SC|M
Studies in Communication and Media

FULL PAPER

Spotting fakes: How do non-experts approach deepfake video detection?

Fälschungen feststellen: Wie können Nicht-Experten Deepfake-Videos erkennen?

Mary Holmes, Klaire Somoray, Jonathan D. Connor, Darcy W. Goodall, Lynsey Beaumont, Jordan Bugeja, Isabelle E. Eljed, Sarah Sai Wan Ng, Ryan Ede & Dan J. Miller

Mary Holmes, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0001-0764-8024

Klaire Somoray (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology & Margaret Roderick Centre for Mental Health Research, 4814, Townsville, Australia. Contact: klaire.somoray@jcu.edu.au. ORCID: https://orcid.org/0000-0001-7521-1425

Jonathan D. Connor (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. Contact: jonathan.connor@jcu.edu.au. ORCID: https://orcid.org/0000-0003-3246-8858

Darcy W. Goodall, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.

Lynsey Beaumont, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.

Jordan Bugeja, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.

Isabelle E. Eljed, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0005-6597-0440

Sarah Sai Wan Ng, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0005-5154-8249

Ryan Ede, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.

Dan J. Miller (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology & Margaret Roderick Centre for Mental Health Research, 4814, Townsville, Australia. Contact: daniel.miller1@jcu.edu.au. ORCID: https://orcid.org/0000-0002-3230-2631

FULL PAPER

Spotting fakes: How do non-experts approach deepfake video detection?

Fälschungen feststellen: Wie können Nicht-Experten Deepfake-Videos erkennen?

Mary Holmes, Klaire Somoray, Jonathan D. Connor, Darcy W. Goodall, Lynsey Beaumont, Jordan Bugeja, Isabelle E. Eljed, Sarah Sai Wan Ng, Ryan Ede & Dan J. Miller

Abstract: Intervening to bolster human detection of deepfakes has proven difficult. Little is known about the behavioural strategies people employ when attempting to detect deepfakes. This paper reports two studies in which non-experts completed a deepfake detection task. As part of the task, participants were presented with a series of short videos – half of which were deepfakes – and asked to categorise each video as either deepfake or authentic. In Study 1 ($N = 391$), an online study, participants were randomly assigned to a control or intervention group (in which they received a list of detection strategies before the detection task). After the detection task, participants elaborated on the approach they employed during the task. In Study 2 ($N = 32$), a laboratory-based study, participants' gaze behaviour (fixations and saccades) was recorded during the detection task. No detection strategies were provided to Study 2 participants. Consistent with prior research, Study 1 participants showed modest detection accuracy ($M = .61$, $SD = .14$) – only somewhat above chance levels (.50) – with no difference between the intervention and control groups. However, content analysis of participants' self-reports revealed that the intervention successfully shifted participants' attention toward cues such as skin texture and facial movements, while the control group more frequently reported relying on intuition (gut feeling) and features such as body language. Study 2 found similar levels of detection accuracy ($M = .65$, $SD = .20$). Participants focused their gaze primarily on the eyes and mouth rather than the body, showing a slight preference for the eyes over the mouth. No differences in gaze were found between authentic and deepfake videos or between correctly and incorrectly categorised videos. The findings suggest interventions can modify detection behaviours (even without improving accuracy). Future interventions may benefit from directing attention from the eyes toward more diagnostic features, such as face–body inconsistencies and the face boundary.

Keywords: Deepfakes, AI-generated media, synthetic media, detection, self-report, eye-tracking

Zusammenfassung: Es hat sich als schwierig erwiesen, Maßnahmen zur Verbesserung der menschlichen Erkennung von Deepfakes zu ergreifen. Über die Verhaltensstrategien, die Menschen bei der Erkennung von Deepfakes anwenden, ist nur wenig bekannt. Dieser Artikel präsentiert zwei Studien, in denen Nicht-Experten eine Deepfake-Erkennungsaufgabe absolvierten. Im Rahmen dieser Aufgabe wurde den Teilnehmern eine Reihe von kurzen Videos gezeigt – von denen die Hälfte Deepfakes waren – und sie wurden gebeten, jedes Video entweder als Deepfake oder als authentisch zu kategorisieren. In Studie 1 ($N = 391$), einer Online-Studie, wurden die Teilnehmer nach dem Zufallsprinzip einer Kontroll- oder Interventionsgruppe zugewiesen (in der sie vor der Erkennungsaufgabe eine Liste mit Erkennungsstrategien erhielten). Nach der Erkennungsaufgabe erläuterten die Teilnehmer den Ansatz, den sie während der Aufgabe verwendet hatten. In Studie 2 ($N = 32$), einer Laborstudie, wurde das Blickverhalten (Fixationen und Sakkaden) der Teilnehmer während der Erkennungsaufgabe aufgezeichnet. Den Teilnehmern von Studie 2 wurden keine Erkennungsstrategien zur Verfügung gestellt. In Übereinstimmung mit früheren Untersuchungen zeigten die Teilnehmer der Studie 1 eine mäßige Erkennungsgenauigkeit ($M = 0{,}61$, $SD = 0{,}14$) – nur geringfügig über dem Zufallsniveau (0,50) – ohne Unterschied zwischen der Interventions- und der Kontrollgruppe. Die Inhaltsanalyse der Selbstauskünfte der Teilnehmer ergab jedoch, dass die Interventionsgruppe ihre Aufmerksamkeit erfolgreich auf Hinweise wie Hautstruktur und Gesichtsbewegungen lenkte, während die Kontrollgruppe häufiger angab, sich auf ihre Intuition (Bauchgefühl) und Merkmale wie Körpersprache zu verlassen. Studie 2 ergab eine ähnliche Erkennungsgenauigkeit ($M = 0{,}65$, $SD = 0{,}20$). Die Teilnehmer richteten ihren Blick hauptsächlich auf die Augen und den Mund und weniger auf den Körper, wobei sie eine leichte Präferenz für die Augen gegenüber dem Mund zeigten. Es wurden keine Unterschiede im Blickverhalten zwischen authentischen und Deepfake-Videos oder zwischen korrekt und falsch kategorisierten Videos festgestellt. Die Ergebnisse deuten darauf hin, dass Interventionen das Erkennungsverhalten verändern können (ohne die Genauigkeit zu verbessern). Zukünftige Interventionen könnten davon profitieren, die Aufmerksamkeit von den Augen auf diagnostischere Merkmale wie Inkonsistenzen zwischen Gesicht und Körper und die Gesichtskonturen zu lenken.

