

# Through the Looking-Glass: AI-Mediated Video Communication Reduces Interpersonal Trust and Confidence in Judgments

Nelson Navajas Fernández\*  
Bauhaus University  
Weimar, Germany

Jeffrey T. Hancock  
Stanford University  
Stanford, United States

Maurice Jakesch  
Bauhaus University  
Weimar, Germany

## Abstract

AI-based tools that mediate, enhance or generate parts of video communication may interfere with how people evaluate trustworthiness and credibility. In two preregistered online experiments (N = 2,000), we examined whether AI-mediated video retouching, background replacement and avatars affect interpersonal trust, people's ability to detect lies and confidence in their judgments. Participants watched short videos of speakers making truthful or deceptive statements across three conditions with varying levels of AI mediation. We observed that perceived trust and confidence in judgments declined in AI-mediated videos, particularly in settings in which some participants used avatars while others did not. However, participants' actual judgment accuracy remained unchanged, and they were no more inclined to suspect those using AI tools of lying. Our findings provide evidence against concerns that AI mediation undermines people's ability to distinguish truth from lies, and against cue-based accounts of lie detection more generally. They highlight the importance of trustworthy AI mediation tools in contexts where not only truth, but also trust and confidence matter.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing; Interaction design theory, concepts and paradigms**; • **Computing methodologies** → **Artificial intelligence**.

## Keywords

AI-mediated communication, video filters, deception detection, trust, credibility, avatars, experiments

## ACM Reference Format:

Nelson Navajas Fernández, Jeffrey T. Hancock, and Maurice Jakesch. 2026. Through the Looking-Glass: AI-Mediated Video Communication Reduces Interpersonal Trust and Confidence in Judgments. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3772318.3790845>

## 1 Introduction

In 2021, a lawyer joined a virtual court hearing in Texas, to everyone's surprise, appearing as a wide-eyed kitten [77]. "I'm here live.

\*Corresponding author e-mail: [nelson.navajas.fernandez@uni-weimar.de](mailto:nelson.navajas.fernandez@uni-weimar.de)



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/2026/04

<https://doi.org/10.1145/3772318.3790845>

"I'm not a cat" he clarified and excused himself with a kitty-worried expression, blinking at the judge. When such mishaps make the news, these incidents demonstrate how deeply AI communication systems can disrupt the expectations and assumptions people hold about mediated communication, particularly in high-stakes contexts where what is said has far-reaching consequences.

However, most AI-based transformations we use in our communication today are less obvious than a cat avatar. With the shift to remote communication accelerated by the COVID-19 pandemic [97, 102], video communication platforms are increasingly used in not only casual but also professional and high-stakes settings [19, 51, 70, 79]. Platforms such as Zoom, Google Meet and Microsoft Teams, integrate a wide range of algorithmic video enhancements and transformations. The available features range from background blurring and replacement to skin improvements, gaze correction and personalized avatars that resemble the speaker.

AI-based video features are widely used and broadly regarded as acceptable [20, 54, 112]. While often presented as convenience features or aesthetic improvements [112], they may alter aspects of communication that are central to impression formation and judgment [37, 46, 53]. As AI-mediated video tools become widely deployed, we need to better understand their potential to shape people's impressions and judgments [35, 37] as well as trust [53, 69], honesty [46, 60, 99] and credibility [59] in online video communication.

Indeed, previous research in computer-mediated communication (CMC) shows that the medium of interaction can significantly shape how we present ourselves and how others perceive us [39]. In video communication, many cues that people rely on in face-to-face settings, such as posture, eye contact and microexpressions [13], are missing or obscured. Even in traditional computer-mediated communication, judging the honesty, credibility and trustworthiness of others is a difficult task [8, 18, 62]. Because such evaluations shape how people interpret and respond to others in everyday communication [62, 103], preserving people's ability to evaluate others in AI-mediated communication remains essential.

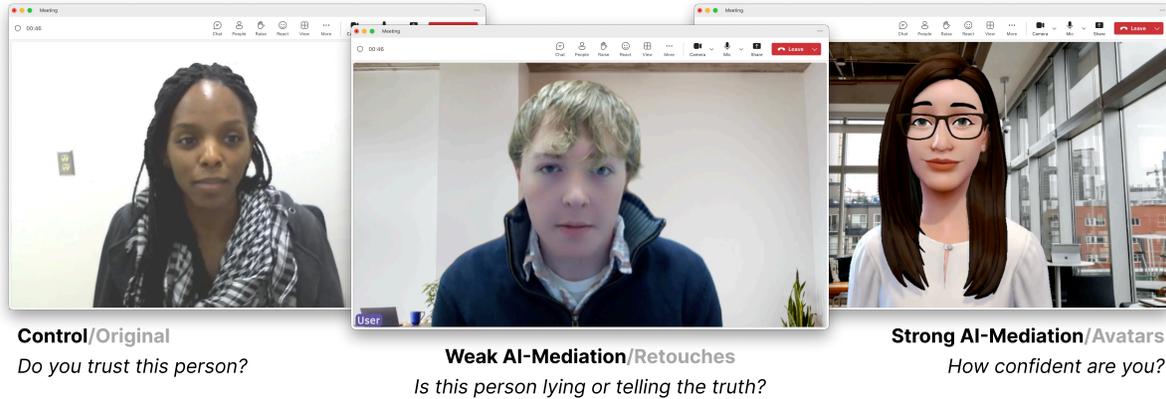
The introduction of AI-based video tools may affect the ways people form impressions of each other and assess credibility, authenticity and trustworthiness [46, 53, 59]. Hancock et al. conceptualized the relevant changes under the framework of AI-mediated communication (AI-MC) [37]: a paradigm shift in computer-mediated communication where a computational agent modifies, augments or generates message content on behalf of a communicator. Prior work in text-based contexts has shown that a mediating AI system that modifies or creates communication can erode interpersonal trust [53] and alter how receivers interpret intentions, credibility and agency [30, 74, 89]. In video communication—the context of the current study—interactions are more dynamic and perceptually

**Study 1: Homogenous environment**

Participants evaluated six videos of the same type:

**Study 2: Mixed environment**

Participants evaluated two videos of each type:



**Figure 1: Study conditions and primary outcome variables.** Participants watched six videos in which video subjects recounted a story about someone they knew, that was either true or false. In the control condition, we embedded the original video in a video call to increase realism. In the weak AI-mediated treatment, we further processed the video using retouching and a virtual background. In the strong AI-mediated treatment, we replaced the subject with an animated avatar. For each video, participants indicated whether they trusted the person in the video, whether they thought the person in the video was lying, and how confident they were in their judgment.

rich than text, further complicating assessments of how the AI tools integrated into widely used platforms [112] may affect judgments of trust, honesty and credibility.

This study investigates how different levels of AI-mediated video communication—ranging from original recordings to weak AI mediation through retouching and virtual backgrounds to strong AI mediation through animated avatars—affect the perceived trustworthiness of the speaker, judgments of truth and confidence in these judgments. In two large online experiments ( $N = 2,000$ ), participants viewed prerecorded videos of others making truthful or deceptive statements. As illustrated in Figure 1, we processed the video stimuli through the integrated AI video features of Microsoft Teams to reflect different degrees of AI mediation: in the (1) control condition, the videos were unaltered, corresponding to regular computer-mediated communication; in the (2) weak AI mediation condition, we enabled skin smoothing, lighting adjustment and virtual backgrounds, corresponding to widely used video transformations; in the (3) strong AI mediation condition, we transformed the speaker into a fully-animated character (avatar) to test the effect of strong AI-based video transformation. In addition to a uniform communication setting in which participants rated six videos of the same mediation type in Study 1, participants in Study 2 encountered different types of AI mediation in a mixed environment, more closely resembling real-world interactions. After watching each video, we asked participants how much they trusted the person and whether the person was telling the truth. They also rated their confidence in their judgments and answered follow-up questions about the cues they relied on in their judgments.

Across both studies, we found that AI-mediated video did not affect deception detection accuracy or participants' overall likelihood of suspecting the other person of lying. However, it consistently

reduced perceived trustworthiness and lowered participants' confidence in their judgments, particularly in the mixed environment (Study 2). The results suggest that AI-mediated video processing meaningfully affects how people evaluate others in online video communication, even when it does not alter their ability to distinguish truths from lies.

Our findings have implications for platform design and for policy debates, as they show how AI mediation shapes people's evaluations of others in online interactions. Even common tools such as avatars and retouching influence how people form trust, credibility and confidence in online interactions, underscoring the need for greater attention to representational consistency, transparency and the context-sensitive use of mediation features.

## 2 Related work

Our work is motivated by the growing integration of AI-mediated communication tools, such as video retouching and avatars, into everyday video communication [78, 109]. Prior work in computer-mediated communication [9, 43] has examined how reduced cues shape trust and impression formation, and deception research has documented the limits of people's ability to detect lies [8] as well as truth-default rates [62]. We draw on this literature and combine it with recent work on AI-mediated communication [37, 46, 53], HCI studies of avatar-mediated communication, and theories of interpersonal judgment, to investigate how AI-mediated video processing affects deception detection, confidence and interpersonal trust.

### 2.1 Deception Detection

Accurately judging whether someone is being honest is essential in mediated interactions [12], where people must assess the reliability

of information provided by others, such as in hiring conversations, interviews, collaborative online work and educational settings. In many contexts, judgments about honesty shape how people interpret, trust and respond to what is communicated [8, 62]. At the same time, decades of research show that deception detection is difficult [15, 104]: people are only slightly better than chance (54% on average) [8, 18] at discerning lies from truth. People also have a strong general tendency to believe what others are saying, known as the "veracity effect" [62, 63].

Previous research has proposed two broad perspectives to explain why deception is complex: On the one hand, cue-based approaches posit that liars reveal themselves through nonverbal cues, so-called "leakage", such as gaze aversion or microexpressions, and that one can discern lies from truth by observing those nonverbal cues [23–25]. However, meta-analyses show that deception cues are inconsistent across studies and are therefore weak and unreliable indicators of deception [18]. Nonetheless, people still rely on them to form their deception judgments [18]. In contrast, context-based perspectives, such as Levine's truth-default theory (TDT), argue that people rely primarily on the plausibility and coherence of what is said and default to believing others unless suspicion is actively triggered, which explains the "veracity effect" [62]. Cue-based theories and Levine's truth-default theory yield different expectations regarding how the disruption or removal of nonverbal cues in AI-mediated communication might influence deception judgments. If deception cues are central to lie detection, AI systems that modify them could alter truth judgment rate or accuracy. In contrast, if judgments are driven mainly by plausibility and coherence, the AI-based disruption or removal of other elements of the communication may have little to no effect.

Our work extends prior research on deception detection and AI-mediated communication by introducing AI-mediated videos into established deception stimuli [66]. Despite extensive research on deception detection, no work has examined how everyday AI video tools, such as retouching or avatars, affect our judgments of deception. It also remains unclear how these tools influence our confidence in those judgments and our perceptions of trustworthiness. Particularly, if AI mediation removes or distorts the nonverbal cues that cue-based theories consider central for deception detection, AI-mediated video could make deceptive statements harder to identify. By systematically comparing judgments of videos with varying degrees of AI-mediated content, we provide empirical evidence that AI mediation in video may shape accuracy and interpersonal dynamics in online video communication.

## 2.2 Trust in Mediated Communication

Trust and belief are not the same psychological processes [47]: While belief is a cognitive judgment about whether a claim is true, trust is a relational, affective stance toward a speaker [40, 47], which is a fundamental precondition for effective human cooperation and interaction [55, 72]. High levels of trust enable conflict resolution, problem solving, and fluency in interaction [22, 98, 110], while low trust undermines learning and collaboration [22, 58]. However, in computer-mediated communication, establishing trust is more difficult than in face-to-face settings [44, 92, 93]. Research shows that trust tends to start at a lower baseline in computer-mediated

settings and is harder to establish when nonverbal cues are missing [9, 108]. Computer-mediated communication enables people to carefully control how they present themselves, and the related reduction in and manipulation of social cues complicates how trust and credibility are assessed [26–29, 36, 38, 43].

The COVID-19 pandemic accelerated the adoption of video communication tools for both casual [42, 54, 101] and high-stakes settings, including hiring [73, 76], telemedicine [7, 67, 70, 96], online exams [107] and legal proceedings [90, 101]. In high-stakes environments, not only does the accuracy of deception judgments matter, but also how credibility and trust are experienced and processed, so understanding how AI-mediated video alters these socio-psychological processes is essential [37].

