khan, A., Brohman, K. & Addas, S. (2022) The anatomy of "fake news": Studying false messages as digital objects. *Journal of Information Technology*. 37 (2), 122–143. doi: 10.1177/02683962211037693.

Kießling, S., Figl, K. & Remus, U. (2021) Human Experts or Artificial Intelligence? Algorithm Aversion in Fake News Detection. In: *European Conference on Information Systems, A Virtual AIS Conference*.

Kietzmann, J., Lee, L. W., McCarthy, I. P. & Kietzmann, T. C. (2020) Deepfakes: Trick or treat? *Business Horizons*. 63 (2), 135-146. doi: 10.1016/j.bushor.2019.11.006.

Kim, T. & Song, H. (2023) Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International Journal of Human–Computer Interaction*. 39 (4), 790-800. doi: 10.1080/10447318.2022.2049134.

Klingbeil, A., Gruetzner, C. & Schreck, P. (2024) Trust and reliance on AI - An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*. 160. doi: 10.1016/j.chb.2024.108352.

Köbis, N. C., Doležalová, B. & Soraperra, I. (2021) Fooled twice: People cannot detect deepfakes but think they can. *iScience*. 24 (11), 103364. doi: https://doi.org/10.1016/j.isci.2021.103364.

Kumari, R., Ashok, N., Ghosal, T. & Ekbal, A. (2022) What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. *Information Processing & Management*. 59 (1), 102740. doi: https://doi.org/10.1016/j.ipm.2021.102740.

Lee, J.-N., Park, Y., Straub, D. W. & Koo, Y. (2019) Holistic Archetypes of IT Outsourcing Strategy: A Contingency Fit and Configurational Approach. *MIS quarterly*. 43 (4), 1201-1225. doi: 10.25300/misq/2019/14370.

Lee, J. D. & See, K. A. (2004) Trust in automation: Designing for appropriate reliance. *Human factors*. 46 (1), 50-80.

Leffrang, D. & Mueller, O. (2024) Algorithmic Advice-Taking Beyond MAE: The Role of Negative Prediction Outliers and Statistical Literacy in Algorithmic Advice-Taking. In: *European Conference on Information Systems*.

Lu, T. & Zhang, Y. (2024) 1 + 1 > 2? Information, Humans, and Machines. *Information systems research*. 0 (0), null. doi: 10.1287/isre.2023.0305.

Madsen, M. & Gregor, S. (2000) *Measuring human-computer trust, 11th australasian conference on information systems*. Citeseer.

Martel, C., Pennycook, G. & Rand, D. G. (2020) Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications.* 5 (1), 47. doi: 10.1186/s41235-020-00252-3.

Mattke, J., Maier, C., Weitzel, T., Gerow, J. E. & Thatcher, J. B. (2022) Qualitative Comparative Analysis (QCA) In Information Systems Research: Status Quo, Guidelines, and Future Directions. *Commun. Assoc. Inf. Syst.* 50, 8.

Mattke, J., Maier, C., Weitzel, T. & Thatcher, J. B. (2021) Qualitative comparative analysis in the information systems discipline: a literature review and methodological recommendations. *Internet Research.* 31 (5), 1493-1517. doi: 10.1108/intr-09-2020-0529.

McGrath, A. (2017) Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass.* 11 (12), e12362.

Mueller, R. O. & Hancock, G. R. (2018) Structural equation modeling. *The reviewer's guide to quantitative methods in the social sciences.* Routledge, pp. 445-456.

Oana, I.-E. & Schneider, C. Q. (2024) A robustness test protocol for applied QCA: theory and R software application. *Sociological Methods & Research.* 53 (1), 57-88.

Oana, I.-E., Schneider, C. Q. & Thomann, E. (2021) *Qualitative comparative analysis using R: A beginner's guide.* Cambridge University Press.

Ochmann, J., Michels, L., Zilker, S., Tiefenbeck, V. & Laumer, S. (2020) The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations. In: *International Conference on Information Systems, India*.

Pennycook, G. & Rand, D. G. (2019) *Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality*, *Proceedings of the National Academy of Sciences*.

Pinski, M. & Benlian, A. (2024) AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans.* 2 (1), 100062. doi: https://doi.org/10.1016/j.chbah.2024.100062.

Ragin, C. C. (2009) Qualitative comparative analysis using fuzzy sets (fsQCA). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques.* 87-122.

Rana, M. S., Nobi, M. N., Murali, B. & Sung, A. H. (2022) Deepfake Detection: A Systematic Literature Review. *IEEE Access.* 10, 25494-25513. doi: 10.1109/access.2022.3154404.

Reich, T., Kaju, A. & Maglio, S. J. (2023) How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology.* 33 (2), 285-302.

Renier, L. A., Schmid Mast, M. & Bekbergenova, A. (2021) To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*. 124, 106879. doi: https://doi.org/10.1016/j.chb.2021.106879.

Rochlin, N. (2017) Fake news: belief in post-truth. *Library Hi Tech.* 35 (3), 386-392. doi: 10.1108/LHT-03-2017-0062.

Saragih, M. & Morrison, B. W. (2022) The Effect of past Algorithmic Performance and Decision Significance on Algorithmic Advice Acceptance. *International Journal of Human–Computer Interaction.* 38 (13), 1228-1237. doi: 10.1080/10447318.2021.1990518.

Schneider, C. Q. & Wagemann, C. (2012) *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis.* Cambridge University Press.

Sniezek, J. & Buckley, T. (1989) *Social influence in the advisor-judge relationship, Annual meeting of the judgment and decision making society, Atlanta, Georgia*.

Telegraph, T. (2022) *Deepfake video shows Volodymyr Zelensky telling Ukrainians to surrender.*

Vaccari, C. & Chadwick, A. (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society.* 6 (1), 2056305120903408.

Vasist, P. N. & Krishnan, S. (2022) Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Communications of the Association for Information Systems.* 51, 590-636. doi: 10.17705/1cais.05126.

Verdoliva, L. (2020) Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing.* 14 (5), 910-932.

Vössing, M., Kühl, N., Lind, M. & Satzger, G. (2022) Designing transparency for effective human-AI collaboration. *Information systems frontiers.* 24 (3), 877-895.

Wang, L., Li, X., Zhu, H. & Zhao, Y. (2024) A dimensional exploration and scale development study of algorithm aversion. *Journal of the Operational Research Society.* 1-22. doi: 10.1080/01605682.2024.2419544.

Wanner, J., Herm, L.-V., Heinrich, K. & Janiesch, C. (2022) The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets.* 32 (4), 2079-2102. doi: 10.1007/s12525-022-00593-5.

Weber, P., Pinski, M. & Baum, L. (2023) Toward an objective measurement of AI literacy.

Wischnewski, M., Krämer, N. & Müller, E. (2023) *Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions*, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

You, S., Yang, C. L. & Li, X. (2022) Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems.* 39 (2), 336-365. doi: 10.1080/07421222.2022.2063553.