Schlagwörter: Deepfakes, KI-generierte Medien, synthetische Medien, Erkennung, Selbstauskunft, Eye-Tracking

1. Introduction

Deepfakes are a form of AI-manipulated media in which an existing person's likeness is inserted into an extant piece of media (be it a static image, piece of audio, or a video). They can be highly realistic. The most common type of deepfakes are "face replacement" deepfakes (Silva et al., 2022). Deepfakes can be used to make it appear as if someone has done or said something they have never done or said. As such, deepfake technology, when used maliciously, can cause serious harms. These harms can occur at the individual and societal level. Examples of individual-level harms include scams (Miller et al., 2025) and the use of non-consensual digitally altered sexual imagery for harassment and extortion (Flynn et al., 2022). Potential societal-level harms include the spread of disinformation and misinformation, manipulation of political campaigns and public opinion, the erosion of trust in democratic institutions and legitimate media reporting (Godulla et al., 2021), and military deception (Smith & Mansted, 2020).

Automated deepfake detection tools have advanced significantly (Abbas & Taeihagh, 2024). However, these technologies are still generally inaccessible to the public. Furthermore, political climate can influence the implementation of these technologies, as demonstrated by Meta's recent decision to discontinue third-party fact-checking on Facebook, Threads, and Instagram (McMahon et al., 2025). Thus, the public is typically still left to their own devices to verify the digital content they consume. There is a serious need to develop behavioural interventions to mitigate the adverse impacts of AI-created content, such as deepfakes (eSafety Australia, 2022; World Economic Forum, 2024).

To date, the development of effective deepfake detection interventions has been hampered by our lack of knowledge of the strategies and processes people employ when attempting deepfake detection. Very little research has examined the specific approaches – conscious or unconscious – that individuals adopt during deepfake detection. Without this foundational knowledge, interventions may be poorly aligned with natural detection behaviours.

The current research sought to address this gap by collecting self-report and eye-tracking data while participants knowingly engaged in a video-based deepfake detection task. This approach provides a more ecologically valid representation of how members of the public process potentially manipulated content when actively searching for deception. To this end, we conducted two complementary studies. Study 1 replicated and extended Somoray and Miller (2023) – discussed below – using an alternative recruitment method. It aimed to further evaluate the efficacy of Somoray and Miller's (2023) passive, visual-anomaly-focused intervention and to examine participants' self-reported strategies for detecting deepfakes. Study 2 investigated implicit detection processes by using eye-tracking methods to directly measure participants' gaze during a deepfake detection task.

2. Literature review

Meta-analytic evidence indicates that the general public typically performs at chance levels on deepfake detection tasks across media modalities, including video (Diel, Lalgi, et al., 2024). Various detection interventions have been developed and tested to improve the public's ability to discern deepfakes (for an overview, see Somoray et al., 2025). Interventions can vary in both focus (e.g., identifying common visual and/or auditory anomalies [also called "artifacts"], increasing motivation to perform well, or assessing the plausibility of message content) and level of interactivity (passive interventions vs. more active interventions in which feedback on performance is provided).

Attempts to increase detection by bolstering motivation have generally proven ineffective (Somoray et al., 2025). For instance, Köbis et al. (2021) found that raising awareness about the dangers of deepfakes or offering cash incentives for correct detections did not enhance detection accuracy. This suggests that the inability to detect deepfakes reflects a skill deficit, rather than a lack of motivation to perform well.

Active interventions have shown some promise. Feedback-based interventions have been found to improve detection for static images (Diel, Teufel, & Bäuerle, 2024; Robertson et al., 2018). However, other studies have failed to replicate these findings when using higher-quality stimuli (Kramer et al., 2019). Tailored media literacy lectures have also been shown to impact perceptions of deepfake video credibility (El Mokadem, 2023), and one-on-one “walk-through” examples have been employed successfully to enhance video detection accuracy (Tahir et al., 2023).

By comparison, passive intervention approaches appear to be less effective. For example, Somoray and Miller (2023) tested a written, visual-anomaly-focused intervention adapted from detection advice provided by the MIT Media Lab. They found the proportion of videos correctly identified on a detection task to be nearly identical in their control (60%) and intervention group (61%). However, these null findings may partly reflect their recruitment method: Via a post on Reddit. If participants happened to share detection strategies in the post thread, this would have “washed out” the effect of their intervention. Bray et al. (2023) and Kramer et al. (2019, Study 2) similarly found that providing anomaly-based detection advice (either once or repeatedly) did not improve detection for static images.

Passive, anomaly-based interventions are simple and scalable, making them attractive options for use in public safety campaigns. However, they currently lack demonstrated efficacy. Refining such interventions requires a clearer understanding of the behaviours people engage in during deepfake detection. Eye-tracking studies have the potential to elucidate this issue. Yet, existing eye-tracking studies in this domain have methodological limitations that may constrain the insights they offer into video detection behaviours. Many have relied on still image stimuli (Caporusso et al., 2020; Cartella et al., 2024) or video stimuli viewed by participants naïve to the study’s purpose (Gupta et al., 2020; Wöhler et al., 2021). Tahir et al. (2021) did incorporate videos in a detection task, but tracked gaze only in relation to static screenshots, not during dynamic viewing. Study 2 in the present research sought to address these issues by recording eye-tracking data during dynamic viewing of video stimuli.