The integration of artificial intelligence into computer-mediated communication as an additional layer of mediating technologies further increases the sender's control over how they present themselves while potentially complicating judgments on the receiver's side [37]. Previous work on AI-mediated communication [37] has shown that, in written contexts, algorithmic modifications complicate how receivers evaluate authenticity and can erode trust and credibility [37, 45, 46, 52, 53, 106]. This effect is particularly strong in settings where there is a mix of human and AI-generated content, where people start to second-guess others' authenticity, a behavior termed the Replicant Effect [53]. We currently do not understand the extent to which these effects apply to the more dynamic and complex medium of video communication.

Research on AI-mediated video so far has focused on the comparatively extreme cases of generated deceptive video such as deepfakes [2, 34, 35, 49]. Here, AI-generated videos are becoming highly realistic and challenging to distinguish from real footage, so people struggle to tell them apart, which leads to increased uncertainty and reduced trust [49, 88, 100, 107]. While deepfakes highlight the risks of realistic synthetic video [35, 88], much less is known about the impacts of commercial everyday forms of AI mediation in video communication, such as retouching, background replacements or synthetic avatars [10], offered through widely used video communication platforms like Zoom, Google Meet and Microsoft Teams.

## 2.3 Theoretical Mechanisms: Expectancy Violations and Uncertainty Reduction

We draw on two theoretical perspectives to contextualize a possible decrease in trust in AI-mediated communication. Expectancy Violations Theory (EVT) [11] argues that people have internalized assumptions about how others should look and behave in social interaction. When visual or behavioral cues deviate from these internalized expectations in mediated communication—such as when facial features are smoothed, backgrounds are replaced or a speaker is replaced by an animated avatar—people may perceive the interaction as less natural or less aligned with normative social scripts, which can reduce perceived trustworthiness [31, 46, 48, 64, 91]. The decrease in trust may be intensified in settings where some people use AI tools and others do not, as differences between mediated and unmediated representations become more salient and thus more likely to violate expectations [11, 83].

Uncertainty Reduction Theory (URT) offers a complementary perspective on how AI-mediated video may influence confidence

in social judgments. Uncertainty Reduction Theory poses that uncertainty in interpersonal encounters is uncomfortable, and that people are motivated to gather information about others to reduce uncertainty and predict others' behavior, attitudes, and intentions [5, 6]. The motivation to reduce uncertainty is heightened in ambiguous interactions, in which people lack access to the full range of interpersonal cues they typically rely on to minimize uncertainty about another person. In mediated communication, viewers rely on visible signals—such as facial expression, gaze direction, and the timing of responses to reduce ambiguity about a speaker's attitudes or intentions [3, 5, 6, 85]. When AI-mediated video alters, smooths or obscures visual cues that could reduce uncertainty in the interaction—e.g., by removing micro-expressions, modifying gaze or reducing facial detail—people might have less information with which to form impressions, increasing uncertainty and lowering the confidence in their own judgments.

## 2.4 Avatar-Mediated Communication in HCI

Avatars are digital representations of users that may be abstract, cartoonish, or human-like. Avatars can be used in online video communication when users prefer privacy, cannot use a camera (e.g., bandwidth or multitasking), or want more control over their visual presentation [78]. The growing integration of avatars into everyday platforms has motivated HCI researchers to examine how such representations affect presence, expressiveness, and interpersonal evaluation in online meetings [83]. A central finding is that high-fidelity avatars are generally more trusted, that is, those avatars that better reflect the real person and their real movements are seen as more trustworthy [83]. Ma et al. [68] find that a critical factor for meeting outcomes is avatar motion fidelity rather than mere visual realism. Webcam-driven head and facial movement supports comfort, emotional clarity, and smooth interaction, while static or synthetic animations make avatars harder to read and increase cognitive load [68]. In mixed environments particularly, where some people appear via video and others via avatar, the synthetic nature of the avatars becomes more salient [56, 83] and people may begin to distrust the avatars [83]. Studies have also shown that in professional settings, avatars can interfere with expectations about workplace appropriateness [50] and that low-fidelity avatars often fall short in conveying facial reactions, gaze and turn-taking cues, leading to lower ratings of professionalism [56].

Research on HCI and avatar-mediated communication highlights that differences in visual fidelity, motion fidelity, and access to non-verbal cues shape how people evaluate others in remote meetings, influencing perceptions of professionalism [50], comfort [68], and trust [83]. However, how such transformations affect judgments of truth or credibility remains an open question. Moreover, the current HCI literature has focused primarily on highly stylized avatar representations [68, 83] and has not examined how less sophisticated forms of AI mediation, such as virtual backgrounds, lighting adjustments, or skin retouching, affect people's judgments.

## 3 Methods

To investigate how AI-mediated video, such as retouching, virtual backgrounds or synthetic avatars, affects people's judgments of

truthfulness and trust in online video communication, we conducted an experiment simulating an online videoconferencing environment. This section provides details on our experimental design, stimuli, procedure, measurements and recruitment.

### 3.1 Hypotheses and Study Design

Our study design is guided by the larger research question: "Does AI-mediated video communication disrupt or interfere with deception judgments and interpersonal trust in online communication?" Based on prior research on deception detection and AI-mediated and avatar-mediated communication, we formulated five hypotheses about how mediating AI systems might affect people's ability to detect lies, confidence in their judgments and trust in others. For each hypothesis, we outline the relevant theoretical mechanisms that motivate the predictions and their direction.

Trust is a central component of interpersonal evaluation in mediated communication [44, 92, 93]. Here, prior CMC research shows that the reductions or distortions of interpersonal cues can decrease perceived trustworthiness [9, 108]. Expectancy Violations Theory (EVT) [11] suggests that visual changes such as avatars or retouching filters may reduce trust when they deviate from what people anticipate as "normal" in a video-call setting. At the same time, research on avatar-mediated communication shows that some mediated representations—particularly those perceived as appropriate or expressive—can be evaluated positively [50]. As prior work indicates that trust could increase or decrease depending on the nature and quality of AI mediation, we formulate the following non-directional hypothesis:

**H1 Interpersonal Trust:** AI-mediated video affects the perceived trustworthiness of the speaker.

Truth judgment rate refer to how often viewers judge statements as true. They are a central outcome in deception detection research [8]. Cue-based perspectives propose that the cognitive effort of lying produces micro-expressions that the receiver can observe to detect deception [23]. If AI mediation reduces access to nonverbal cues, for example, by replacing the speaker with a synthetic avatar, viewers may judge statements as more truthful because they fail to detect cues of deception. In contrast, Levine's truth-default theory argues that people generally default to judging statements as true unless suspicion is triggered [62]. AI mediation could increase uncertainty or suspicion and result in breaking out of that truth-default state, which would reduce truth judgment rate; that is, people would believe the AI-mediated speaker less often, so we hypothesize that:

**H2 Truth judgment rate:** AI-mediated video affects the rate at which participants judge what is said as true.

In deception detection research, humans achieve only slightly better-than-chance accuracy (54%) at correctly classifying veracity judgments [8]. Similar to truth judgment rate (H2), cue-based theories might predict that removing or disrupting visual cues used to detect deception could impair people's ability to make accurate truth-lie judgments, thereby reducing accuracy. In contrast, Levine's truth-default theory predicts that accuracy could improve when people rely less on nonverbal cues and more on message content [62]. Reducing visible cues under AI mediation could strengthen the intuition to shift to content-based information rather than visual cues, thereby increasing accuracy. Furthermore, meta-analyses

show that accuracy often remains unchanged regardless of viewing conditions [8, 18], which would predict a flat performance in accuracy across all levels of AI mediation. We hypothesize:

**H3 Judgment accuracy:** AI-mediated video affects the rate at which participants judge truths as truths (truth accuracy, true positives) and lies as lies (lie accuracy, true negatives).

Confidence in judgments is a subjective metric that captures how sure people feel about their judgments [16]. Uncertainty Reduction Theory (URT) [5, 6] suggests that when familiar interpersonal cues (e.g., facial expressions) are reduced or altered, uncertainty increases, which may lower people’s confidence in their own judgments in AI-mediated interactions. At the same time, prior avatar and mediated-communication studies suggest that when representations are appropriate or easy to interpret, people may feel more certain in their evaluations [68, 83]. As prior lines of work predict both decreases and increases in confidence, we propose the following non-directional hypothesis:

**H4 Judgment confidence:** AI-mediated video affects participants’ confidence in their deception judgments.

In real-world interactions, people may often encounter a mix of original video communications interspersed with retouch effects, virtual backgrounds and avatars. Such mixed environments may increase the salience of the AI mediation and may affect how participants react to the use of AI [53]. Prior HCI and avatar-mediation work shows that differences in environment influence how people evaluate one another [83]. From a theoretical perspective, Expectancy Violations Theory (EVT) [11] suggests that a mixed environment where people communicate with different levels of AI mediation alongside original video representations may reinforce both expectation violations and a sense of uncertainty. We hypothesize:

**H5 Interaction with type of environment:** The impact of AI-mediated video on accuracy, trust and confidence is stronger in settings where people see a mix of different types of AI mediation compared to settings where everyone uses the same AI tools.

To test the above hypotheses empirically, we designed a study consisting of two experiments: a between-subjects experiment (Study 1) to test H1 to H4 in an environment where participants encounter a single type of AI mediation only; and a within-subjects experiment (Study 2), where participants encounter different types of AI mediation to test H5 in addition to H1 to H4. We preregistered the hypotheses together with the study design and analysis plan before data collection <sup>1</sup>.

### 3.2 Stimuli and Experimental Treatments

We structured the experiment as two complementary studies conducted concurrently. In the experiments, participants evaluated six videos in a videoconferencing platform setting. We processed the videos to different levels of AI-based transformations, with video subjects telling true or fabricated stories about another person. We asked participants to judge whether video subjects were telling the truth or lying.

<sup>1</sup>AI Video Filters and Deception Judgments – Study 1 & 2 Preregistration (#239571), submitted 2025-07-23 on AsPredicted.

We considered several video deception datasets for studying deception across different lie stakes. The Miami University Deception Detection Database (MU3D) [66] provides truthful and deceptive videos. The Bag-of-Lies dataset [33] includes multimodal signals such as video, audio, and biometrics in low to medium-stakes laboratory settings. DOLOS [32] presents medium-stakes deception from incentivized game-show interactions with richly annotated audiovisual data. Other high-stakes datasets include courtroom trial recordings [87] and political deception videos [105], where consequences add complexity.

For our present work, we used the Miami University Deception Detection Database [66], which contains webcam-recorded videos of 80 subjects, equally divided by race (black/white) and gender (male/female). Each subject is featured in four videos, in which they either make a truthful or a deceptive statement about their social relationships, under a positive or negative valence. The dataset captures unscripted, conversational speech with direct camera eye contact and natural behavior, closely mimicking the dynamics of online video communication platforms and aligning well with the purpose of the current study. As the valence dimension is irrelevant to the current study and would have introduced unnecessary variation, we focused on positive-valence videos only, yielding a 160-video base set comprising 80 lies and 80 truths.

We embedded these videos into a video communication platform, Microsoft Teams, and further processed them to reflect different levels of AI mediation (see Figure 1) in addition to the control condition:

**T1 Control condition:** In the control condition, participants saw the original, unaltered recording embedded in a Microsoft Teams interface.

**T2 Weak AI mediation condition:** In the weak AI mediation treatment, the videos were preprocessed with the skin smoothing, lighting adjustments and virtual backgrounds features offered by Microsoft Teams.

**T3 Strong AI mediation condition:** In the strong AI mediation treatment, we further processed the videos with Microsoft Teams’ avatar feature that replaced the person in the video with a fully synthetic representation of the speaker. For the treatment, 80 digital avatars were manually designed in the Microsoft Avatars App by the first author and a research assistant to resemble the speaker in the original video.

By using the integrated video-processing features of a video communication tool commonly used in professional settings, we can study comparatively common and realistic stimuli. In contrast to previous work studying more extreme forms of AI-mediated video communication, such as deepfakes [35, 100], our treatments provide insights into a communication setting encountered daily by millions of people. The calibration of our treatments was further informed by in-person testing and an initial pilot study with a subset of 16 videos from 8 video subjects and N = 100 participants.

### 3.3 Procedure and Measurements

Before beginning the study, all participants provided informed consent and read brief instructions explaining the main task. They also completed an attention-check question to confirm their understanding of the task before proceeding to the main task. The

main task consisted of evaluating six short, prerecorded video clips (approximately 40 seconds long) [66], preprocessed according to the treatment conditions described above. Videos were balanced for veracity, ensuring that each participant viewed exactly three truthful and three deceptive statements. Additionally, we balanced video subjects across gender and race. Although we had multiple videos per speaker, we ensured that each participant saw each speaker at most once. Participants could watch each video only once using standard playback controls, except for the replay option. We limited the number of videos to six per participant to balance sufficient exposure to each condition while minimizing participant fatigue.