3. Study 1

3.1 Method

3.1.1 Design

Study 1 employed an online between-subjects experimental design in which participants were randomly assigned to either an intervention (receiving a list of written detection strategies) or control condition, before completing a deepfake detection task. The Human Research Ethics Committee of James Cook University granted ethical approval to conduct the study. The study was preregistered on OSF (https://osf.io/vutb8) on April 21, 2023, before data collection.

3.1.2 Stimulus materials and measures

The same written detection strategies were used as in Somoray and Miller (2023). These strategies were sourced from the MIT Media Lab (https://www.media.mit.edu/projects/detect-fakes/overview/). These strategies are provided in the Supplementary Material (Table S1 in OSF file).

The detection task involved the presentation of 20 stimulus videos. The same videos were used as in Somoray and Miller (2023). They were originally sourced from the Deepfake Detection Challenge (DFDC) Dataset (Dolhansky et al., 2020). All videos were 10 seconds in length and featured regular people rather than public figures. Stimulus videos depicted an equal number of male and female models and included models of various skin tones.

Each participant saw exactly 10 deepfake and 10 authentic videos. Before the detection task, participants were informed as to how many videos they would be presented with and what proportion would be deepfakes. Two sets of videos were created. That is, Set A contained the authentic version of Video 1, whereas Set B contained the deepfaked version, et cetera. Participants were randomly assigned to receive either Set A or B. The order of the presentation of videos within sets was randomised to mitigate order effects.

Participants responded to each video with one of two binary options: *This video is a deepfake* or *this video is real*. Detection accuracy was calculated as the number of videos correctly categorised divided by the total number of videos categorised (e.g., correctly categorising 13 out of 20 videos would give a detection accuracy score of .65). After the detection task, participants were presented with an open-ended question asking what strategies they employed during the task. The wording of this question differed between conditions: Control condition = "What strategy/s did you use when doing the detection activity?"; intervention condition = "Which, if any, of the strategies provided at the beginning of the experiment helped you the most during the detection activity? Additionally, what other strategy/s, if any, did you use during the detection activity?" Participants were also asked about their perceptions of their susceptibility to deepfake-based scams and misinformation. These findings are reported elsewhere (Dornbusch et al., 2025).

3.1.3 Procedure

Following Somoray and Miller (2023), participants were randomly assigned to either the intervention (provided with a list of written detection tips) or control condition. Participants were then given information about the detection task and presented with two comprehensive check questions, which they were required to answer correctly before they could start the detection task. These questions concerned the definition of deepfakes and the proportion of deepfaked videos in the detection task video set (50%). They were also informed that, at the end of the study, they would receive a score reflecting the number of videos they correctly categorised. Participants then completed the detection task before being asked to provide demographic information. Participants were able to watch each video as many times as they wished. After the detection task, participants were debriefed and provided with their detection accuracy score.

3.1.4 Recruitment and Participants

Participants were recruited via a student participation scheme at the authors' institution and by sharing the study via the authors' professional networks and snowball recruitment. Student participants were provided with course credit in exchange for their participation. Recruitment occurred from April 2023 to February 2024.

The study was accessed by 474 people. Participant data were removed if participants: 1) did not provide consent, 2) did not attempt the detection task, 3) spent on average under 10 seconds watching each stimulus video, or 4) indicated that this was not their first time participating in the study. This left a final sample of 391 participants. Demographic characteristics of the sample are reported in Table 1.

Table 1. Participant demographics for Study 1 (*N* = 391) and Study 2 (*N* = 32)

Variable	Study 1	Study 2
	M (*SD*)	
Age	25.80 (10.24)	26.32 (7.95)
	n (%)	
Gender		
Male	116 (29.7)	11 (34.4%)
Female	271 (69.3%)	21 (65.6%)
Non-binary	3 (0.8%)	-
Country of residence		
Australia	213 (54.5%)	32 (100.0%)
Singapore	159 (40.7%)	-
China	8 (2.0%)	-
Other countries	11 (2.8%)	-
Highest level of education		
High school graduate	166 (42.5%)	12 (37.5%)
TAFE/other vocational studies	43 (11.0%)	4 (12.5%)
Undergraduate degree	137 (35.0%)	6 (18.8%)
Some postgraduate study or a postgraduate degree	45 (11.5%)	10 (31.3%)

3.1.5 Codebook development

Quantitative content analysis was used to analyse responses to the open-ended question. A codebook was developed to facilitate this process. Initially, three investigators independently coded 10% of responses while blinded to the experimental condition, generating potential coding categories (e.g., *voice*, *blur*, *gut feeling*) and organising these into putative groupings (e.g., *visual artefacts*, *feeling*). The investigators then met to develop a pilot codebook containing groupings, codes, definitions, and examples. To test the codebook's reliability, two authors independently coded an additional 10% of responses. The coders had a 75% agreement in categorising these responses, demonstrating moderate intercoder reliability (Burla et al., 2008). Following this, the raters met to make necessary modifications to the pilot codebook. For example, a code was added (e.g., *skin – general – any mention of wrinkles, blemishes, smoothness or agedness of the skin, without specification as to whether this is on the face or body*). The finalised codebook is provided in the Supplementary Material (Table S2 in OSF file). The remainder of responses were coded by one investigator. To prevent rater drift, coding was completed in blocks with regular codebook review.

3.2 Results

3.2.1 Detection accuracy

In the overall sample, mean detection accuracy was .61 (*SD* = .14), suggesting that participants, on average, correctly identified 12 out of the 20 videos. This is above the degree of accuracy that would be expected by chance alone (.50). The poorest performers correctly categorised 4 out of 20 videos (.20), while the best performers correctly categorised 19 out of 20 videos (.95). The control (*M* = .61, *SD* = .14) and intervention groups (*M* = .60, *SD* = .14) did not differ on detection accuracy, $t(389) = 0.46$, $p = .646$, Cohen's $d = 0.05$.

3.2.2 Content analysis of self-reported detection strategies

Of the 392 participants, 47 did not respond to the open-ended question and were therefore removed from the content analysis. Analysis of participant responses indicated that most participants reported employing more than one strategy. A total of 640 detection strategies were reported across the 345 participants who responded to the open-ended question. Table 2 provides the percentage of participants who reported each type of strategy for the whole sample and broken down by experimental condition. Colour gradient heat-mapping (green for higher values, white for lower values) is used to visualise which strategies were more commonly reported. Across the entire sample, the most frequently reported detection strategy was to look for *visual attributes* (this coding category was defined as "Any mention of shadows, lighting, textures, colours or resolution. This does NOT include glitches or blurring"; for definitions for all codes see Table 2) with just over a third of participants giving a response which could be categorised under this code. Other popular strategies (reported by > 10% of the overall sample) included: *Body movement*; *face movement – eyes*; and *facial features* – eyes.

Table 2. Frequency of self-report strategies in overall Study 1 sample and by condition

Grouping	Code	Codebook description	Overall sample #	Overall sample %	Condition	
				(*N* = 344)	Control % (*n* = 178)	Intervention % (*n* = 166)
Audio artefacts	Audio attribute	Any mention of sound quality that is NOT about background noise or voice.	8	2.3%	3.9%	0.6%
	Background noise	Any mention of background noise or environmental noise.	4	1.2%	1.1%	1.2%
	Voice	Any mention of tone/modulation of voice or accents.	20	5.8%	10.7%	0.6%
Body	Body feature	Any mention of a body feature other than the face and does NOT mention skin. This includes neck, collarbone or hands, “composition” or body.	7	2.0%	1.1%	3.0%
	Body movement	Any mention of abnormality regarding body language, demeanour or posture, body movement or hand gestures.	54	15.7%	24.7%	6.0%
Face	Face expression	Any mention of facial expression. This includes mood, expression, and emotions.	48	14.0%	16.3%	11.4%
	Facial feature - eyes	Any mention of the eye area that does NOT mention movement (e.g., blinking).	48	14.0%	8.4%	19.9%
	Facial feature - general	Any mention of a holistic evaluation of the face that is NOT to do with the expression of emotion.	36	10.5%	10.1%	10.8%
	Facial feature - hair	Any mention of facial hair or hair line. Eyebrows/lashes don’t count as hair.	15	4.4%	1.7%	7.2%
	Facial feature - mouth	Any mention of movement around the mouth, including lips and teeth.	14	4.1%	3.4%	4.8%
	Facial movement - eyes	Any mention of the movement of the eyes, including blinking or eyebrow movement.	52	15.1%	9.6%	21.1%
	Facial movement - general	Any mention of the overall movement of the face such as angle that does NOT specify a particular facial feature.	32	9.3%	9.0%	9.6%
	Facial movement - mouth	Any mention of the movement of the mouth, including teeth, lips.	9	2.6%	3.4%	1.8%
Feeling	Gut feeling	Anything relating to participant’s emotional reaction or feelings. This includes “uncanny valley” “the vibe being off” etc.	19	5.5%	10.1%	0.6%
Skin	Discrepancy of the skin on face and body	Any mention of discrepancy between the skin on face and body. For example, any mention of the agedness of the skin not matching across face and body.	1	0.3%	0.0%	0.6%
	Skin - face	Any mention of wrinkles, blemishes (e.g., moles), smoothness, folding or agedness of the person in the facial area. This does NOT include other parts of the body (see below). The term “complexion” can be considered to be referring to the face.	18	5.2%	1.7%	9.0%
	Skin - general	Any mention of wrinkles, blemishes, smoothness or agedness of the skin, without specification as to whether this is on the face or body.	28	8.1%	1.1%	15.7%
Sync	Sync between voice and mouth movement	Any mention of the discrepancy between the synchronisation of the voice/speech and mouth movement. For example, any mention of match or mismatch of voice/speech and mouth movement.	17	4.9%	7.3%	2.4%
Visual artefacts	Blur	Any mention of blurring or smoothness, including the face.	15	4.4%	3.9%	4.8%
	Glitch	Any mention of glitching, including the face.	38	11.0%	14.0%	7.8%
	Visual attribute	Any mention of shadows, lighting, textures, colours or resolution. This does NOT include glitches or blurring.	123	35.8%	28.7%	43.4%
Other	None/unsure	Code for people who said no, unsure, or N/A.	16	4.7%	1.7%	7.8%
	Not covered by above codes	Participant gives a response not covered by an existing coding unit.	9	2.6%	3.4%	1.8%
	Unclear response	Response is unclear or ambiguous.	8	2.3%	4.5%	0.0%

As seen in Table 2, differences between the control and intervention group were observed for some codes. Compared to participants in the control condition, participants in the intervention group more frequently reported engaging in strategies falling under the following codes: *Visual attribute* (control = 28.7%, intervention = 43.4%); *skin – general* (control = 1.1%, intervention = 15.7%); *facial features – eyes* (control = 8.4%, intervention = 19.9%); *facial movement – eyes* (control = 9.6%, intervention = 21.1%); and *skin – face* (control = 1.7%, intervention = 9.0%). In contrast, the control group more frequently reported strategies falling under codes such as *body movement* (control = 24.7%, intervention = 6.0%); *voice* (control = 10.7%, intervention = 0.6%); and *gut feeling* (control = 10.1%, intervention = 0.6%).

3.3 Discussion

In Somoray and Miller (2023) the intervention group – who received a list of strategies they could apply to aid themselves in the detection task – did not outperform the control group. There are a number of possible reasons for this lack of an effect, including 1) recruitment via social media platforms undermining the validity of the experimental manipulation (e.g., if information was shared to the control group in discussion threads), 2) the detection guidance provided to participants being ineffective (e.g., incorrect or difficult to apply), or 3) intervention-group participants choosing not to apply the strategies outlined in the provided detection guidance.