After each video, participants answered three questions:

- O1 Veracity Judgment:** "Do you think this person is lying or telling the truth?" (Binary choice: Lie / Truth)
- O2 Judgment Confidence:** "How confident are you in your judgment?" (5-point Likert scale: Not at all confident to Extremely confident)
- O3 Trustworthiness:** "How trustworthy does the person in the video seem?" (5-point Likert scale: Not at all trustworthy to Extremely trustworthy)

We adjusted the scale directions across participants to mitigate order bias. In addition to the three main outcome variables of judgment, judgment confidence and trustworthiness, participants also answered an open-ended exploratory question for the last video only, where they explained their judgment ("Why do you think this person is lying or telling the truth?") and completed a multiple-choice question indicating what specific cues influenced their judgment ("Which of the following most influenced your judgment of whether the person was telling the truth?"). We allowed participants to select up to three cues from a set of eight cues we assembled based on a review of categories of cues reported in prior deception research: visual nonverbal cues, vocal/paraverbal cues, content-based cues and global demeanor [18, 61].

Finally, participants estimated their own judgment success ("Out of the six videos you watched, how many do you believe you correctly assessed?"). As a manipulation check, participants also indicated how many of the six videos they believed featured an avatar. They answered an open-ended exploratory question ("Did you notice anything unusual or artificial about the videos?"). The study concluded with demographic and exploratory questions about participants' use of online video communication platforms and their experience with AI-based video features. After submitting demographic information, participants received a detailed debriefing statement explaining the study's purpose.

### 3.4 Analysis Approach

We analyzed trust (H1), truth judgment rate (H2), accuracy (H3) and judgment confidence (H4) using separate linear mixed-effects models for each study, with AI mediation level as a fixed effect and random intercepts for participants to account for repeated measurements. To test the effect of the environment on trust, truth judgment rate, accuracy, and confidence (H5), we fitted a combined mixed-effects model across both studies with a fixed-effect interaction between AI mediation level (weak or strong) and environment type (homogeneous or mixed). As a further robustness check, we also estimated an extended version of these models that included

demographic and experience-related covariates: age, gender, education, race, English proficiency, prior experience with video tools, prior experience with AI tools, frequency of AI interaction, and general trust in AI. We report descriptive statistics of means and standard deviations alongside model estimates ( $\beta$  coefficients, 95% CIs, p-values). Hypotheses are evaluated based on the model results.

### 3.5 Recruitment

In Study 1 (between-subjects design), participants were randomly assigned to one of the treatment conditions and evaluated six videos of the same type. The between-subjects design isolates the impact of each manipulation, allowing for robust conceptual comparisons. In Study 2 (within-subjects design), participants viewed two videos per condition in random order, reflecting real-world variability in mediated communication and allowing us to capture how participants reacted to AI features in settings where some, but not all, people use them. To allow for valid comparisons across studies, participants for Study 1 and Study 2 were recruited concurrently from the same sample and randomly assigned to one of the two study designs.

A total of 2,000 participants were recruited through Prolific [81]. Eligibility criteria required that participants be 18 years or older and be resident in the United States. We determined the sample size based on a bootstrapped power analysis conducted before data collection. Note that our sample differs from the preregistration, as we initially planned to recruit only 1,000 participants. The initial power analysis, based on pilot study data, estimated the required sample size to achieve approximately 80% power to detect small changes ( $d = .2$ ) in trust and confidence. After collecting an initial 1,000 answers, we realized that relevant effect sizes in accuracy (3-5%, corresponding to a Cohen's  $d = .06$  to  $.1$ ) were substantially smaller than estimated, and the initially planned sample would leave us with variations in accuracy that were difficult to interpret. Increasing the sample size to  $N = 2,000$  enabled us to detect larger accuracy differences ( $d = .1$ ) with approximately 70% power. The larger sample should also have improved the robustness and interpretability of our overall study. As no changes in accuracy were detected with the increased sample, we see no risk of false positives.

We compensated participants \$2 for their participation, which, on average, took about 10 minutes, corresponding to a \$12 hourly rate. To encourage attentive responding and to raise the stakes of the scenario, we offered participants a \$2 bonus if they classified at least 5 videos correctly, doubling their base payment. 289 participants received the bonus payment. Participants ranged in age from 21 to 77 ( $M = 41$ ,  $SD = 13.7$ ,  $Median = 37$ ). Male participants represented 61.3% and female participants represented 38.7% of the sample. 65.8% of the sample self-identified as White or Caucasian; 22.5% as Black or African American; 4.5% as Asian; 4.5% as Latino or Hispanic; and 2.7% as Indigenous, Middle Eastern, North African, or mixed race. Most participants were highly educated, with 42.3% holding a bachelor's degree and 23.4% a master's degree. The majority were native English video subjects (90.1%), with the remainder reporting advanced or intermediate English proficiency.

## 4 Results

In this section, we present the empirical results from the experiments and analyze how AI-mediated video processing influenced interpersonal trust, judgment accuracy and confidence. The results reveal that AI mediation affects perceptions of trust and the confidence with which people make judgments, but has limited effects on actual judgment accuracy or on the tendency to believe others.

**Interpersonal trust (H1):** Figure 2 shows participants' trust in the person in the video across different types of AI mediation. The left panel shows trust ratings in homogeneous environments (Study 1), where participants encountered only one type of mediation, with the different types of mediation shown on the x-axis. In the control condition on the far left, where participants saw only original videos without AI mediation, the video subjects received an average trust rating of 0.51 (SD = 0.265), corresponding to "moderately trustworthy". Subjects using retouch effects received similar trust ratings ( $M = 0.504$ ,  $SD = 0.263$ ), while video subjects using avatar filters received slightly lower trust ratings ( $M = 0.485$ ,  $SD = 0.257$ ). We fitted a linear mixed model to predict reported trust by mediation type in Study 1 with a per-participant random fixed effect to account for the repeated measures design. The effect of avatar-based mediation on trust is statistically significant and negative ( $\beta = -0.03$ , 95% CI [-0.05, -0.002],  $t(5797) = -2.15$ ,  $p = .032$ ). In contrast, the effect of the weak AI mediation condition is statistically non-significant ( $\beta = -0.006$ , 95% CI [-0.03, 0.02],  $t(5797) = -0.52$ ,  $p = .604$ ). We provide further details on the model in Table 1 in the Appendix.

In Study 2, where participants encountered different types of AI mediation in a mixed environment (right panel in Figure 2), the effect of AI mediation on trust was more substantial. Video subjects in the original video, shown in the left column, received an average trust rating of 0.499 (SD = 0.264), similar to the trust ratings in the Study 1 control group. Video subjects who used retouches with virtual backgrounds received slightly lower trust ratings ( $M = 0.477$ ,  $SD = 0.259$ ), whereas video subjects who used avatars as the stronger AI mediation received substantially lower trust ratings ( $M = 0.419$ ,  $SD = 0.264$ ). We fitted a linear mixed model with a per-participant random fixed effect to predict reported trust by mediation type in Study 2. Compared to the control condition, the effect of AI mediation on perceived trust of the speaker in the mixed environment is statistically significant and negative for the retouch condition ( $M = .477$ ,  $\beta = -0.02$ , 95% CI [-0.04, -0.006],  $t(6235) = -2.81$ ,  $p = .005$ ) and for avatars ( $M = 0.419$ ,  $\beta = -0.08$ , 95% CI [-0.09, -0.07],  $t(6235) = -10.51$ ,  $p < .001$ ). We provide further details on the model in Table 2 in the Appendix.

Next, we analyzed the extent to which the effects of different types of AI mediation differed across environments by comparing results from Study 1 and Study 2 (H5). We fitted a linear mixed model predicting reported trust, with mediation type and environment type as predictors, across Study 1 and Study 2, including a per-participant random fixed effect to account for the repeated-measures design.

The interaction term for the avatar condition was statistically significant and negative, indicating that the trust penalty for avatars was substantially larger in mixed environments than in homogeneous ones ( $\beta = -0.05$ , 95% CI [-0.08, -0.03],  $t(12034) = -4.05$ ,  $p <$

.001). We found no reliable interaction for retouch videos ( $\beta = -0.02$ , 95% CI [-0.04, 0.01],  $t(12034) = -1.13$ ,  $p = .257$ ). The main effect of the environment was non-significant ( $\beta = -0.01$ , 95% CI [-0.03, 0.008],  $t(12034) = -1.14$ ,  $p = .253$ ), showing that baseline trust in the control condition did not differ between environments. We provide further details on the model in Table 3 in the Appendix.

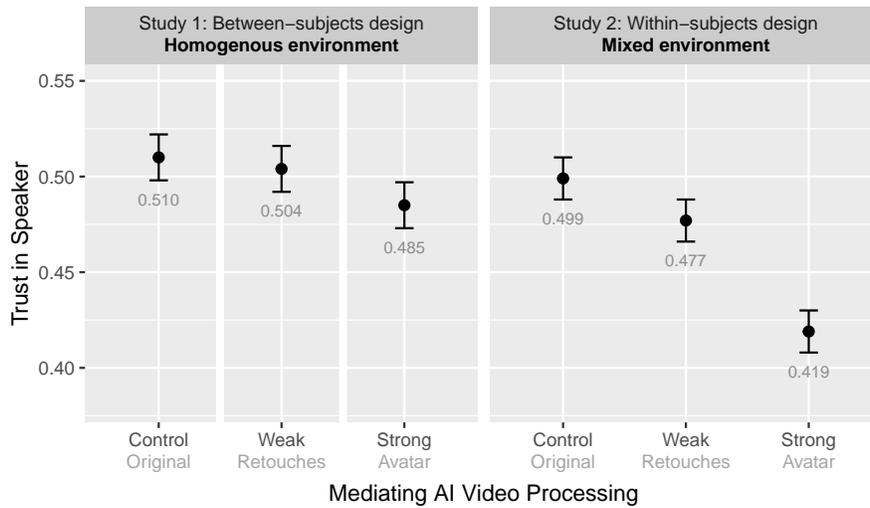
Overall, the results support the hypothesis that AI-mediated video influences perceived trustworthiness (H1). Considering the effect of the environment (H5), trust in avatar video subjects in mixed environments was significantly lower than in the homogeneous setting.

**Truth judgment rate (H2):** Although we observed a reduction in trust through AI mediation, we did not observe changes in how often participants thought the person in the video was lying. Figure 3 shows participants' truth judgment rate, that is, how often participants indicated that they thought the person in the video was telling the truth. In the control condition, in which participants were shown the original video, they believed video subjects were telling the truth about 60% of the time (59.9% in Study 1 and 59.7% in Study 2). Note that this rate is significantly higher than the ground truth frequency shown in grey, aligning with other studies showing that people are truth-biased [62, 63]. However, we observed no significant difference in truth judgment rates when video subjects used retouches (58.1% and 57.6%) or avatar filters (61.5% and 57.7%) across both studies. Given our overall sample size, we would have expected to detect a change of about 3-5% most of the time.

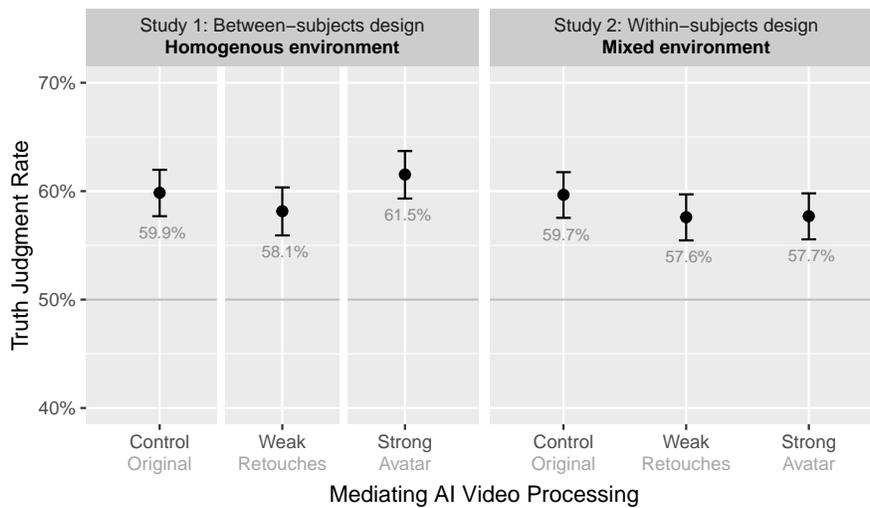
We fitted linear mixed models to predict reported truth judgment rate by mediation type in Study 1 and in Study 2 with a per-participant random fixed effect. In neither study was the effect of avatar-based mediation on truth judgment rate statistically significant. The effect of weak AI mediation (retouch) on truth judgment rate is statistically non-significant in Study 1 ( $\beta = -0.02$ , 95% CI [-0.05, 0.01],  $t(5797) = -1.09$ ,  $p = .278$ ) and Study 2 ( $\beta = -0.02$ , 95% CI [-0.05, 0.009],  $t(6235) = -1.35$ ,  $p = .176$ ), as is the effect in the avatar condition for Study 1 ( $\beta = 0.02$ , 95% CI [-0.01, 0.05],  $t(5797) = 1.07$ ,  $p = .284$ ) and Study 2 ( $\beta = -0.02$ , 95% CI [-0.05, 0.01],  $t(6235) = -1.29$ ,  $p = .19$ ). Furthermore, we fitted a linear mixed model to predict the reported truth judgment based on mediation type and environment type across Study 1 and Study 2 with a per-participant random fixed effect. None of the interaction terms were statistically significant, indicating no effects on the type of environment (H5). We provide further details on the models in Table 1, 2 and 3 in the Appendix.