The overall samples' detection accuracy in the current study was virtually identical to that observed in Somoray and Miller (2023) – Study 1: *M* = .61, *SD* = .14; Somoray and Miller (2023): *M* = .61, *SD* = .13. Consistent with Somoray and Miller (2023), the Study 1 intervention group performed almost identically to the control group. This suggests that the lack of an experimental effect observed in Somoray and Miller (2023) is not solely attributable to the authors' recruitment approach.

Content analysis of participants' self-reports does suggest that the intervention influenced participants' behaviours. The intervention group appeared to focus on areas reflective of those highlighted in the detection tips they were provided with. For instance, the intervention group were more likely to self-report examining the skin on the models' faces for anomalies, reflecting one of the detection strategies ("Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? Deepfakes are often incongruent on some dimensions."). In contrast, participants in the control group were more likely to self-report relying on their "gut feeling" or irrelevant features such as the model's body language (a likely ineffective strategy, given that deepfakes are typically face manipulations). This suggests that the non-significant results observed in Study 1 and in Somoray and Miller (2023) were not due to participants in the intervention group simply ignoring the detection strategies provided to them. This casts doubt on whether these strategies are fit for purpose.

4. Study 2

4.1 Method

4.1.1 Design

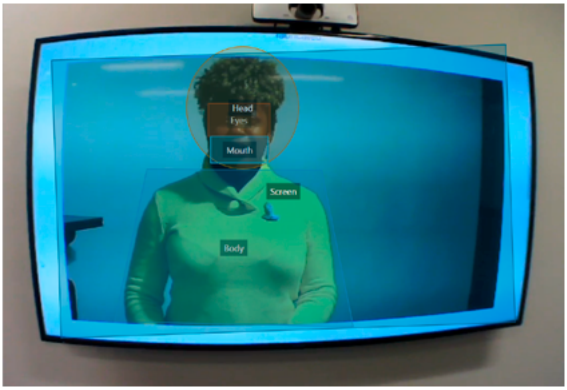
Study 2 was an in-person laboratory study in which participants completed a detection task while their gaze behaviours were recorded. Unlike in Study 1, a detection intervention was not introduced. The Human Research Ethics Committee of James Cook University granted ethical approval to conduct the study.

4.1.2 Materials, measures, and apparatus

The detection task was similar to that used in Study 1. This time, however, five practice trial videos were presented prior to the presentation of ten detection task videos. As in Study 1, stimulus videos were sourced from the DFDC dataset, although the specific videos used differed. Following the Study 1 procedure, two sets of videos were created, and the order in which videos were presented within sets was randomised. The videos depicted models of various genders and skin tones. All models were non-public figures.

The same detection accuracy index was used as in Study 1. Three gaze variables were analysed as part of this study: Average fixation duration (the average duration of participants' fixations, measured in milliseconds), fixation count (the frequency with which participants fixated their gaze), and saccade count (the frequency with which participants made saccades, i.e., shifted their gaze between fixations). These variables were recorded in relation to five areas of interest (AOIs): The screen, the stimulus model's body, the stimulus model's head, the stimulus model's eye area, and the stimulus model's mouth area. However, we report results only for the eyes, mouth, and body AOIs (as the eyes and mouth AOIs are situated within the head AOI and all other AOIs are situated within the screen AOI). These AOIs are depicted in Figure 1.

Figure 1. An example of the areas of interest (eyes, mouth, body, head, and screen) created during data processing. Image representative of participants' field of view.



Eye-tracking information was recorded using Pupil Labs' *Pupil Invisible* model glasses. These glasses fit like normal prescription eyeglasses, allowing for naturalistic movement. They record the movement of each eye. The specifications of this equipment are provided in the Supplementary Material hosted on OSF. Participants completed the study sitting in a chair approximately 57cm away from a 70cm HD TV screen. The stimulus videos took up most of the screen. Figure S1 (in OSF file) depicts the experimental setup.

4.1.3 Procedure

At the start of the study, participants were informed that exactly half of the detection task videos were deepfakes and that they would receive a detection accuracy score at the study's conclusion. Participants then completed the same comprehension check questions as in Study 1. The eye-tracking glasses were then calibrated to the participant, and the five practice trial videos were presented. Following each practice trial, participants received feedback indicating whether their detection decision was correct or incorrect. Participants then completed the detection task. Unlike in Study 1, participants were not permitted to watch stimulus videos more than once. Following the detection task, participants completed demographic questions and received a debriefing that included their detection accuracy score. Investigators read from a pre-established script when explaining the study procedure to participants.

4.1.4 Recruitment and participants

As in Study 1, participants were recruited via undergraduate student recruitment channels (in exchange for course credit) and snowball recruitment within the researchers' personal and professional networks. Recruitment occurred from December 2023 to August 2024. Student participants were offered course credit for their participation, and non-student participants were entered into a prize draw for a gift card. Those who require eyeglasses for up-close work were excluded (unless wearing contact lenses), as the eye-tracking glasses do not fit comfortably over regular eyeglasses. Demographic characteristics of the sample are reported in Table 1.

4.1.5 Data processing

Eye-tracking data was processed using Pupil Labs' iMotions 10 software. This involved manually creating AOIs and moving these to match the movement of the stimulus video model (e.g., moving the eyes AOI to the left as the stimulus video model moved their head to the left of screen). This was done for all AOIs, for all 10 detection task videos, for each participant. Practice trial videos were excluded from this process, as this data was not included in the analysis. Further technical details of the data processing are provided in the Supplementary Material in OSF.

4.2 Results

Mean detection accuracy in the overall sample was .65 (*SD* = .20), indicating that participants, on average, correctly identified 6.5 out of 10 videos. This is above the degree of accuracy that would be expected by chance alone (.50). The best performer correctly categorised all 10 videos (1.00), while the worst performer correctly categorised 2 videos only (.20).

Table 3 presents descriptive statistics for average fixation duration, fixation count, and saccade count broken down by AOI (body, eyes, mouth), video authenticity (deepfake or authentic), and decision (correct or incorrect categorisation of video) for the overall sample. The table suggests that participants' visual attention was directed predominantly towards the eyes and mouth, rather than towards the body.

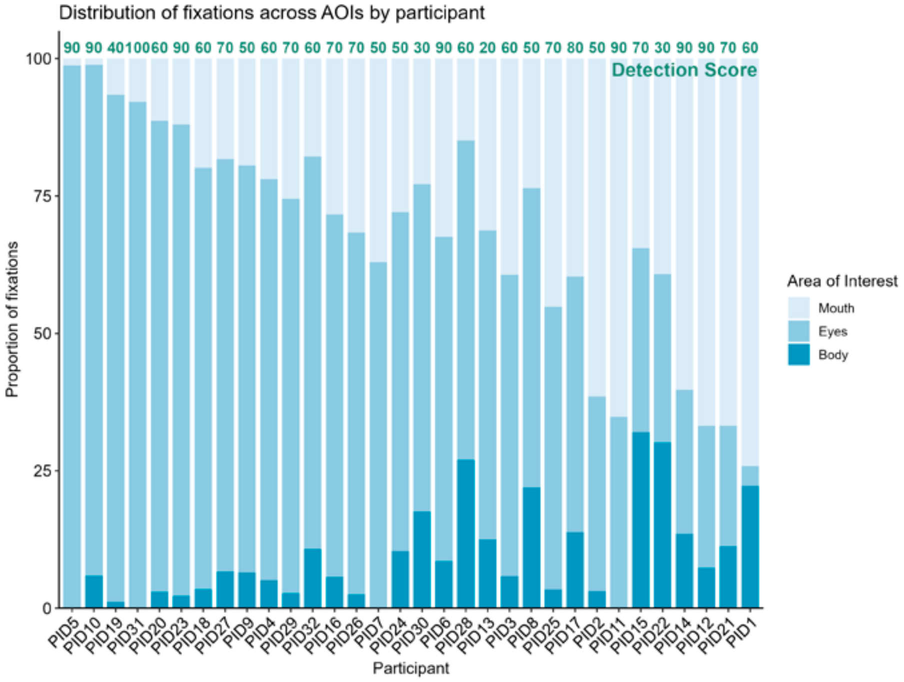
Table 3. Descriptive statistics for gaze variables across Study 2 sample (*N* = 32) by video authenticity, decision, and area of interest (AOI)

AOI	All Videos	Video authenticity		Video decision	
		Deepfake	Authentic	Correct	Incorrect
Average fixation duration (ms)					
Body	85.70 (119.19)	84.84 (118.97)	86.57 (119.78)	82.30 (120.72)	92.08 (116.55)
Eyes	395.91 (232.78)	415.29 (232.87)	376.41 (231.77)	387.39 (231.06)	411.86 (236.18)
Mouth	352.14 (263.12)	353.87 (276.02)	350.39 (250.33)	349.95 (256.02)	356.24 (277.07)
Fixation count					
Body	1.61 (2.54)	1.55 (2.63)	1.66 (2.46)	1.45 (2.41)	1.89 (2.76)
Eyes	10.64 (6.73)	11.12 (6.84)	10.15 (6.60)	10.85 (7.07)	10.24 (6.04)
Mouth	5.77 (5.54)	5.35 (5.44)	6.18 (5.62)	6.29 (5.97)	4.78 (4.48)
Saccade count					
Body	1.93 (3.15)	1.86 (3.11)	2.01 (3.19)	1.76 (3.01)	2.27 (3.38)
Eyes	12.53 (10.58)	12.90 (10.42)	12.15 (10.76)	12.78 (11.00)	12.05 (9.78)
Mouth	6.70 (7.59)	6.30 (7.67)	7.10 (7.51)	7.20 (7.78)	5.76 (7.16)

A 3 (AOI) × 2 (video authenticity) × 2 (video decision) repeated measures ANOVA was conducted for each outcome variable (average fixation duration, fixation count, saccade count). To account for potential interactions, each ANOVA included four interaction terms: AOI × authenticity; AOI × decision; authenticity × decision; and AOI × authenticity × decision. The details of these analyses are provided in the Supplementary Material (Tables S3–S11 in OSF file). These analyses indicated that participants had significantly longer fixations when looking at the eyes and mouth, relative to the body ($p < .001$ in both cases). The difference in average fixation length between the eyes and mouth AOIs was non-significant ($p > .999$). Further, participants made significantly more fixations on the eyes than the mouth (p = .033) or body (p < .001). They also fixated more frequently on the mouth than the body (p < .001). Similarly, participants engaged in more saccades in the eyes AOI compared to the body AOI (p < .001). Saccades were more frequent in the mouth AOI compared to the body AOI (p = .006) but not the eyes AOI (p = .122). All reported p-values have been Bonferroni corrected. Average fixation duration, fixation count, and saccade count did not differ between correctly or incorrectly categorised videos or between deepfake and authentic videos.

While, on average, the eyes tended to attract the most visual attention, there did appear to be individual differences around this. In Figure 2, it can be seen that some participants focused almost exclusively on the eyes, some focused almost exclusively on the mouth, and others spent a roughly equal amount of time on each AOI. Detection accuracy was unrelated to proportion of time spent looking at the eyes, $r(30)$ = .10, p = .603, mouth, $r(30)$ = .04, p = .826, or body, $r(30)$ = -.34, p = .055. In the latter case, results are bordering on significance, which could suggest that those who spent more time looking at the body tended to perform worse on the detection task.