We observed no differences in truth judgment rates across mediation types and environments. The results do not support H2 but are consistent with Levine's truth-default theory, which posits that people have a consistent truth bias and default to believe others unless clear suspicion is triggered [62].

**Judgment accuracy (H3):** Figure 4 shows participants' judgment accuracy across conditions, that is, how often they rated truths as truths and lies as lies. The black data points at the center of the graph show the overall accuracy on the combined set of stimuli, half of which contained truths and half of which contained lies. The grey reference line shows the baseline accuracy that participants would have achieved by random responses (50%). In line with findings in related work, participants were slightly better than random at telling truths from lies, with 52.3% accuracy in the Study



**Figure 2: AI mediation affects interpersonal trust, particularly in mixed environments. Average interpersonal trust by video mediation type and study design with 95% confidence intervals;  $N = 2,000$  ratings per data point. Participants in the left panel (Study 1) rated six videos of the same mediation type, whereas participants in the right panel (Study 2) watched two videos of each mediation type in random order.**

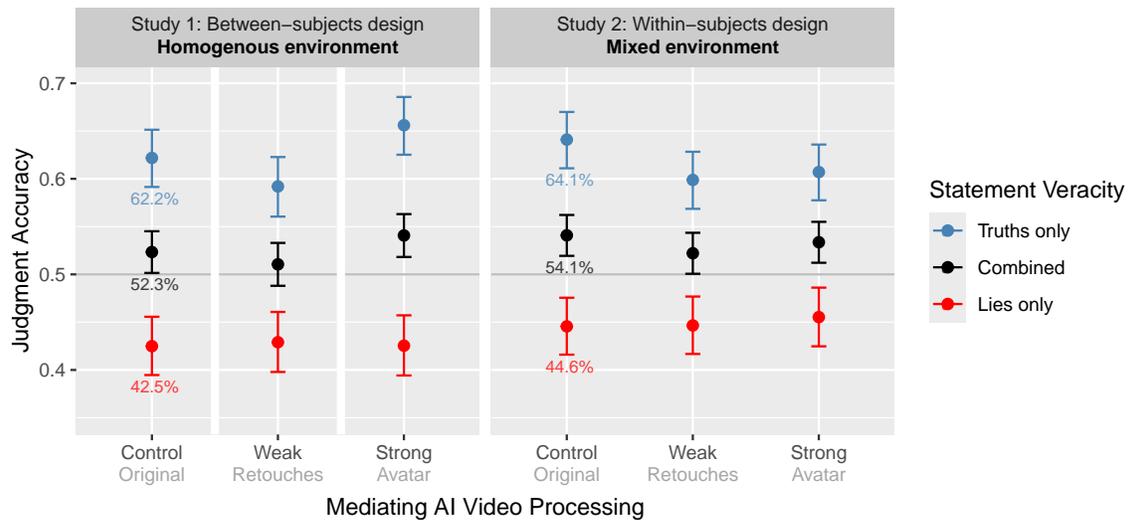


**Figure 3: Truth judgment rates were unaffected by mediating AI mediation. Percentage of times participants thought the person in the video told the truth, with 95% Wilson confidence intervals;  $N = 2,000$  judgments per data point. Across both studies, participants exhibit a bias towards truth judgment rate (57.6 to 61.5%) independent of AI mediation, while the stimuli dataset contained 50% truths and 50% lies.**

1 control group and 54.1% in the control group of Study 2. This rate did not change significantly across videos with retouching (51.1% and 52.2%) or avatar (54.1% and 53.4%) in Study 1 and Study 2.

We fitted two linear mixed models to predict judgment accuracy by mediation type in Study 1 and Study 2, with a per-participant random fixed effect. Participants achieved higher accuracy on videos in which subjects told the truth (shown in blue), with 62.2% in the control condition in Study 1 and 64.1% in the control condition in Study

2. While this level of accuracy is significantly higher ( $\beta = 0.20$ , 95% CI [0.17, 0.23],  $p < .001$ ) than the accuracies participants achieved in videos with lies (42.5% and 44.5% respectively), this difference largely reflects the general truth bias in participants' judgments observed above. In Study 2, we observed a non-significant reduction in accuracy in the retouch condition (59.9%,  $\beta = -0.02$ , 95% CI [-0.05, 0.01],  $t(6235) = -1.22$ ,  $p = .224$ ).



**Figure 4: Judgment accuracy by AI mediation type and statement veracity. Percentage of times participants correctly identified a truth as a truth or a lie as a lie, with 95% Wilson confidence intervals;  $N = 1,000-2,000$  judgments per data point. Participants identified about 60-65% of truths correctly, but only about 42-45% of lies. Except for truths in the avatar treatment in Study 1, AI mediation did not affect judgment accuracy.**

We tested the mediating effect of the environment type (H5) by fitting a linear mixed-effects model predicting accuracy from mediation type, study environment and their interaction. The model showed no statistically significant effects of either retouch ( $\beta = -0.007$ , 95% CI [-0.08, 0.06],  $p = .844$ ) or avatar mediation ( $\beta = 0.04$ , 95% CI [-0.03, 0.11],  $p = .238$ ) relative to the control condition. The interaction terms assessing whether mediation effects differed between homogeneous and mixed environments were also non-significant for both retouch ( $\beta = -0.006$ , 95% CI [-0.05, 0.04],  $p = .791$ ) and avatar conditions ( $\beta = -0.02$ , 95% CI [-0.07, 0.02],  $p = .270$ ). We provide the full model details in Table 3 in the Appendix.

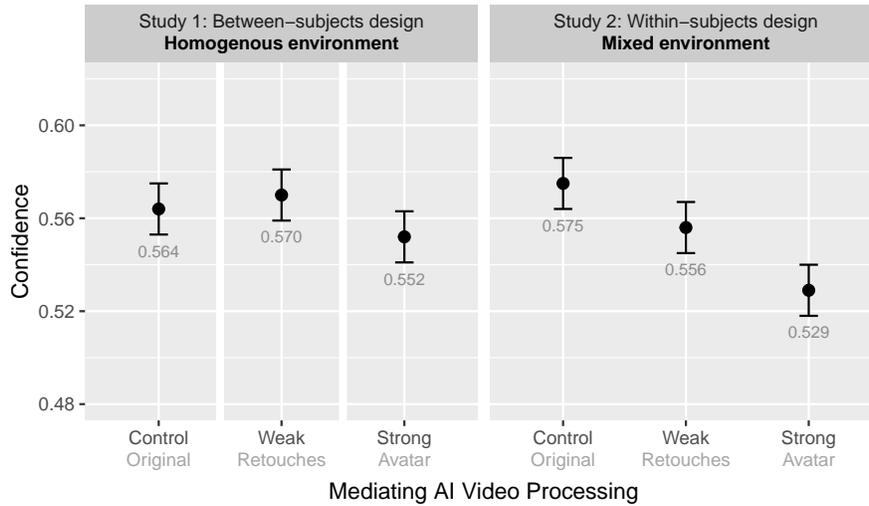
Participants achieved 51–54.5% overall accuracy, corresponding to just 1–4.5% above chance. Consequently, only relatively large AI mediation effects that meaningfully improved or reduced this limited margin of accuracy would be detectable. Although our studies were adequately powered to detect changes of 3–5%, they were not sufficiently sensitive to reliably capture smaller effects of 1–2%. Our results show no meaningful differences in accuracy across mediation types (H3) or environments (H5), aligning with prior meta-analysis work on deception detection research showing that human deception detection accuracy remains stable across viewing conditions and is only slightly above chance [8, 18].

**Judgment confidence (H4):** We fitted a linear mixed model to predict the reported judgment confidence based on mediation type in Study 1 and Study 2 with a per-participant random fixed effect to account for the repeated measures design. While AI mediation did not affect participants' judgment accuracy, it did affect their confidence in their judgments, as shown in Figure 5. In the control conditions, participants reported a mean confidence of 0.56 (SD = 0.258) in Study 1 and 0.57 (SD = 0.253) in Study 2, which falls between moderately confident (0.5) and very confident (0.75) on the Likert scale. In Study 1, where participants encountered only

one type of mediation, their confidence did not change significantly when they encountered videos with retouching ( $M = 0.57$ ,  $SD = 0.255$ ,  $\beta = 0.006$ , 95% CI [-0.02, 0.03],  $t(5797) = 0.44$ ,  $p = .661$ ) or avatar filters ( $M = 0.55$ ,  $SD = 0.248$ ,  $\beta = -0.01$ , 95% CI [-0.04, 0.02],  $t(5797) = -0.86$ ,  $p = .391$ ). In Study 2, however, where participants encountered a mix of different mediation types, they were significantly less confident in their judgments when evaluating videos with retouches ( $M = 0.556$ ,  $SD = 0.249$ ) and avatar filters ( $M = 0.529$ ,  $SD = 0.258$ ). Linear mixed-effects models confirm a significant negative effect compared to the control condition for retouch ( $M = 0.556$ ,  $\beta = -0.02$ , 95% CI [-0.03, -0.006],  $t(6235) = -2.94$ ,  $p = .003$ ) and avatar filters ( $M = 0.529$ ,  $\beta = -0.05$ , 95% CI [-0.06, -0.03],  $t(6235) = -7.10$ ,  $p < .001$ ). Details on the models are provided in Table 1 and Table 2 in the Appendix.

To test whether the effect of AI mediation on confidence differs between homogeneous and mixed environments (H5), we fitted a linear mixed-effects model with an interaction between level of AI mediation and environment, including a random intercept for participants (see Table 3). While the interaction for retouches was not statistically significant ( $\beta = -0.03$ , 95% CI [-0.05, 0.0043],  $t(12034) = -1.67$ ,  $p = .095$ ), the model shows a statistically significant interaction in the avatar condition ( $\beta = -0.03$ , 95% CI [-0.06, -0.004],  $t(12034) = -2.23$ ,  $p = .026$ ), suggesting that confidence drops more sharply for avatar-mediated videos in mixed environments than in homogeneous ones, aligning with findings from prior work on HCI and avatar-mediated communication research [83].

Judgment confidence decreased for AI-mediated video subjects, particularly for avatars and in mixed environments, consistent with H4 and H5. While confidence remained stable in homogeneous settings (Study 1), it declined in mixed environments in which participants had to compare differently mediated videos side by side,



**Figure 5: AI mediation affects judgment confidence, but only in mixed environments. Average reported confidence in judgment by video mediation type and study design with 95% confidence intervals;  $N = 2,000$  ratings per data point. Participants encountering different types of AI-mediated content and original content in Study 2 (right panel) were less confident in their judgments, particularly for strong AI-mediated content.**

aligning with prior work on avatar mediation [83] and Uncertainty Reduction Theory.

Figure 6 summarizes the answers participants gave when asked which cues or elements most influenced their judgments in the last video. We coded each cue option as a binary indicator (selected = 1, not selected = 0) and report the percentage of participants selecting each cue along with Wilson 95% confidence intervals. In a multiple-choice question, participants could select up to three cues, such as gaze and eye contact, body language, or speech fluency, shown on the y-axis on the left of Figure 6. The x-axis shows how often participants selected a cue, depending on the AI mediation type of the relevant video. Control and retouch videos are shown in dark and light grey, respectively, and avatar videos in red. We gathered the set of cues available from three major categories of deception research to cover the most reported cues according to prior work: visual nonverbal cues, vocal or paraverbal cues, content-based cues and global demeanor [18, 61].

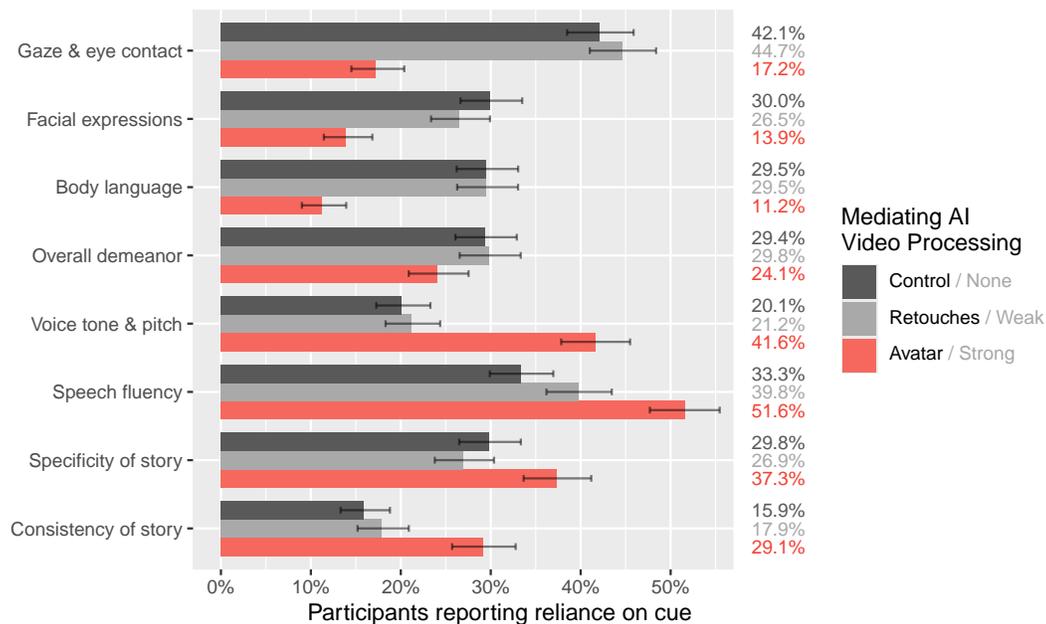
In the control condition, participants substantially relied on gaze and eye contact (42.1%), facial expressions (30%), and body language (29.5%) to make their judgments. The retouch condition (weak AI mediation) closely resembles the control condition in the use of nonverbal cues such as gaze and eye contact (44.7%), facial expressions (26.5%), and body language (29.5%). By contrast, when the speaker used an avatar instead, participants reported relying less on these nonverbal cues (17.2%, 13.9%, and 11.2%, respectively). Instead, participants shifted their attention from visual cues to audio and content-based cues, relying on voice tone and pitch (41.6%), speech fluency (51.6%), story specificity (37.3%), and consistency (29.1%) more than participants in the control group (20.1%, 33.3%, 29.8%, and 15.9%, respectively). Overall, these results show that while AI mediation did not affect participants' judgment accuracy, it did change how they arrived at their judgments and how confident

they felt about them, providing mechanisms that help explain why participants felt less confident in their judgments and trusted the speaker less in strongly AI-mediated videos.

#### 4.1 Manipulation Checks

As a manipulation check, we asked participants to indicate how many of the six videos they watched featured a fully-animated avatar after the main task. In Study 2 (mixed environment), participants rated exactly two avatar videos and, on average, reported a similar number ( $M = 2.22$ ,  $SD = 0.58$ ). In Study 1 (homogeneous environment), participants in the avatar condition (strong AI mediation) correctly recognized that all or most of the videos featured avatars ( $M = 5.78$ ,  $SD = 0.86$ ). In the open-ended question ("Did you notice anything unusual or artificial about the videos?"), participants also frequently commented on the avatar videos in the mixed environment (Study 2), describing them as "not a real person," "weird," "off," "unnatural," or "hard to read." Several participants noted that the avatar "looked artificial" or that the speaker disappeared to be "behind an avatar": "I immediately don't trust people who use avatars. They are generally 'social snipers' who hide behind anonymity so they don't have to be responsible for their actions" (Participant N474). Comments on the virtual background or visual filters from the retouching condition were less frequent. They typically referred to mild visual artifacts, such as "blurred background," "border around the person," or "the lighting seemed edited".

Overall, our two manipulation checks suggest that strong AI mediation was salient to participants, particularly in the mixed environment of Study 2. The qualitative responses indicate that participants noticed strong AI-mediated videos and occasionally noted minor artifacts in weak mediation conditions.



**Figure 6: Participants relied more on content-based cues and less on expressions and body language when evaluating AI-mediated content. Percentage of participants indicating reliance on different cues based on mediation type, with 95% Wilson confidence intervals;  $N = 2,000$ . Participants relied substantially on gaze, expressions and body language (top three rows) when evaluating the original and retouched videos, cues that were lost under strong AI mediation. Instead, participants relied more on voice, fluency, and content consistency (bottom four rows) when evaluating avatar videos.**

## 4.2 Robustness Checks

To increase the robustness of our results, we fitted an extended linear mixed-effects model across both studies with covariates including age, gender, race, education, English proficiency, subjective trust in AI systems, and prior experience with online video communication platforms, video filters and AI mediation, in addition to the independent variables of mediation type and study environment (see Appendix Table 4).

All significant effects reported in the previous studies remain significant in the extended model, after including the covariates: the model predicts that the negative effect of avatars on trust remains significant ( $\beta = -0.027$ ,  $p < .05$ ), particularly in the mixed environment ( $\beta = -0.080$ ,  $p < .001$ ). A somewhat smaller but statistically significant trust reduction was also predicted for retouched videos in mixed environments ( $\beta = -0.021$ ,  $p < .01$ ). Similarly, a significant decrease in confidence was predicted for retouched videos in mixed environments ( $\beta = -0.019$ ,  $p < .01$ ), with an even larger reduction for avatars in mixed environments ( $\beta = -0.046$ ,  $p < .001$ ). As in the main analysis, truth judgment rates and accuracy remained unaffected by mediation type.

## 5 Discussion

In our experiments, AI-mediated video communication substantially affected how people evaluated each other. Particularly in the strong AI-mediated treatment, where video subjects used synthetic avatars, participants' trust in the speaker and their confidence in their truth-lie judgments were reduced. While AI mediation

changed the cues on which participants relied for their judgments, it did not affect how often participants suspected others of lying, nor did it improve or impair judgments accuracy. The observed decreases in trust and confidence were moderated by the type of environment, with larger decreases in mixed environments, where participants encountered a mix of original and AI-mediated videos.

In the following, we discuss three possible interpretations of why AI-mediated video might undermine trust without triggering suspicion; how accuracy remains stable while cue reliance changes; and how confidence in judgment drops even though judgment accuracy does not change. We finish by outlining implications for design and policy.

### 5.1 AI-Mediated Video Reduces Interpersonal Trust Without Raising Suspicion

AI-mediated video consistently reduced interpersonal trust (H1), particularly when video subjects were replaced by avatars in a mixed environment alongside more natural, unaltered video subjects (H5). The findings align with prior work on avatar-mediated communication in which low-fidelity representations are trusted less than natural faces [68, 83]. The observed effect of reduced trust in strong AI mediation was amplified in mixed environments (Study 2), aligning with Expectancy Violations Theory (EVT), which posits that deviations from the internalized social expectations decrease trust in the speaker [11]. The mixed environment may have made AI mediation more salient and highlighted participants' expectations regarding how a speaker should present themselves on video,

leading to more negative trust evaluations. Here, our findings align with and extend the Replicant Effect [53], in which, in a mix of human-generated and AI-generated content, the trustworthiness of subjects suspected of using generated content decreases as people begin to question each other's humanity. Our findings show that the Replicant Effect holds in the more dynamic medium of video and that even after people have become more accustomed to various forms of AI systems in recent years, AI-mediated communication still decreases interpersonal trust. Our work also highlights that trust is reduced even under weaker forms of AI mediation, such as retouching and virtual backgrounds.

Surprisingly, although participants found AI-mediated video subjects less trustworthy, they believed them just as often as they believed those in the original video. The stability in truth and lie rates is consistent with Levine's truth-default theory [62], which posits that, by default, people believe others are telling the truth unless something triggers suspicion. While the visual unfamiliarity and synthetic nature of the avatars may have disrupted trust, it was not enough to trigger suspicion and override the default behavior of believing others.

The apparent contradiction of reduced trust in avatars yet continued belief in them is best understood by distinguishing trust and belief as psychologically distinct processes with different implications [47]. While trust reflects an affective and relational judgment about another person's characteristics, belief is a cognitive judgment about whether a statement is factually true [47]. Our findings extend prior work on HCI and avatar-mediated communication [68, 83] by showing that AI-mediated speakers can be believed without being trusted. That distinction matters because AI mediation can erode the interpersonal foundations of communication, such as trust, social connection and willingness to cooperate [82, 94], without undermining social epistemology and without leading to a mediation environment in which people begin to question mediated statements.

## 5.2 AI-Mediated Video Affects Cue Reliance But Not Deception Detection Accuracy

The present work extends prior work on the effects of avatar-mediated communication [83] by examining their effect on lie detection. Here, our results largely align with well-documented findings from deception research [8, 18]. In both studies, participants identified truths and lies with roughly 52–54% accuracy, replicating the average truth-lie accuracy of 54% reported in [8]. Participants were also more accurate at identifying truths than lies, an asymmetry known in deception research as the veracity effect [63]. However, accuracy rates did not differ meaningfully between the control condition, in which video subjects were unaltered, and the retouch or avatar conditions, in which facial expressions, eye contact and other visual behaviors were altered or removed.

The stable accuracy across conditions, even under strong AI mediation, contradicts the predictions of cue-based deception detection theories [23, 24]. These theories posit that deception is detected by leakage, that is, by observing a set of nonverbal cues involuntarily produced by the cognitive effort of lying. As avatar-mediated communication would substantially reduce the nonverbal cues that

might give away a liar, we would expect accuracy to decrease in environments with reduced leakage. Our studies, however, show that accuracy neither decreased nor improved in the avatar condition.

While we do not observe an effect of avatars on deception detection, the stable accuracy across conditions aligns with a common finding in lie detection research: people are not good at detecting lies [18], regardless of the type of mediation. Instead, the stable accuracy across conditions supports a heuristic view of lie detection [8, 61], which holds that judgments are driven by content-based cues like plausibility, coherence and fact-checking, rather than leakage through involuntary nonverbal cues. These content-based cues, such as plausibility, coherence, and fact-checking, are arguably less affected by the use of retouching, backgrounds and video avatars.

While accuracy was not affected by AI mediation, participants shifted to rely on speech fluency, story consistency, and specificity, rather than to facial expressions or body language. This finding is again paralleled by the central tenet of Levine's truth-default theory [62], which holds that content rather than demeanor is the basis for detecting deception. Based on Levine's truth-default theory, one might have expected the shift to content-based cues to lead to even greater accuracy, as when people pay more attention to content cues, they are more likely to notice inconsistencies. However, accuracy remained stable across conditions, which may be due to the nature of the interactions we studied: in our tasks, participants judged statements about unknown people, with no means to verify the claims or compare them with existing knowledge, limiting the potential to challenge the default assumption of truth. Although participants shifted to rely on content-based and verbal cues, without information to contextualize or fact-check the statements, participants may have rarely had the chance to notice inconsistencies that might have broken their truth-default bias.

## 5.3 AI-Mediated Video Complicates Judgments and Reduces Judgment Confidence

Participants' subjective confidence in their own judgments decreased in the mixed environment (Study 2), particularly for the avatar condition—again, despite their truth-lie rates and accuracy staying the same.

The decrease in confidence in judgments for avatars aligns with prior findings on avatar-mediated communication [83] and is supported by both cue-based theories and Uncertainty Reduction Theory. While decades of deception research have shown that nonverbal cues like gaze or facial expressions are poor indicators of lying [18], people continue to rely on them, or at least, they believe they depend on them. When these familiar cues are removed or disrupted, people feel less confident about what to rely on instead. In this way, these cues serve a social-psychological function: they reduce uncertainty and help individuals feel familiar in social interactions. Uncertainty Reduction Theory [6] posits that people are motivated to seek information that reduces ambiguity in social communication. AI-mediated communication interrupts that process by stripping away or synthetically simulating these social cues. This disruption holds even after people have gotten more accustomed to AI tools, and even for commonly used transformations such as lighting corrections and virtual backgrounds. Even if the removal of familiar cues does not impair accuracy, it leaves participants feeling

less sure of themselves because they can't rely on some signals that they would usually rely on to make interpretation easier.

The most pronounced decline in judgment confidence is in Study 2, in which participants saw all three types of mediation within the same environment. The mix of mediation types likely heightened the salience of visual disruption and made avatars feel more unpredictable or "out of place". Here, Expectancy Violation Theory [11] posits that people have internalized expectations for how social interactions should look and feel, and when those expectations are violated, such as encountering a synthetic face after two natural ones, it creates friction in the interpretation of the communication. While friction does not make the person seem more deceptive, it may make the interaction more challenging to process. Previous research suggests that a disruption in processing fluency leads to lower confidence in judgments [11, 111]. This interpretation aligns with participants' open-ended responses in our study, in which they describe avatars as "hard to read" or "off-putting", suggesting that a breakdown in fluency of interpretation makes people feel less confident about what to make of the communication.

## 5.4 Implications

Our findings show that AI-mediated video did not affect truth judgment rates or detection accuracy. As such, our findings challenge concerns that "with their ability to alter users' appearances dramatically, beauty filters can facilitate deception" [71]. Similarly, AI-mediated video has been discussed as a potential threat to people's ability to judge honesty accurately [57, 84]. Here, although we find that AI-mediated communication can affect trust and confidence, it does not substantially affect people's ability to detect lies in video communication. We note, however, that even under AI mediation, people's ability to detect deception remains close to chance.

However, consistent with prior work on trust in avatars [83], we show that speakers using AI-mediated video were trusted less and participants felt less confident in their judgments [17, 41, 80]. The mismatch between judgment performance and relational trust and confidence is an important implication [4, 14] in contexts where the feeling of certainty of knowing can be as important as the decision itself. In high-stakes settings such as remote hiring, clinical evaluations or legal proceedings, lower confidence may lead to hesitation, increased caution or reduced assertiveness [21, 65, 86, 95]. Our results also broaden the debate about the risks of AI video beyond deepfakes [35, 88]: even widely used filters like retouching and virtual backgrounds affect evaluations of trust and credibility in ways that matter [1, 37]. Our results show that the central risk of AI-mediated video may lie in the erosion of trust and confidence in judgments, particularly when mixed environments make mediation salient.

When designing video communication platforms, the question is how new AI-based features that may improve aesthetics and convenience might also undermine trust and confidence. Here, further research is needed to understand what elements are required for representational consistency within calls, what forms of disclosure might mitigate reductions in trust, and higher-fidelity or more expressive avatars may preserve the aspects of communication that people feel they need to feel confident in their judgments [68, 83].

Our findings also highlight the need for more context-sensitive forms of AI mediation: communication tools could offer users different AI mediation levels depending on the situation—for example, realistic appearances in professional calls and more stylized options in informal chats—to adjust the communication to the expectations of the context and to calibrate the affordances of the communication based on context-specific needs.

## 5.5 Limitations and Future Directions

Participants evaluated short, prerecorded videos from [66] in which video subjects made simple personal statements about people unknown to the participants. Despite being incentivized with a bonus payment, the simulated scenario did not constitute a high-stakes situation for participants, and speakers in the video likely felt minimal pressure to lie. A low-stakes context may have muted the behavioral leakage or suspicion triggers on which cue-based theories depend. Future work should examine whether AI-mediated video has a different impact on high-stakes lies, where participants are more motivated to detect deception and speakers are under pressure. Such settings could include hiring, legal evaluations or sensitive interpersonal disclosures, where stakes and incentives are real and carry consequences.

Furthermore, AI-mediated communication may be perceived differently in ongoing teams, family calls or workplace meetings. As the video subjects were strangers to the participants, the findings may not generalize to communication with more familiar groups, where prior work shows that familiarity can attenuate negative impressions of mediated cues [75] and that contextual knowledge can increase lie-detection accuracy. Future work could examine how AI mediation affects trust and confidence in contexts where familiarity and existing relationships shape impressions and social expectations.

In addition, the experimental setup lacks the fluid interactive nature of real-time video communication. In live conversations, participants can ask follow-up questions, interpret timing and adapt to social feedback. More research is needed to investigate how AI-mediated appearances affect interpersonal dynamics in live or semi-structured conversations, particularly in collaborative or conflict-prone contexts such as negotiations or interviews. Finally, although participants at the time of our study (August 2025) had substantial exposure to AI tools and weaker forms of AI-mediated video, such as retouching or virtual backgrounds, strong AI-mediated video is still unevenly adopted. As AI-mediated video and synthetic appearances become more normalized, user expectations and reactions may shift. Future work should track how perceptions of AI-mediated video evolve, including through longitudinal studies and cross-cultural comparisons, to assess whether increased exposure to AI reinforces or attenuates the observed effects.

## Acknowledgments

We thank our research assistant, José Agostinho, for assisting the first author with processing the video stimuli and generating the avatars.

We acknowledge the use of ChatGPT for reviewing the author's original writing and for proposing phrasing improvements to increase clarity. All manuscript text was written and finalized by the authors.

## References

- [1] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. Trust in AI: Progress, Challenges, and Future Directions. *Humanities and Social Sciences Communications* 11, 1 (Nov. 2024), 1568. doi:10.1057/s41599-024-04044-8
- [2] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. 2024. Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 2 (March 2024), 719–728. doi:10.1609/aaai.v38i2.27829
- [3] Leslie A. Baxter and Dawn O. Braithwaite. 2008. *Engaging Theories in Interpersonal Communication: Multiple Perspectives*. SAGE.
- [4] Charles Bellemare and Alexander Sebald. 2019. *Self-Confidence and Reactions to Subjective Performance Evaluations*. Technical Report. IZA - Institute of Labor Economics. jstor:resrep66649
- [5] Charles R. Berger and James J. Bradac. 1982. *Language and Social Knowledge: Uncertainty in Interpersonal Relations*. E. Arnold.
- [6] Charles R. Berger and Richard J. Calabrese. 1975. Some Explorations in Initial Interaction and Beyond: Toward a Developmental Theory of Interpersonal Communication. *Human Communication Research* 1, 2 (Dec. 1975), 99–112. doi:10.1111/j.1468-2958.1975.tb00258.x
- [7] Bokolo Anthony Jnr. 2020. Use of Telemedicine and Virtual Care for Remote Treatment in Response to COVID-19 Pandemic. *Journal of Medical Systems* 44, 7 (June 2020), 132. doi:10.1007/s10916-020-01596-5
- [8] Charles F. Bond and Bella M. DePaulo. 2006. Accuracy of Deception Judgments. *Personality and Social Psychology Review* 10, 3 (Aug. 2006), 214–234. doi:10.1207/s15327957pspr1003\_2
- [9] Nathan Bos, Judy Olson, Darren Gergle, Gary Olson, and Zach Wright. 2002. Effects of Four Computer-Mediated Communications Channels on Trust Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. Association for Computing Machinery, New York, NY, USA, 135–140. doi:10.1145/503376.503401
- [10] Michael Boyle, Christopher Edwards, and Saul Greenberg. 2000. The Effects of Filtered Video on Awareness and Privacy. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/358916.358935
- [11] Judee Burgoon. 2015. Expectancy Violations Theory. doi:10.1002/9781118540190.wbeic102
- [12] J.K. Burgoon, G.A. Stoner, J.A. Bonito, and N.E. Dunbar. 2003. Trust and Deception in Mediated Communication. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of The*. 11 pp.–. doi:10.1109/HICSS.2003.1173792
- [13] Brian L. Connelly, S. Trevis Certo, R. Duane Ireland, and Christopher R. Reutzel. 2011. Signaling Theory: A Review and Assessment. *Journal of Management* 37, 1 (Jan. 2011), 39–67. doi:10.1177/0149206310388419
- [14] Isabelle Dauriche, Hugh Rabagliati, and Kenny Smith. 2021. Subjective Confidence Influences Word Learning in a Cross-Situational Statistical Learning Task. *Journal of Memory and Language* 121 (Dec. 2021), 104277. doi:10.1016/j.jml.2021.104277
- [15] BM DePaulo. 1985. Deceiving and Detecting Deceit. *The Self and Social Life* (1985), 323–370.
- [16] Bella M. DePaulo, Kelly Charlton, Harris Cooper, James J. Lindsay, and Laura Muhlenbruck. 1997. The Accuracy-Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review* 1, 4 (Nov. 1997), 346–357. doi:10.1207/s15327957pspr0104\_5
- [17] Bella M. DePaulo, Kelly Charlton, Harris Cooper, James J. Lindsay, and Laura Muhlenbruck. 1997. The Accuracy-Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review* 1, 4 (Nov. 1997), 346–357. doi:10.1207/s15327957pspr0104\_5
- [18] Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to Deception. *Psychological Bulletin* 129, 1 (2003), 74–118. doi:10.1037/0033-2909.129.1.74
- [19] Nicola Döring, Katrien De Moor, Markus Fiedler, Katrin Schoenberg, and Alexander Raake. 2022. Videoconference Fatigue: A Conceptual Analysis. *International journal of environmental research and public health* 19, 4 (2022), 2061. doi:10.3390/ijerph19042061
- [20] Nicola Döring, Katrien De Moor, Markus Fiedler, Katrin Schoenberg, and Alexander Raake. 2022. Videoconference Fatigue: A Conceptual Analysis. *International Journal of Environmental Research and Public Health* 19, 4 (Jan. 2022), 2061. doi:10.3390/ijerph19042061
- [21] Kit S. Double and Damian P. Birney. 2024. Confidence Judgments Interfere with Perceptual Decision Making. *Scientific Reports* 14, 1 (June 2024), 14133. doi:10.1038/s41598-024-64575-7
- [22] Amy Edmondson. 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44, 2 (June 1999), 350–383. doi:10.2307/2666999
- [23] Paul Ekman. 1997. Lying and Deception. In *Memory for Everyday and Emotional Events*. Psychology Press.
- [24] Paul Ekman and Wallace V. Friesen. 1969. Nonverbal Leakage and Clues to Deception†. *Psychiatry* 32, 1 (Feb. 1969), 88–106. doi:10.1080/00332747.1969.11023575
- [25] Paul Ekman and Wallace V. Friesen. 1974. Detecting Deception from the Body or Face. *Journal of Personality and Social Psychology* 29, 3 (1974), 288–298. doi:10.1037/h0036006
- [26] Nicole B. Ellison and Jeffrey T. Hancock. 2013. Profile as Promise: Honest and Deceptive Signals in Online Dating. *IEEE Security & Privacy* 11, 5 (Sept. 2013), 84–88. doi:10.1109/MSP.2013.120
- [27] Eyal Ert, Aliza Fleischer, and Nathan Magen. 2016. Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb. *Tourism Management* 55 (Aug. 2016), 62–73. doi:10.1016/j.tourman.2016.01.013
- [28] Motahhare Esлами, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. doi:10.1145/2858036.2858494
- [29] Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. 2014. *Media Technologies: Essays on Communication, Materiality, and Society*. MIT Press.
- [30] Ella Glikson and Omri Asscher. 2023. AI-mediated Apology in a Multilingual Work Context: Implications for Perceived Authenticity and Willingness to Forgive. *Computers in Human Behavior* 140 (2023), 107592.
- [31] G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. 2021. Mental Models and Expectation Violations in Conversational AI Interactions. *Decision Support Systems* 144 (May 2021), 113515. doi:10.1016/j.dss.2021.113515
- [32] Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. 2023. Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning. arXiv:2303.12745 [cs] doi:10.48550/arXiv.2303.12745
- [33] Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-Of-Lies: A Multimodal Dataset for Deception Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [34] Adib Habbal, Mohamed Khalif Ali, and Mustafa Ali Abuzaraida. 2024. Artificial Intelligence Trust, Risk and Security Management (AI TRISM): Frameworks, Applications, Challenges and Future Research Directions. *Expert Systems with Applications* 240 (April 2024), 122442. doi:10.1016/j.eswa.2023.122442
- [35] Jeffrey T. Hancock and Jeremy N. Bailenson. 2021. The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (March 2021), 149–152. doi:10.1089/cyber.2021.29208.jth
- [36] Jeffrey T. Hancock and Jamie Guillory. 2015. Deception with Technology. In *The Handbook of the Psychology of Communication Technology*. John Wiley & Sons, Ltd, Chapter 12, 270–289. doi:10.1002/9781118426456.ch12
- [37] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. doi:10.1093/jcmc/zmz022
- [38] Jeffrey T. Hancock, Catalina Toma, and Nicole Ellison. 2007. The Truth about Lying in Online Dating Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 449–452. doi:10.1145/1240624.1240697
- [39] Jeffrey T. Hancock, Michael T. Woodworth, and Saurabh Goorha. 2010. See No Evil: The Effect of Communication Medium and Motivation on Deception Detection. *Group Decision and Negotiation* 19, 4 (July 2010), 327–343. doi:10.1007/s10726-009-9169-7
- [40] Russell Hardin. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.
- [41] Nigel Harvey. 1997. Confidence in Judgment. *Trends in Cognitive Sciences* 1, 2 (May 1997), 78–82. doi:10.1016/S1364-6613(97)01014-0
- [42] Aubree A. Herman, Sydney E. Brammer, and Narisra M. Punyanunt-Carter. 2025. Face Off: Exploring College Students' Perceptions Regarding Face Filters on TikTok. *Media Watch* 16, 1 (Jan. 2025), 93–107. doi:10.1177/09760911241291950
- [43] Susan C. Herring. 2002. Computer-Mediated Communication on the Internet. *Annual Review of Information Science and Technology* 36, 1 (2002), 109–168. doi:10.1002/aris.1440360104
- [44] N. Sharon Hill, Kathryn M. Bartol, Paul E. Tesluk, and Gosia A. Langa. 2009. Organizational Context and Face-to-Face Interaction: Influences on the Development of Trust and Collaborative Behaviors in Computer-Mediated Groups. *Organizational Behavior and Human Decision Processes* 108, 2 (March 2009), 187–201. doi:10.1016/j.obhdp.2008.10.002
- [45] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*

- (CHI EA '18). Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3170427.3188487
- [46] Jess Hohenstein and Malte Jung. 2020. AI as a Moral Crumple Zone: The Effects of AI-mediated Communication on Attribution and Trust. *Computers in Human Behavior* 106 (May 2020), 106190. doi:10.1016/j.chb.2019.106190
- [47] Richard Holton. 1994. Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy* 72, 1 (March 1994), 63–76. doi:10.1080/00048409412345881
- [48] Joo Wha Hong, Qiyao Peng, and Dmitri Williams. 2021. Are You Ready for Artificial Mozart and Skrillex? An Experiment Testing Expectancy Violation Theory and AI Music. *New Media & Society* 23, 7 (July 2021), 1920–1935. doi:10.1177/1461444820925798
- [49] Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (March 2021), 188–193. doi:10.1089/cyber.2020.0174
- [50] Kori M. Inkpen and Mara Sedlins. 2011. Me and My Avatar: Exploring Users' Comfort with Avatars for Workplace Communication. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 383–386. doi:10.1145/1958824.1958883
- [51] Tim Jacks. 2021. Research on Remote Work in the Era of COVID-19. *Journal of Global Information Technology Management* 24, 2 (April 2021), 93–97. doi:10.1080/1097198X.2021.1914500
- [52] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15. arXiv:2302.00560 [cs] doi:10.1145/3544548.3581196
- [53] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. doi:10.1145/3290605.3300469
- [54] Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, Graeme McLean, Yvonne Rogers, Paul Marshall, and Luk Warlop. 2022. "What Lies behind the Filter?" Uncovering the Motivations for Using Augmented Reality (AR) Face Filters on Social Media and Their Effect on Well-Being. *Computers in Human Behavior* 128 (March 2022), 107126. doi:10.1016/j.chb.2021.107126
- [55] Gareth R. Jones and Jennifer M. George. 1998. The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. *Academy of Management Review* 23, 3 (July 1998), 531–546. doi:10.5465/amr.1998.926625
- [56] Sasa Junuzovic, Kori Inkpen, John Tang, Mara Sedlins, and Kristie Fisher. 2012. To See or Not to See: A Study Comparing Four-Way Avatar, Video, and Audio Conferencing for Work. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work (GROUP '12)*. Association for Computing Machinery, New York, NY, USA, 31–34. doi:10.1145/2389176.2389181
- [57] Nimmi Kanji. [n. d.]. Why AI Filters Can Take a Toll on Our Self-Esteem | TELUS. <https://www.telus.com/en/wisere/resources/content/article/why-ai-filters-can-take-a-toll-on-our-self-esteem>.
- [58] Sandra Kiffin-Petersen and John Cordery. 2003. Trust, Individualism and Job Characteristics as Predictors of Employee Preference for Teamwork. *The International Journal of Human Resource Management* 14, 1 (Feb. 2003), 93–116. doi:10.1080/09585190210158538
- [59] Jihyun Kim, Kelly Merrill Jr., Kun Xu, and Stephanie Kelly. 2022. Perceived Credibility of an AI Instructor in Online Education: The Role of Social Presence and Voice Features. *Computers in Human Behavior* 136 (Nov. 2022), 107383. doi:10.1016/j.chb.2022.107383
- [60] Margarita Leib, Nils Köbis, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch. 2024. Corrupted by Algorithms? How AI-generated and Human-written Advice Shape (Dis)Honesty. *The Economic Journal* 134, 658 (Feb. 2024), 766–784. doi:10.1093/ej/thead056
- [61] Timothy R. Levine. [n. d.]. Scientific Evidence and Cue Theories in Deception Research: Reconciling Findings From Meta-Analyses and Primary Experiments. ([n. d.]).
- [62] Timothy R. Levine. 2014. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology* 33, 4 (Sept. 2014), 378–392. doi:10.1177/0261927X14535916
- [63] Timothy R. Levine, Hee Sun Park, and Steven A. McCornack. 1999. Accuracy in Detecting Truths and Lies: Documenting the "Veracity Effect". *Communication Monographs* 66, 2 (June 1999), 125–144. doi:10.1080/03637759909376468
- [64] Zijian Lew and Joseph B. Walther. 2023. Social Scripts and Expectancy Violations: Evaluating Communication with Human or AI Chatbot Interactants. *Media Psychology* 26, 1 (Jan. 2023), 1–16. doi:10.1080/15213269.2022.2084111
- [65] Zan Liu. 2024. The Asymmetric Impact of Decision-Making Confidence on Regret and Relief. *Frontiers in Psychology* 15 (April 2024). doi:10.3389/fpsyg.2024.1365743
- [66] E. Paige Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. 2019. Miami University Deception Detection Database. *Behavior Research Methods* 51, 1 (Feb. 2019), 429–439. doi:10.3758/s13428-018-1061-4
- [67] Heather Lukas, Changhao Xu, You Yu, and Wei Gao. 2020. Emerging Telemedicine Tools for Remote COVID-19 Diagnosis, Monitoring, and Management. *ACS Nano* 14, 12 (Dec. 2020), 16180–16193. doi:10.1021/acsnano.0c08494
- [68] Fang Ma, Ju Zhang, Lev Tankelevitch, Payod Panda, Torang Asadi, Charlie Hewitt, Lohit Petikam, James Clemons, Marco Gillies, Xueni Pan, Sean Rintel, and Marta Wilczkowiak. 2025. Nods of Agreement: Webcam-Driven Avatars Improve Meeting Outcomes and Avatar Satisfaction Over Avatar-Driven or Static Avatars in All-Avatar Work Videoconferencing. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW142:1–CSCW142:28. doi:10.1145/3711040
- [69] Xiao Ma, Jeffrey T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 2397–2409. doi:10.1145/2998181.2998269
- [70] Devin M. Mann, Ji Chen, Rumi Chunara, Paul A. Testa, and Oded Nov. 2020. COVID-19 Transforms Health Care through Telemedicine: Evidence from the Field. *Journal of the American Medical Informatics Association* 27, 7 (2020), 1132–1135.
- [71] Bernard Marr. [n. d.]. Picture Perfect: The Hidden Consequences Of AI Beauty Filters. <https://www.forbes.com/sites/bernardmarr/2023/06/09/picture-perfect-the-hidden-consequences-of-ai-beauty-filters/>.
- [72] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model Of Organizational Trust. *Academy of Management Review* 20, 3 (July 1995), 709–734. doi:10.5465/amr.1995.9508080335
- [73] Julie M. McCarthy, Donald M. Truxillo, Talya N. Bauer, Berrin Erdogan, Yiduo Shao, Mo Wang, Joshua Liff, and Cari Gardner. 2021. Distressed and Distracted by COVID-19 during High-Stakes Virtual Interviews: The Role of Job Interview Anxiety on Performance and Reactions. *Journal of Applied Psychology* 106, 8 (2021), 1103–1117. doi:10.1037/apl0000943
- [74] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–14. doi:10.1145/3449091
- [75] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–14. doi:10.1145/3449091
- [76] Dena F. Mujtaba and Nihar R. Mahapatra. 2025. Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. arXiv:2405.19699 [cs] doi:10.48550/arXiv.2405.19699
- [77] BBC News. 2021. Lawyer Gets Stuck with Cat Filter during Virtual Court Case. <https://www.bbc.com/news/av/world-us-canada-56005428>. BBC (2021).
- [78] Kristine L. Nowak and Jesse Fox. 2018. Avatars and Computer-Mediated Communication: A Review of the Definitions, Uses, and Effects of Digital Representations. *Review of Communication Research* 6 (2018), 30–53. doi:10.12840/issn.2255-4165.2018.06.01.015
- [79] Brid O'Connell, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human-Computer Interaction* 8, 4 (Dec. 1993), 389–428. doi:10.1207/s15327051hci0804\_4
- [80] Marcus O'Connor. 1989. Models of Human Behaviour and Confidence in Judgement: A Review. *International Journal of Forecasting* 5, 2 (Jan. 1989), 159–169. doi:10.1016/0169-2070(89)90083-6
- [81] Stefan Palan and Christian Schitter. 2018. Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* 17 (March 2018), 22–27. doi:10.1016/j.jbef.2017.12.004
- [82] Ye Pan and Anthony Steed. 2017. The Impact of Self-Avatars on Trust and Collaboration in Shared Virtual Environments. *PLOS ONE* 12, 12 (Dec. 2017), e0189078. doi:10.1371/journal.pone.0189078
- [83] Payod Panda, Molly Jane Nicholas, Mar Gonzalez-Franco, Kori Inkpen, Eyal Ofek, Ross Cutler, Ken Hinckley, and Jaron Lanier. 2022. AllTogether: Effect of Avatars in Mixed-Modality Conferencing Environments. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3533406.3539658
- [84] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns* 5, 5 (May 2024), 100988. doi:10.1016/j.patter.2024.100988
- [85] Malcolm R. Parks and Mara B. Adelman. 1983. Communication Networks and the Development of Romantic Relationships: An Expansion of Uncertainty Reduction Theory. *Human Communication Research* 10, 1 (Sept. 1983), 55–79. doi:10.1111/j.1468-2958.1983.tb00004.x
- [86] Andrea L. Patalano and Zachary LeClair. 2011. The Influence of Group Decision Making on Indecisiveness-Related Decisional Confidence. *Judgment and Decision Making* 6, 2 (Feb. 2011), 163–175. doi:10.1017/S1930297500004113
- [87] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception Detection Using Real-life Trial Data. In *Proceedings of the*

- 2015 ACM on International Conference on Multimodal Interaction. ACM, Seattle Washington USA, 59–66. doi:10.1145/2818346.2820758
- [88] Claudiu Popa, Rex Pallath, Liam Cunningham, Hewad Tahiri, Abiram Kesavara-jah, and Tao Wu. 2025. Deepfake Technology Unveiled: The Commoditization of AI and Its Impact on Digital Trust. doi:10.48550/ARXIV.2506.07363
- [89] Zoe A. Purcell, Mengchen Dong, Anne-Marie Nussberger, Nils Köbis, and Maurice Jakesch. 2024. People Have Different Expectations for Their Own versus Others' Use of AI-mediated Communication Tools. *British Journal of Psychology* (Sept. 2024), bjop.12727. doi:10.1111/bjop.12727
- [90] Dana Remus and Frank Levy. 2017. Can Robots Be Lawyers: Computers, Lawyers, and the Practice of Law. *Georgetown Journal of Legal Ethics* 30 (2017), 501.
- [91] Minjin (MJ) Rheu, Yue (Nancy) Dai, Jingbo Meng, and Wei Peng. 2024. When a Chatbot Disappoints You: Expectancy Violation in Human-Chatbot Interaction in a Social Support Context. *Communication Research* 51, 7 (Oct. 2024), 782–814. doi:10.1177/00936502231221669
- [92] Catherine M Ridings, David Gefen, and Bay Arinze. 2002. Some Antecedents and Effects of Trust in Virtual Communities. *The Journal of Strategic Information Systems* 11, 3 (Dec. 2002), 271–295. doi:10.1016/S0963-8687(02)00021-5
- [93] Elena Rocco. 1998. Trust Breaks down in Electronic Contexts but Can Be Repaired by Some Initial Face-to-Face Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '98*. ACM Press, Los Angeles, California, United States, 496–502. doi:10.1145/274644.274711
- [94] Oliver Schilke and Martin Reimann. 2025. The Transparency Dilemma: How AI Disclosure Erodes Trust. *Organizational Behavior and Human Decision Processes* 188 (May 2025), 104405. doi:10.1016/j.obhdp.2025.104405
- [95] Lydia Schooler, Mesude Okhan, Sara Hollander, Maureen Gill, Yoonseo Zoh, M.J. Crockett, and Hongbo Yu. 2024. Confidence in Moral Decision-Making. *Collabra: Psychology* 10, 1 (Aug. 2024), 121387. doi:10.1525/collabra.121387
- [96] Sachin Sharma, Raj Rawal, and Dharmesh Shah. 2023. Addressing the Challenges of AI-based Telemedicine: Best Practices and Lessons Learned. *Journal of education and health promotion* 1 (2023), 338.
- [97] Kristen M. Shockley, Tammy D. Allen, Hope Dodd, and Aashna M. Waiwood. 2021. Remote Worker Communication during COVID-19: The Role of Quantity, Quality, and Supervisor Expectation-Setting. *Journal of applied psychology* 106, 10 (2021), 1466.
- [98] Tony L. Simons and Randall S. Peterson. 2000. Task Conflict and Relationship Conflict in Top Management Teams: The Pivotal Role of Intragroup Trust. *Journal of Applied Psychology* 85, 1 (2000), 102–111. doi:10.1037/0021-9010.85.1.102
- [99] Hung-Yue Suen and Kuo-En Hung. 2024. Revealing the Influence of AI and Its Interfaces on Job Candidates' Honest and Deceptive Impression Management in Asynchronous Video Interviews. *Technological Forecasting and Social Change* 198 (Jan. 2024), 123011. doi:10.1016/j.techfore.2023.123011
- [100] John Twomey, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 2023. Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets That Discuss Deepfakes in the Russian Invasion of Ukraine. *PLOS ONE* 18, 10 (Oct. 2023), e0291668. doi:10.1371/journal.pone.0291668
- [101] Deedra Vargo, Lin Zhu, Briana Benwell, and Zheng Yan. 2021. Digital Technology Use during COVID-19 Pandemic: A Rapid Review. *Human Behavior and Emerging Technologies* 3, 1 (2021), 13–24. doi:10.1002/hbe.2.242
- [102] Deedra Vargo, Lin Zhu, Briana Benwell, and Zheng Yan. 2021. Digital Technology Use during COVID -19 Pandemic: A Rapid Review. *Human Behavior and Emerging Technologies* 3, 1 (Jan. 2021), 13–24. doi:10.1002/hbe.2.242
- [103] Aldert Vrij. 2000. *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. Wiley, Chichester.
- [104] Aldert Vrij. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*. John Wiley & Sons.
- [105] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. 2024. Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 21 (March 2024), 23053–23058. doi:10.1609/aaai.v38i21.30349
- [106] Joseph B. Walther, Tracy Loh, and Laura Granka. 2005. Let Me Count the Ways: The Interchange of Verbal and Nonverbal Cues in Computer-Mediated and Face-to-Face Affinity. *Journal of Language and Social Psychology* 24, 1 (March 2005), 36–65. doi:10.1177/0261927X04273036
- [107] Mika Westerlund. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9, 11 (Jan. 2019), 39–52. doi:10.22215/TIMREVIEW/1282
- [108] Jeanne M. Wilson, Susan G. Straus, and Bill McEvily. 2006. All in Due Time: The Development of Trust in Computer-Mediated and Face-to-Face Teams. *Organizational Behavior and Human Decision Processes* 99, 1 (Jan. 2006), 16–33. doi:10.1016/j.obhdp.2005.08.001
- [109] Bowen Yi, Da Shi, and Gang Li. 2026. Real Me or Digital Me? Consumers' Consumption Responses to Online Avatars. *International Journal of Hospitality Management* 132 (Jan. 2026), 104394. doi:10.1016/j.ijhm.2025.104394
- [110] Dale E. Zand. 1997. *The Leadership Triad: Knowledge, Trust, and Power*. Oxford University Press.
- [111] Yuke Zhou and Ning Jia. 2023. The Impact of Item Difficulty on Judgment of Confidence—A Cross-Level Moderated Mediation Model. *Journal of Intelligence* 11, 6 (June 2023), 113. doi:10.3390/jintelligence11060113
- [112] Zoom. 10/20/2023 11:36:03 AM. Zoom Virtual Backgrounds, Filters, and Virtual Avatars. <https://www.zoom.com/en/products/virtual-meetings/features/backgrounds-filters/>.

## A Appendix

### A.1 Statistical models

In the following, we provide the complete regression tables referenced in the results section, offering a comprehensive overview of all model specifications and coefficients presented in our analysis. Our statistical reporting draws on four different models: We fitted a linear mixed model (formula: outcome AI mediation, estimated using REML and nloptwrap optimizer) to predict the outcome (reported trust, truth judgment rate, judgment accuracy and confidence) based on the type of AI mediation (original, retouches, or avatar-based) with a per-subject random fixed effect in the homogeneous environment in Study 1. The details are reported in Table 1. We calculated an equivalent model for the mixed environment in Study 2 only, with statistics reported in Table 2. We then fitted a linear mixed model across studies (estimated using REML and nloptwrap optimizer) to predict the outcome based on the interaction of the AI mediation and environment type (formula: outcome environment + environment:ai-mediation) with a per-subject random fixed effect across both studies. The details are reported in Table 3. Finally, we fitted an extended version of the previous model that also included covariates for participant age, gender, education, race, and experience with AI. The model details are reported in Table 4.

**Table 1: Study 1 Linear mixed model predicting the outcome based on mediation type with a per-participant random fixed effect**

	Trust	Truth	Accuracy	Confidence
(Intercept)	0.510*** (0.008)	0.599*** (0.011)	0.523*** (0.011)	0.564*** (0.010)
MediationRetouch	-0.006 (0.012)	-0.017 (0.016)	-0.013 (0.016)	0.006 (0.014)
MediationAvatar	-0.025* (0.012)	0.017 (0.016)	0.017 (0.016)	-0.012 (0.014)
SD (Intercept Subject)	0.116	0.000	0.000	0.160
SD (Observations)	0.235	0.490	0.499	0.197
Num.Obs.	5802	5802	5802	5802
R2 Marg.	0.002	0.001	0.001	0.001
R2 Cond.	0.197			0.398
AIC	571.3	8219.8	8435.6	-808.1
BIC	604.6	8253.1	8469.0	-774.8
ICC	0.2			0.4
RMSE	0.22	0.49	0.50	0.18

Note. \* p <0.05, \*\* p <0.01, \*\*\* p <0.001.

**Table 2: Study 2 Linear mixed model predicting the outcome based on mediation type with a per-participant random fixed effect**

	Trust	Truth	Accuracy	Confidence
(Intercept)	0.499*** (0.006)	0.597*** (0.011)	0.541*** (0.011)	0.575*** (0.006)
VideoTypeRetouch	-0.021** (0.008)	-0.021 (0.015)	-0.019 (0.015)	-0.019** (0.006)
VideoTypeAvatar	-0.080*** (0.008)	-0.020 (0.015)	-0.007 (0.015)	-0.046*** (0.006)
SD (Intercept Subject)	0.092	0.000	0.045	0.146
SD (Observations)	0.246	0.493	0.497	0.207
Num.Obs.	6240	6240	6240	6240
R2 Marg.	0.016	0.000	0.000	0.005
R2 Cond.	0.137		0.008	0.335
AIC	851.4	8912.1	9061.0	-482.1
BIC	885.1	8945.8	9094.7	-448.4
ICC	0.1		0.0	0.3
RMSE	0.24	0.49	0.50	0.19

Note. \* p <0.05, \*\* p <0.01, \*\*\* p <0.001.

**Table 3: Study 1 and 2 Linear mixed model predicting the outcome based on mediation and environment type with a per-participant random fixed effect**

	Trust	Truth	Accuracy	Confidence
(Intercept)	0.510*** (0.008)	0.599*** (0.011)	0.523*** (0.011)	0.564*** (0.009)
ConditionRetouch	-0.006 (0.011)	-0.017 (0.016)	-0.013 (0.016)	0.006 (0.014)
ConditionAvatar	-0.025* (0.011)	0.017 (0.016)	0.017 (0.016)	-0.012 (0.014)
ConditionMixed	-0.011 (0.010)	-0.002 (0.015)	0.017 (0.016)	0.011 (0.011)
ConditionMixed × VideoTypeRetouch	-0.021** (0.007)	-0.021 (0.015)	-0.019 (0.015)	-0.019** (0.006)
ConditionMixed × VideoTypeAvatar	-0.080*** (0.007)	-0.020 (0.015)	-0.007 (0.015)	-0.046*** (0.006)
SD (Intercept Subject)	0.104	0.000	0.000	0.153
SD (Observations)	0.241	0.492	0.499	0.202
Num.Obs.	12 042	12 042	12 042	12 042
R2 Marg.	0.014	0.001	0.000	0.004
R2 Cond.	0.169			0.366
AIC	1438.6	17 128.1	17 493.6	-1278.2
BIC	1497.8	17 187.3	17 552.8	-1219.1
ICC	0.2			0.4
RMSE	0.23	0.49	0.50	0.19

Note. \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001.

**Table 4: Robustness check Study 1 and 2 Linear mixed model predicting each outcome from mediation and environment type with a per-participant random intercept and controls for demographic and experience covariates**

	Trust	Truth	Accuracy	Confidence
(Intercept)	0.347***	0.546***	0.565***	0.445***
ConditionRetouch	-0.006	-0.019	-0.011	0.003
ConditionAvatar	-0.027*	0.014	0.018	-0.020
ConditionMixed	-0.012	-0.002	0.019	0.009
AgeNum	0.000	0.001	0.000	0.000
GenderMale	-0.005	-0.002	-0.003	0.024**
GenderNon-binary	-0.014	0.021	-0.004	-0.051
EducationNum	0.005*	0.008*	0.002	0.004
RaceBlack or African American	0.045**	0.001	0.013	0.074***
	(0.015)	(0.021)	(0.022)	(0.018)
RaceIndigenous or Native	0.074*	0.123**	-0.007	0.094**
	(0.031)	(0.044)	(0.045)	(0.036)
RaceMiddle Eastern or North African	0.051	0.100	0.125	0.049
	(0.063)	(0.092)	(0.093)	(0.075)
RaceMultiracial or Mixed Race	0.039*	-0.010	0.017	-0.011
	(0.019)	(0.028)	(0.029)	(0.023)
RacePacific Islander	0.097	0.158	0.057	0.119
	(0.081)	(0.117)	(0.119)	(0.096)
RaceWhite or Caucasian	0.058***	0.015	0.025	0.044**
	(0.013)	(0.019)	(0.020)	(0.016)
EnglishLevelNum	0.002	-0.071	-0.064	-0.068
	(0.033)	(0.047)	(0.048)	(0.039)
ExperienceVideoNum	0.036**	-0.003	0.012	0.043**
	(0.012)	(0.018)	(0.018)	(0.015)
ExperienceAI Num	0.014	0.014	0.000	0.075***
	(0.015)	(0.021)	(0.021)	(0.017)
AIinteractionNum	0.032*	-0.002	0.030	0.081***
	(0.014)	(0.020)	(0.020)	(0.016)
AITrustNum	0.090***	0.104***	-0.039	0.058***
	(0.014)	(0.021)	(0.021)	(0.017)
ConditionMixed × VideoTypeRetouch	-0.021**	-0.021	-0.019	-0.019**
	(0.007)	(0.015)	(0.015)	(0.006)
ConditionMixed × VideoTypeAvatar	-0.080***	-0.020	-0.007	-0.046***
	(0.007)	(0.015)	(0.015)	(0.006)
SD (Intercept Subject)	0.098	0.000	0.000	0.142
SD (Observations)	0.241	0.491	0.499	0.202
Num.Obs.	12 042	12 042	12 042	12 042
R2 Marg.	0.034	0.006	0.001	0.055
R2 Cond.	0.171			0.368
AIC	1445.1	17 202.4	17 616.2	-1372.1
BIC	1630.0	17 387.3	17 801.1	-1187.2
ICC	0.1			0.3
RMSE	0.23	0.49	0.50	0.19

Note. \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001.