Figure 2. Distributions of fixations across areas of interest, along with detection accuracy scores for all participants



4.3 Discussion

Study 2 suggests that, when trying to determine the authenticity of videos, participants show a strong preference for looking at eyes of stimulus models rather than the body, and a moderate preference for the eyes rather than the mouth. Participants' apparent focus on the eyes is somewhat inconsistent with past studies, which have found that attention is often directed away from the eyes and towards other regions of the face when participants view high-quality deepfakes (Wöhler et al., 2021). This said, visualisation of the data (Figure 2) suggests that there was a subset of participants who adopted a "mouth-focused" approach. Detection accuracy was unrelated to proportion of time spent looking at the eyes and proportion of time spent looking at the mouth. Spending a greater proportion of time looking at the body may be associated with poorer detection performance.

Participants exhibited similar gaze patterns regardless of whether they correctly or incorrectly categorised videos, as evidenced by the lack of main effects for video decision. Participants may have employed a consistent visual search strategy across all videos – such as rapidly scanning the eye area for anomalies – with variable success depending on the presence and detectability of visual cues (i.e., some videos may contain obvious anomalies that others do not).

Participants exhibited similar gaze patterns when viewing authentic versus deepfake videos, as evidenced by the absence of main effects for video authenticity. This indicates that participants did not subconsciously modify their visual behaviour in response to deepfake content, at least not in ways captured by our gaze measurements. These findings contrast with previous research (Gupta et al., 2020; Wöhler et al., 2021), which documented distinct gaze patterns when participants unknowingly viewed deepfake videos. A critical methodological difference may explain this discrepancy: Unlike previous studies, participants in our experiment were explicitly aware they were in a deepfake detection task. This awareness may have resulted in participants adopting a more deliberate visual search strategy, which they actively applied to all videos.

5. Overall discussion

These studies sought to investigate laypeople's behaviour when faced with the problem of trying to identify deepfake videos. This was done through the analysis of participants' self-reports of the strategies they employed on a deepfake detection task (Study 1) and gaze data collected during a detection task (Study 2). Many of the findings are relevant to those seeking to design better deepfake detection training modules.

First, both studies corroborate prior research indicating that deepfake detection is difficult for most individuals (Diel, Lalgi, et al., 2024; Köbis et al., 2021; Somoray & Miller, 2023), with participants performing only marginally better than chance. Importantly, this poor performance occurred despite participants being explicitly warned that they would encounter deepfakes. For this reason, we should expect "real-world" detection rates to be even lower than those observed in Studies 1 and 2.

Second, Study 1 suggests that the provision of written detection tips is not enough to meaningfully bolster detection rates (as also found in Somoray & Miller, 2023). These null findings are consistent with the results of other studies into the efficacy of passive, anomaly-based detection interventions (Bray et al., 2023; Kramer et al., 2019). However, the findings do indicate that people will shift their behaviour on detection tasks in response to detection instructions. That is, the detection approaches self-reported by participants in the intervention condition showed a greater alignment with the detection instructions than those of the control group. This highlights the possibility of improving the public's detection abilities through passive detection interventions (even if the specific advice tested in Study 1 was itself ineffective).

Third, a consistent finding across both studies is that many people gravitate towards the face region, particularly the eyes, when trying to ascertain video veracity. This may reflect Western cultural norms around eye contact (Senju et al., 2013). Most deepfakes involve face replacement (Silva et al., 2022) – imposing the face of a target person onto a model, while leaving the model's body unadjusted. Thus, a focus on the face is advisable during deepfake detection. However, for this same reason, assessing for discrepancies between the face and body may also be informative (e.g., looking for discrepancies in the agedness of the skin on the face and hands). Future training modules may benefit from overtly directing participants towards this strategy, while, ideally, also providing visual examples. It is also worth noting that the eyes may be less diagnostic than other face regions. Areas such as the boundary of the face (which may show visual peculiarities, particularly during head movement) or the lips (which may reveal errors in audio-mouth synchronisation) could provide more reliable cues. Future studies should investigate whether explicitly directing participants' attention away from the eyes and towards other regions of the face improves detection accuracy.

Future research should also investigate the gaze behaviour of deepfake detection experts. Across most domains, experts demonstrate more efficient and selective visual scanning than novices, strategically directing attention to task-relevant areas and maintaining longer fixations on critical information (Brams et al., 2019). The domain of medicine represents a notable exception, where experts exhibit more extensive visual spans. The gaze patterns of superior detectors could reveal optimal visual strategies for deepfake identification. To facilitate the identification of individuals with exceptional detection abilities, population norms should be established by administering standardised video sets to large representative samples.

In interpreting the study's findings, it is important to consider the choice of stimulus videos, which all depicted non-public figures discussing mundane topics. This is both a strength and limitation of the study. It is a strength in that it minimises the influence of prior knowledge or contextual biases, forcing participants to rely on visual and auditory cues. By controlling for these factors, the study provides a clearer picture of deepfake detection behaviour "in a vacuum" and offers insight into how people identify manipulated content when contextual information is limited.

At the same time, the use of non-public figures discussing mundane topics limits ecological validity. In real-world scenarios, videos often feature public figures or emotionally salient messages, where context cues and prior knowledge and attitudes play an important role. For example, when assessing videos of public figures such as politicians, detectors can draw on visual and auditory cues while *also* evaluating whether the message content aligns with what they know of the figure's beliefs ("Would this person ever say something like this?"). Familiarity with the deepfaked subject may even enhance ability to pick up on visual anomalies (Thaw et al., 2020). These factors would likely increase detection performance. Conversely, the use of known figures discussing charged topics may, in some instances, undermine performance. For example, detectors are less likely to correctly identify deepfakes when message content aligns with their existing personal beliefs (Sütterlin et al., 2023). Holding strong prior attitudes towards the deepfaked subject may also influence detection decisions (Ng, 2023).

Several other limitations also warrant consideration. First, the study did not account for individual differences in perceptual expertise that may influence deepfake detection ability. Future research should explore whether factors such as experience with digital media production moderate gaze behaviour and detection accuracy. Second, while this study examined overall gaze patterns, it did not differentiate between deepfakes of varying sophistication. Research suggests that deepfake quality impacts detection performance (see Somoray et al., 2025) and that individuals unconsciously adjust their visual behaviour based on deepfake quality (Wöhler et al., 2021), warranting further investigation of this factor. Finally, in both studies, individuals with a particular interest in deepfakes may have been more inclined to participate, introducing the possibility of sampling bias. If greater familiarity with, or interest in, deepfakes is linked to enhanced detection performance, the sample's performance may have been greater than that of the general population. However, this concern is somewhat mitigated by the recruitment of student participants, who were likely motivated to participate by external factors (e.g., course credit) rather than a specific interest in deepfakes.

GenAI declaration

Generative AI (Claude 4.0 and ChatGPT-5) was used for basic copy-editing.

Supplementary material

A supplementary material file can by found on OSF: https://osf.io/tzpd7/files/osfstorage/6922940d5f3279069d76fc29. All other materials associated with the study can be found on the OSF page for Study 1: https://osf.io/tzpd7.

References

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, *252*. https://doi.org/10.1016/j.eswa.2024.124260

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M., & Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin*, *145*(10), 980–1027. https://doi.org/10.1037/bul0000207

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1). https://doi.org/10.1093/cybsec/tyad011

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research, 57*(2), 113–117. https://doi.org/10.1097/01.NNR.0000313482.33917.7d

Caporusso, N., Zhang, K., & Carlson, G. (2020). Using eye-tracking to study the authenticity of images produced by generative adversarial networks. *2020 International Conference on Electrical, Communication, and Computer Engineering* (ICECCE), 1–6. https://doi.org/10.1109/ICECCE49384.2020.9179472

Cartella, G., Cuculo, V., Cornia, M., & Cucchiara, R. (2024). Unveiling the truth: Exploring human gaze patterns in fake images. *IEEE Signal Processing Letters*, 1–5. https://doi.org/10.1109/LSP.2024.3375288

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*. https://doi.org/10.1016/j.chbr.2024.100538

Diel, A., Teufel, M., & Bäuerle, A. (2024). *Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy*. OSF. https://doi.org/10.31219/osf.io/muwnj

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Cristian Canton Ferrer. (2020). *The DeepFake Detection Challenge (DFDC) Dataset*. arXiv. https://doi.org/10.48550/arxiv.2006.07397

Dornbusch, A., Tye, T., Somoray, K., & Miller, D. J. (2025). *Third person effects and the base-rate fallacy: Cognitive biases in deepfake detection* [Manuscript in preparation].

El Mokadem, S. S. (2023). The effect of media literacy on misinformation and deep fake video detection. *Arab Media & Society*, *35*, 53–78. https://www.arabmediasociety.com/

Flynn, A., Powell, A., Scott, A. J., & Cama, E. (2022). Deepfakes and digitally altered imagery abuse: A cross-country exploration of an emerging form of image-based sexual abuse. *The British Journal of Criminology*, *62*(6), 1341–1358. https://doi.org/10.1093/bjc/azab111

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes: An interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media*, *10*(1), 72–96. https://doi.org/10.5771/2192-4007-2021-1-72

Gupta, P., Chugh, K., Dhall, A., & Subramanian, R. (2020). The eyes know it: FakeET- An eye-tracking database to understand deepfake perception. *Proceedings of the 2020 International Conference on Multimodal Interaction,* 519–527. https://doi.org/10.1145/3382507.3418857

Köbis, N, C., Doležalová, B., & Soraperra., I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11). https://doi.org/10.2139/ssrn.3832978

Kramer, R. S., Mireku, M. O., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, *4*(1). https://doi.org/10.1186/s41235-019-0181-4

McMahon, L., Kleinman, Z., & Subramanian, C. (2025, January 8). Facebook and Instagram get rid of fact checkers. *BBC News*. https://www.bbc.com/news/articles/cly74mpy8klo

Miller, D. J., Somoray, K., & Stevens, H. (2025). *A shallow history of deepfakes*. SSRN. http://dx.doi.org/10.2139/ssrn.5130379

Ng, Y. L. (2023). An error management approach to perceived fakeness of deepfakes: The moderating role of perceived deepfake targeted politicians' personality characteristics. *Current Psychology*, *42*, 25658–25669. https://doi.org/10.1007/s12144-022-03621-x

Robertson, D. J., Mungall, A., Watson, D. G., Wade, K. A., Nightingale, S. J., & Butler, S. (2018). Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research: Principles and Implications*, *3*. https://doi.org/10.1186/s41235-018-0113-8

Senju, A., Vernetti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., & Johnson, M. H. (2013). Cultural background modulates how we look at other persons' gaze. *International Journal of Behavioral Development*, *37*(2), 131–136. https://doi.org/10.1177/0165025412465360

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, *4*. https://doi.org/10.1016/j.fsisyn.2022.100217

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*. https://doi.org/10.1016/j.chb.2023.107917

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, *2025*. https://doi.org/10.1155/hbe2/1833228

Smith, H., & Mansted, K. (2020). *Weaponised deep fakes: National security and democracy* [Policy brief]. Australian Strategic Policy Institute. https://www.aspi.org.au/report/weaponised-deep-fakes

Sütterlin, S., Ask, T. F., Mägerle, S., Glöckler, S., Wolf, L., Schray, J., Chandi, A., Bursac, T., Khodabakhsh, A., Know, B. J., Canham, M., & Lugo, R. G. (2023). Individual deep fake recognition skills are affected by viewer's political orientation, agreement with content and device used. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition: 17th International Conference, Held as Part of the 25th HCI International Conference, Copenhagen, Denmark, Proceedings: Vol. 14019 (pp. 269–284). Springer, Cham.* https://doi.org/10.1007/978-3-031-35017-7_18

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445699

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2020). Is it real? A study on detecting deepfake videos. *Proceedings of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.366

Wöhler, L., Zembaty, M., Castillo, S., & Magnor, M. (2021). Towards understanding perceptual differences between genuine and face-swapped videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445627

World Economic Forum (2024). *The Global Risks Report 2024*. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf