Face the Fake: Game-Based Deepfake Awareness on Interactive Public Displays

Doruntina Murtezaj
doruntina.murtezaj@unibw.de
University of the Bundeswehr Munich
Munich, Germany
LMU Munich
Munich, Germany

Anton Hantschmann
anton.hantschmann@campus.lmu.de
LMU Munich
Munich, Germany

Viktorija Paneva
viktorija.paneva@ifi.lmu.de
LMU Munich
Munich, Germany

Florian Alt
florian.alt@ifi.lmu.de
LMU Munich
Munich, Germany
University of the Bundeswehr Munich
Munich, Germany

Abstract

Deepfakes pose a growing societal challenge, creating a need for accessible interventions that help people critically reflect on synthetic media. We present *Face the Fake*, a game-based application that brings deepfake awareness to an interactive public display. The application invites passersby to distinguish between authentic and manipulated videos and provides immediate feedback and detection strategies. We examined its use through an exploratory two-phase field deployment in a university setting, combining interaction logs with post-interaction survey data. The findings indicate that participants engaged with the application, perceived the detection strategies as informative, and reported increased motivation to learn about deepfakes and greater confidence in recognizing them. The unattended deployment further illustrates the opportunities and practical challenges associated with voluntary interaction in public settings. This work contributes the design and implementation of *Face the Fake*, insights from its real-world deployment, and implications for designing future public AI-literacy interventions.

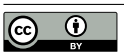
CCS Concepts

• **Security and privacy → Usability in security and privacy**; • **Human-centered computing → Empirical studies in HCI**.

Keywords

Public Displays, Usable Security, Deepfake Detection, Media Literacy, Game-Based Learning

ACM Reference Format:
Doruntina Murtezaj, Anton Hantschmann, Viktorija Paneva, and Florian Alt. 2026. Face the Fake: Game-Based Deepfake Awareness on Interactive Public Displays. In *25th International Conference on Mobile and Ubiquitous Multimedia (MUM '26), December 01–04, 2026, Glasgow, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3861188.3861230



This work is licensed under a Creative Commons Attribution 4.0 International License.
MUM '26, Glasgow, United Kingdom
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2791-7/2026/12
https://doi.org/10.1145/3861188.3861230

1 Introduction

Advances in generative artificial intelligence (AI) have made it increasingly easy to create realistic synthetic media. Deepfakes—videos, images, or audio generated or manipulated using AI techniques such as face swapping or reenactment—can convincingly depict events or statements that never occurred [34]. Although such technologies have legitimate creative applications, their increasing accessibility raises concerns about information integrity and public trust. Deepfakes have already been associated with misinformation, political manipulation, defamation, fraud, and extortion, illustrating their potential consequences for both individuals and society [2, 26, 44].

Efforts to address these risks have primarily focused on technical detection methods, platform-level measures, and regulation. However, technical countermeasures alone cannot fully prepare people for encountering synthetic media in everyday contexts, particularly as generation and detection techniques continue to evolve. Educational interventions therefore represent an important complementary approach: rather than attempting to identify manipulated content on users' behalf, they can support critical reflection on the credibility, context, and characteristics of digital media. Prior research has explored workshops, games, and other interactive formats for teaching AI and media-literacy concepts [3, 42]. Media-literacy interventions have likewise used feedback, active reasoning, and exposure to manipulation techniques to support more informed judgments about potentially misleading content [19].

Many existing interventions, however, require people to deliberately access an online resource, participate in a workshop, or enter a formal learning setting. This may limit their ability to reach individuals who would not actively seek out information about deepfakes. Interactive public displays offer a complementary setting for such interventions. Situated in shared environments, they can support brief, voluntary, and low-barrier encounters during people's everyday routines [11]. Their visibility can also make emerging security and media-literacy issues publicly accessible rather than confining

them to individual devices or formal educational contexts. At the same time, interactions with public displays are often opportunistic and short, creating distinct challenges for attracting attention, communicating unfamiliar concepts, and sustaining engagement.

Despite extensive research on public displays and growing interest in game-based misinformation interventions, limited work has examined how interactive public displays can be used for situated deepfake awareness. It therefore remains unclear how passersby engage with this type of intervention, how they perceive the educational content embedded within it, and what opportunities and practical challenges emerge when deepfake awareness is brought into a public setting.

To investigate this space, we developed *Face the Fake*, a game-based application deployed on an interactive public display. The application presents authentic and manipulated videos and invites users to decide which video is a deepfake. Following each decision, it provides immediate feedback and concise detection strategies intended to encourage reflection on visual and contextual cues. In this way, the intervention combines experiential exposure to synthetic media with conceptual guidance in a walk-up-and-use format.

We examined *Face the Fake* through an exploratory two-phase field deployment in a university building. In the first phase, 42 passersby were actively recruited to interact with the display and complete a post-interaction survey. In the second phase, the installation was deployed unattended, allowing us to examine voluntary use under less facilitated conditions. We combined interaction logs and survey responses to characterize engagement patterns, in-game behavior, participants' perceptions of the detection strategies, and their experiences of the public-display format. The following research questions guided the study:

RQ1 How do people engage with *Face the Fake* during a public-display deployment, and what patterns emerge in their in-game behavior?

RQ2 How do participants perceive the game format and the presented detection strategies as supporting deepfake awareness?

RQ3 What opportunities and practical challenges emerge when using interactive public displays for deepfake-awareness interventions?

Contribution Statement. This paper makes three contributions: (1) *System Design and Implementation*: We present *Face the Fake*, a functional public-display application that combines deepfake classification, immediate feedback, and detection strategies in a brief, game-based interaction. (2) *Empirical Insights*: Through a two-phase field deployment, we characterize interaction patterns, self-reported experiences, and differences between facilitated and unattended use of the installation. (3) *Design and Deployment Implications*: We derive evidence-linked implications for designing and studying future public-display interventions for AI and media literacy.

2 Background and Related Work

This section reviews research on deepfakes, media-literacy interventions, and interactive public displays, establishing the foundations and motivation for our study.

2.1 The Evolution and Societal Impact of Deepfakes

2.1.1 Deepfakes as Synthetic Media. Deepfakes have emerged as a prominent socio-technical phenomenon, combining advances in generative artificial intelligence with substantial societal implications. The term refers to digitally generated or manipulated media that use deep-learning techniques to imitate authentic video, images, or audio [16]. Deep learning enables computational models to identify and reproduce patterns in large datasets, supporting techniques such as face swapping, attribute manipulation, reenactment, and the synthesis of identities or expressions [22, 32]. As a result, synthetic media can appear sufficiently realistic to make distinguishing manipulated from authentic content difficult for human viewers [18].

The term *deepfake* gained widespread attention in 2017, when open-source face-manipulation tools were used to create synthetic videos from publicly available material [23]. Since then, the production and dissemination of deepfakes have accelerated due to advances in generative models and the increasing accessibility of the required tools and computational resources [15, 38, 44]. What began as a relatively specialized form of media manipulation has consequently developed into a broader challenge for interpreting digital content and assessing its authenticity.

2.1.2 Societal Implications. Deepfake technology has a dual character. It enables constructive applications in entertainment, education, accessibility, and personalized media. Examples include recreating or translating voices, producing visual effects, supporting virtual try-on applications, and generating personalized audiovisual content [9, 44]. These applications illustrate that synthetic-media technologies are not inherently harmful and that their implications depend substantially on their purpose and context of use.

At the same time, deepfakes can facilitate misinformation, political manipulation, non-consensual sexual content, financial fraud, extortion, and defamation [2, 16, 36]. Documented incidents include manipulated media targeting political candidates, fabricated statements attributed to public figures, and impersonation attempts in organizational communication [21]. Beyond the immediate harm caused by individual deceptive artifacts, the growing awareness that audiovisual evidence can be fabricated may also weaken trust in authentic media [40]. Deepfakes therefore pose not only a detection problem but also a broader challenge for how people assess and trust digital information.

Responses to these risks have frequently focused on automated detection methods that identify technical traces of manipulation [38]. However, detection techniques must continuously adapt to changing generation methods and cannot fully address the situations in which people encounter synthetic media without access to specialized tools. Technical detection therefore represents only one part of a broader response. Educational interventions can complement it by helping people critically examine the source, context, and characteristics of potentially manipulated content.

2.1.3 Media Literacy and Game-Based Interventions. Media literacy emphasizes people's ability to assess, contextualize, and question digital content rather than relying exclusively on automated detection. In the context of misinformation and synthetic media,

literacy-oriented interventions seek to foster reflective judgment, appropriate skepticism, and awareness of common manipulation tactics rather than perfect classification accuracy [24].

Prior work distinguishes between reactive interventions, such as fact-checking after exposure, and preventive approaches that prepare people before they encounter misleading content. Inoculation theory has informed preventive misinformation interventions by proposing that controlled exposure to manipulation strategies can strengthen resistance to subsequent deceptive messages [24]. Instead of teaching judgments about individual examples alone, such interventions draw attention to recurring manipulation techniques, inconsistencies, implausible details, and contextual cues.

Interactive interventions can operationalize these principles through active participation, immediate feedback, and opportunities for reflection. Games are particularly suited to creating environments in which users can experiment, make mistakes, and reconsider their judgments. A systematic review by Kiili et al. [24] indicates that game-based misinformation interventions can encourage active reasoning, pattern recognition, and reflection. For example, *Fakey* uses a simulated social-media environment to expose players to credibility cues and trade-offs involved in judging online information [29]. In the context of deepfakes, Maathuis et al. [27] emphasize the value of presenting diverse manipulation techniques, balancing difficulty, and encouraging reflection beyond binary right–wrong feedback.

These studies establish games as a promising format for misinformation and deepfake awareness. However, existing interventions have predominantly been designed for web-based, mobile, classroom, or workshop settings. Comparatively little is known about bringing such interventions into shared public environments, where encounters are brief, voluntary, and influenced by the surrounding social and spatial context. *Face the Fake* explores this gap by combining exposure to authentic and manipulated videos, immediate feedback, and detection strategies in a walk-up-and-use public-display application.

2.2 Public Displays as Interactive Infrastructure

Public displays have become a ubiquitous feature of modern urban environments – ranging from public squares and university campuses to hospitals, shopping centers, and transportation hubs [11]. They bring digital content into shared physical spaces and into public view, bridging the digital and social spheres [35]. Earlier generations of displays were largely passive, presenting static advertisements or looping videos [8]. Over time, advances in sensing and interaction design have transformed them into interactive media capable of engaging passersby in dynamic ways [20]. Public displays are particularly suited for informal and situated learning scenarios, where users engage opportunistically rather than through deliberate instruction, making them a promising medium for lightweight media literacy interventions [33].

2.2.1 Spatial and Temporal Models of Interaction. Research on *Interactive Public Displays (IPDs)* has established both spatial and temporal models of interaction. Spatial models such as Prante's *three zones of interaction* [37] or proxemics-based approaches by Greenberg et al. [17] conceptualize user behavior in terms of distance and bodily orientation relative to the display. Temporal models, in contrast, describe how users transition through different engagement stages. Brignull and Rogers' *public interaction flow model* [8] and Müller et al.'s *audience funnel* [30] describe how passersby move from peripheral awareness to direct interaction, emphasizing the need to overcome participation thresholds. Together, these models highlight that interaction with public displays is a gradual process rather than a binary on–off activity, shaped by curiosity, attention, and social context.

Despite these insights, barriers to interaction remain a key challenge. Users may hesitate to approach displays due to uncertainty about interactivity, social inhibitions, or fear of disrupting others. Cheung et al. [10] addressed these barriers by integrating personal mobile devices as socially safe entry points to public screens, reducing the effort needed to initiate or withdraw from interaction. This hybrid approach illustrates how linking personal and public devices can improve accessibility and user comfort in shared environments.

2.2.2 Social Dynamics and User Behavior Around Displays. Interaction with public displays is inherently social. The *honeypot effect* [8, 31, 45] describes how observing others interact can attract additional users, creating chained participation. Similarly, display configuration and spatial layout influence engagement: concave displays attract fewer users, while flat configurations promote stronger group involvement [39]. Concepts such as the *sweet spot* [5] and *f-formations* [28] further explain how spatial arrangements shape user positioning and collaboration, particularly in multi-user scenarios. Equally important are the reasons why users disengage. Studies have identified factors such as interruption, discomfort, or social pressure as common triggers for leaving a display [41, 45]. Alt et al. [4] demonstrated how users' intention to leave a display can be detected and used as a trigger for measures to make users re-engage. These findings highlight that sustaining engagement in public spaces requires sensitivity to both environmental and social context. For open-ended or playful applications, designing for graceful disengagement–or turning leaving moments into reflective opportunities–remains an open challenge.

2.2.3 Public Displays for Security and Media Literacy. Murtezaj et al. [33] introduced the concept of *Public Security User Interfaces (PSUIs)*–interactive installations designed to raise cybersecurity awareness and promote secure digital behavior. PSUIs aim to (1) create awareness of security risks (e.g., password reuse, phishing detection), (2) trigger protective actions, (3) provide direct control, and (4) spark public conversation about digital security. PSUIs are typically multimodal, combining touch, gesture, and visual or auditory feedback. Our work extends this line of research from cybersecurity to media literacy by exploring how an interactive public display can support situated engagement with AI-generated content. *Face the Fake* combines playful classification, feedback, and detection strategies; our evaluation examines engagement and perceptions rather than its effectiveness in building resilience.

3 Face the Fake: System Design and Implementation

Face the Fake is a game-based public-display application in which users assess authentic and manipulated videos, receive immediate

feedback, and explore strategies for evaluating synthetic media. The system was designed for brief, voluntary interaction in public settings and to support an exploratory field deployment. This section describes its design rationale, interaction flow, content, and technical implementation.

3.1 Design Goals

The design goals were derived through a literature-informed synthesis. Research on walk-up-and-use public displays motivated low-barrier onboarding and support for brief, self-directed encounters [11, 12]; game-based misinformation interventions motivated active judgment and reflection [24, 29]; and work on conceptual scaffolding and feedback in security education motivated immediate feedback and explanatory guidance [7, 13, 46]. We adapted these principles to the intended public-display deployment, in which participation was voluntary, users received no prior instruction, and interaction duration was self-selected. This process resulted in four design goals:

DG1: Enable walk-up-and-use interaction.
Public-display users may encounter the application without prior instructions. An animated start screen signals interactivity, while short tutorial overlays introduce navigation and video controls during first use.

DG2: Support active judgment and immediate feedback.
Rather than presenting information passively, the application asks users to classify videos and immediately reveals whether their decisions were correct.

DG3: Encourage reflection beyond binary classification.
Authentic and manipulated videos are presented side by side after each decision, followed by concise detection strategies. This structure connects the user's judgment to visual comparison and conceptual guidance.

DG4: Accommodate brief and flexible engagement.
Each video constitutes a self-contained round, allowing users to stop after one judgment or continue playing. Progress information, optional session resumption, and access to the study survey support different levels of engagement.

These goals informed a recurring challenge–feedback–reflection loop. We use the term *game-based* to describe this combination of repeated classification challenges, immediate performance feedback, progress information, and voluntary continuation. The application was not designed as a comprehensive training program or as a test of expert detection performance.

3.2 Interaction Flow

The application begins in an attract mode that uses animation and a call to action to communicate that the public display is interactive. After starting, users receive a short explanation of the task and are shown a video clip. For each round, the application randomly selected either the authentic or manipulated version of the current video pair with equal probability for the classification task. The corresponding alternative was withheld until the subsequent comparison view. The core interaction follows a challenge–feedback–reflection loop (Figure 1). Users classify a video, compare the authentic and manipulated versions, explore detection strategies, and review their in-game activity and performance. A comparison view subsequently presents the authentic and manipulated versions side by side, allowing users to inspect differences between them.

Following the comparison, users can access cards presenting strategies for evaluating deepfake content. These strategies address perceptual cues, such as facial inconsistencies, irregular lighting, unnatural blinking, as well as contextual verification, such as checking whether a depicted event or statement can be corroborated through other sources. The strategies are framed as fallible heuristics rather than deterministic indicators because individual artifacts may disappear as generation techniques improve [1, 6, 14].

Users can then continue with another video or access an analytics view summarizing the number of completed rounds and their in-game accuracy. Navigation follows a sequential horizontal structure using swipe gestures and arrow buttons. Animated overlays demonstrate the relevant gestures when users first encounter a view, reducing the need for external instructions.

Two auxiliary functions support longer engagement and study participation. Users can save their progress and receive a four-digit PIN with which to resume a session. They can also open a dynamically generated QR code leading to the post-interaction survey. The QR-code URL contains an anonymous session identifier that links survey responses with the corresponding interaction logs.

3.3 Video and Strategy Content

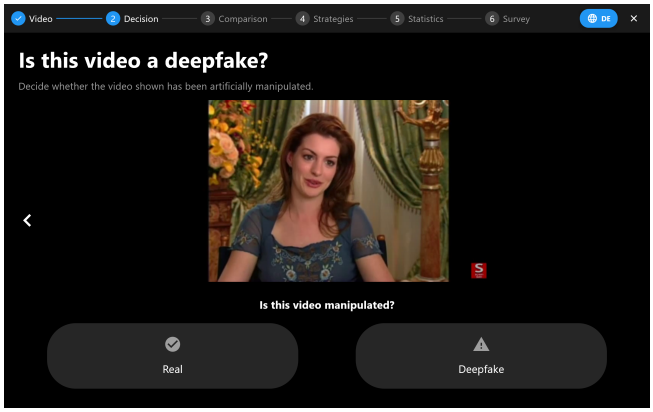
The application used eleven authentic–manipulated video pairs selected from the *Celeb-DF (v2)* dataset [25]. This dataset contains authentic videos of public figures and corresponding high-quality face manipulations and is widely used in deepfake-detection research. We selected paired clips to enable direct comparison between authentic and manipulated representations of the same identity. The selected material therefore focuses specifically on facial video manipulation and does not represent the full range of synthetic media, such as generated audio, images, or text.

The subset was selected to provide visually comparable real–fake pairs and variation across depicted individuals and observable artifacts. It was intended to support exposure and reflection rather than benchmark human classification performance. The clips were not formally calibrated by difficulty, and the strategies were not treated as guarantees that a particular video could be identified correctly. Instead, the application combined perceptual cues with broader contextual-verification strategies to avoid presenting deepfake recognition as a deterministic visual task.

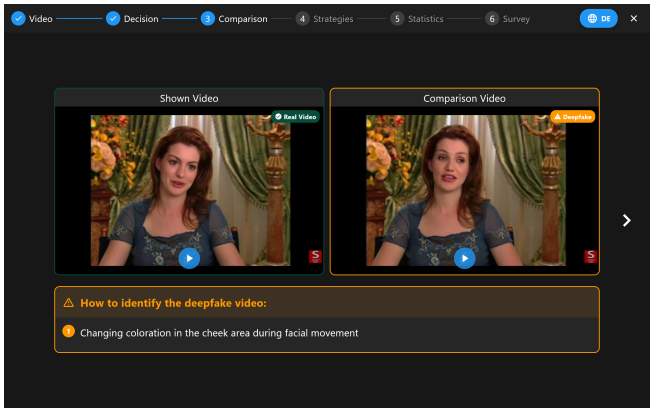
Video assets and metadata are defined in an external JSON configuration. Each entry specifies the video and thumbnail files, its real or manipulated status, and an identifier linking the two members of a pair. This configuration allows the video set to be expanded or replaced without modifying the application logic.

3.4 Technical Implementation

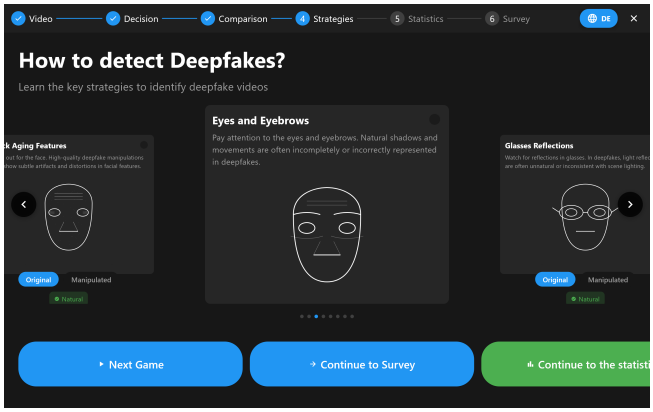
Face the Fake was implemented in Flutter and deployed as a native Windows application on a 55-inch touchscreen connected to a Windows 11 workstation. Its architecture separates the user interface, application state, and data access using the BloC pattern. Video content is loaded from the JSON configuration, while anonymous session data—including completed rounds, correct decisions, timestamps, the initial confidence rating, and use of the save and resume



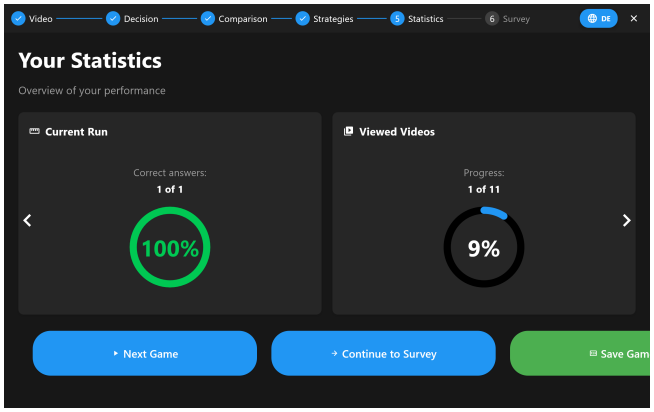
(a) Video classification



(b) Feedback and comparison



(c) Detection strategies



(d) Performance summary

Figure 1: Core interaction flow of *Face the Fake*. Users classify a video as authentic or manipulated (a), compare the authentic and manipulated versions and receive feedback on observable cues (b), explore strategies for evaluating deepfake content (c), and review a summary of their in-game activity and performance (d).

functions—were stored locally. The application supports German and English and restricts unintended access to operating-system functions to maintain continuous operation in the public setting. This modular structure enables researchers to replace the video material and detection strategies or deploy the application on other compatible devices. The system can be made available upon request; the video assets remain subject to the conditions of the original dataset.

4 Field Evaluation

We conducted an exploratory field evaluation to examine how people engaged with *Face the Fake*, how they perceived its game-based format and detection strategies, and what opportunities and challenges emerged during its deployment on a public display. The evaluation focused on short-term interaction patterns and self-reported experiences rather than long-term learning or transferable detection ability.

4.1 Study Design and Procedure

The study comprised two consecutive phases conducted in the same location. Phase 1 was a four-day facilitated deployment during which members of the research team approached passersby and invited them to use the application and complete a post-interaction survey. This phase enabled us to link interaction logs with participants' survey responses and obtain feedback about their experience.

Phase 2 was a nine-day unattended deployment during which the display remained freely accessible without active recruitment or researcher supervision. This phase provided a complementary view of voluntary interaction under less facilitated conditions. Both phases ran on weekdays between 10:00 and 16:00. Because Phase 2 interaction was anonymous, we could not determine whether individual users participated more than once or had previously encountered the installation during Phase 1.

The installation was situated in a university corridor with regular pedestrian traffic and nearby seating and was accessible to students, staff, and visitors. The display was positioned for use while standing and was visible to passersby approaching from the corridor.

4.2 Data Collection

The application automatically recorded an anonymous log for each session, including the number of completed rounds, the number of correct real–deepfake decisions, timestamps for the first and final rounds, and use of the save and resume functions. Aggregate in-game accuracy and interaction duration were derived from these values. Before the first round, users rated their perceived ability to recognize deepfake videos on a five-point scale. This initial rating was stored in the interaction log.

For Phase 2, the unit of analysis was an interaction session. A new session began when the application was initiated from the attract screen and received a new anonymous session identifier. An active session ended when the user explicitly exited the interaction or after 30 seconds of inactivity, at which point the application returned to the attract screen. Accordingly, the 30 analyzed sessions represent interactions rather than confirmed unique users, since the anonymous logging did not differentiate repeated or collaborative use.

Afterwards, users could access a Qualtrics survey by scanning a QR code. A six-character session identifier embedded in the survey URL allowed responses to be linked to the corresponding interaction log without collecting identifying information. The survey contained 20 questions addressing prior experience with deepfakes and public displays, perceived information gain, perceptions of the detection strategies, learning motivation, reasons for beginning and ending the interaction, demographics, and optional feedback. The post-interaction survey also repeated the five-point self-assessment of perceived deepfake-recognition ability, enabling within-participant comparison with the rating collected before play. Attitudinal items used seven-point agreement scales, while the survey also included multiple-choice and open-response questions. The complete questionnaire is provided in Appendix A.

4.3 Participants and Recorded Sessions

In Phase 1, 42 participants interacted with the application and completed the survey. In Phase 2, 30 additional interaction sessions were retained for behavioral analysis, but only three corresponding surveys were submitted. Because these responses were insufficient to characterize perceptions during Phase 2, they were excluded from the survey analysis. Phase 2 is therefore represented through anonymous interaction logs only. Participation was voluntary, and participants received no compensation.

4.4 Data Analysis

The analysis was organized around the three research questions. We used descriptive statistics to characterize participant demographics, interaction duration, completed rounds, in-game accuracy, strategy access, and survey responses. Phase 1 and Phase 2 logs were analyzed separately because the phases differed in recruitment and because demographic and survey data were unavailable for most Phase 2 sessions.

Changes in self-assessed recognition ability were examined within participants by comparing the ratings collected before and after interaction. To explore whether these changes were related to in-game behavior, we calculated Spearman rank correlations between participants' pre–post self-assessment difference and their overall in-game accuracy and number of completed rounds. These analyses were exploratory and were not interpreted as evidence of learning or causal effects. Participants who completed more rounds were self-selected, and their performance may have differed before or independently of continued play. Furthermore, because the application recorded aggregate session outcomes rather than the sequence of correct and incorrect decisions, we could not examine changes in accuracy across successive rounds.

Comparisons involving self-reported strategy access were likewise interpreted descriptively because only a small number of Phase 1 participants did not access the strategies and prior deepfake experience differed between the groups. Multiple-choice responses concerning attention and reasons for ending play were summarized as frequencies and percentages. Because participants could select multiple responses, these categories were not treated as mutually exclusive. Responses to open-ended questions were used descriptively to contextualize quantitative patterns; no formal thematic analysis was conducted. For the visualization of the seven-point agreement ratings, we report approximate 95% confidence intervals calculated as $M \pm 1.96(SD/\sqrt{n})$.

4.5 Ethics

The study was approved by the Ethics Committee of the Faculty of Mathematics, Informatics and Statistics at LMU Munich (approval no. EK-MIS-2025-0418-FT-d01). Participants were informed that participation was voluntary and anonymous and that they could withdraw at any time. The application did not collect directly identifying information. Interaction logs and survey responses were linked only through anonymous session identifiers and stored in accordance with the EU General Data Protection Regulation. Interaction with the public installation was based on informed implied consent, while survey participation required explicit confirmation.

5 Results

We report findings from the interaction logs and post-interaction survey, organized around the three research questions. Unless otherwise stated, survey findings refer to the 42 participants in Phase 1. Phase 2 findings are based on anonymous interaction logs from 30 sessions.

5.1 Data Quality and Participant Characteristics

In Phase 1, 48 interaction sessions were recorded. Forty-two sessions could be linked to survey responses and constituted the Phase 1 analysis sample ($N = 42$); the remaining six unlinked sessions were excluded. One participant omitted only the final optional-feedback page, and their otherwise complete response was retained. Responses entered under "Other" for field of study were recategorized based on the reported disciplines. Two application crashes affected the comparison view during Phase 1; interaction records generated before and after these crashes were merged using their session identifiers.

After incomplete logs resulting from application crashes were excluded, 30 Phase 2 sessions remained for the behavioral analysis. Three corresponding survey responses were submitted, but these were excluded from the survey analysis because they were

insufficient to characterize perceptions during the unattended deployment.

Phase 1 participants had a mean age of $M = 26.1$ years ($SD = 10.2$, range 18–59). The sample was predominantly female (73.8%, $n = 31$), with 26.2% identifying as male ($n = 11$). Most participants held a high-school diploma or lower qualification (69.0%, $n = 29$). The most frequently represented fields were Humanities & Languages (40.5%, $n = 17$), Social & Educational Sciences (21.4%, $n = 9$), and Law & Business (16.7%, $n = 7$). Most participants had previously encountered deepfakes (78.6%, $n = 33$), whereas 42.9% ($n = 18$) reported that they had not previously interacted with a public display.

5.2 RQ1: Engagement and In-Game Behavior

Phase 1 participants completed an average of $M = 6.24$ rounds ($Mdn = 5$, $SD = 3.92$, range 1–13) and interacted with the application for $M = 5.45$ minutes ($Mdn = 5.00$, $SD = 3.69$, range 0–20.1). Their mean in-game classification accuracy was 75.2%. Recorded durations of zero could occur in single-round sessions because duration was calculated from the timestamps of the first and final rounds.

Interactions during the unattended Phase 2 deployment were shorter and involved fewer rounds. Phase 2 sessions included an average of $M = 4.40$ rounds ($Mdn = 3$, $SD = 3.84$, range 1–13) and lasted $M = 3.12$ minutes ($Mdn = 2.20$, $SD = 3.56$, range 0–15.9). Mean in-game classification accuracy was 69.2% ($Mdn = 76.4\%$, $SD = 34.3$ percentage points, range 0–100%). These differences are reported descriptively because the phases differed in recruitment and available participant information. The save function was used in one session during each phase, but no subsequent session resumed previously saved progress.

5.3 RQ2: Perceptions of the Game Format and Detection Strategies

Before beginning the first round, Phase 1 participants rated their perceived ability to recognize deepfake videos at $M = 2.98$ ($Mdn = 3$, $SD = 0.81$, range 1–4) on a five-point scale. The post-interaction rating was $M = 3.79$ ($Mdn = 4$, $SD = 0.87$, range 2–5), corresponding to a mean within-participant difference of $\Delta M = 0.81$. Ratings increased for 61.9% ($n = 26$) of participants, remained unchanged for 26.2% ($n = 11$), and decreased for 11.9% ($n = 5$), as illustrated in Figure 2 (b).

Exploratory analysis showed that larger increases in self-assessed recognition ability were moderately associated with higher overall in-game accuracy ($r_s = .35$, $p = .024$). Descriptively, participants whose self-assessment increased tended to show higher in-game accuracy than those whose ratings remained unchanged or decreased, although the groups were unequal in size, particularly the decreased-rating group ($n = 5$). Changes in self-assessment were not clearly associated with the number of completed rounds ($r_s = .25$, $p = .106$). Because the logs contained aggregate session outcomes, we could not determine whether accuracy changed across successive rounds. These findings concern perceived recognition ability and performance on the presented in-game material rather than an objective or transferable measure of detection skill.

Most participants (85.7%, $n = 36$) reported accessing the detection strategies. Among these participants, agreement that the strategies were new to them averaged $M = 5.72$ ($SD = 1.14$) on a seven-point scale. Across the Phase 1 sample, agreement that the application provided new information about recognizing deepfakes averaged $M = 6.17$ ($SD = 1.12$).

Participants also agreed that the playful format increased their motivation to learn about recognizing deepfakes ($M = 5.93$, $SD = 1.07$). Figure 2 (c) summarizes these self-reported outcomes. When asked which elements they found motivating, participants most frequently selected the detection strategies (61.9%, $n = 26$) and videos of public figures (57.1%, $n = 24$), followed by the performance overview (31.0%, $n = 13$), the save function (11.9%, $n = 5$), and other features (4.8%, $n = 2$). Multiple selections were permitted.

5.4 RQ3: Opportunities and Challenges of the Public-Display Deployment

During the facilitated Phase 1 deployment, spare time was the most frequently reported reason for engaging with the installation (57.1%, $n = 24$). The animation and playful format were each selected by 23.8% ($n = 10$), followed by an existing interest in deepfakes (21.4%, $n = 9$) and being approached by another person (19.0%, $n = 8$).

Limited time was the most frequently reported reason for ending an interaction (40.5%, $n = 17$). Thirteen participants (31.0%) could not identify a particular reason, nine (21.4%) had viewed all available videos, and three (7.1%) reported losing interest. Table 1 summarizes these responses. Because participants could select multiple responses, percentages are not mutually exclusive and do not sum to 100%.

The unattended phase recorded 30 voluntary sessions, showing that the installation was used without active recruitment. However, only three interactions proceeded to the optional survey. Consequently, Phase 2 provides evidence about interaction behavior but not about users' motivations, demographics, or perceptions of the public-display medium.

6 Discussion

Our exploratory deployment examined how people engaged with a game-based deepfake-awareness application on a public display, how they perceived its educational components, and what opportunities and challenges arose in facilitated and unattended use. The findings characterize brief, situated encounters with deepfake-awareness content; they do not establish objective or transferable improvements in detection ability.

6.1 Interpretation of the Findings

RQ1: Engagement and In-Game Behavior. Both deployment phases recorded multi-round interactions with the deepfake-awareness application, including voluntary use during the unattended phase. During the facilitated phase, interactions lasted approximately five minutes and included several classification rounds. The unattended phase also recorded voluntary use, although sessions were shorter and involved fewer rounds. These findings are consistent with public-display research showing that interaction is often brief, opportunistic, and shaped by the surrounding situation [11].

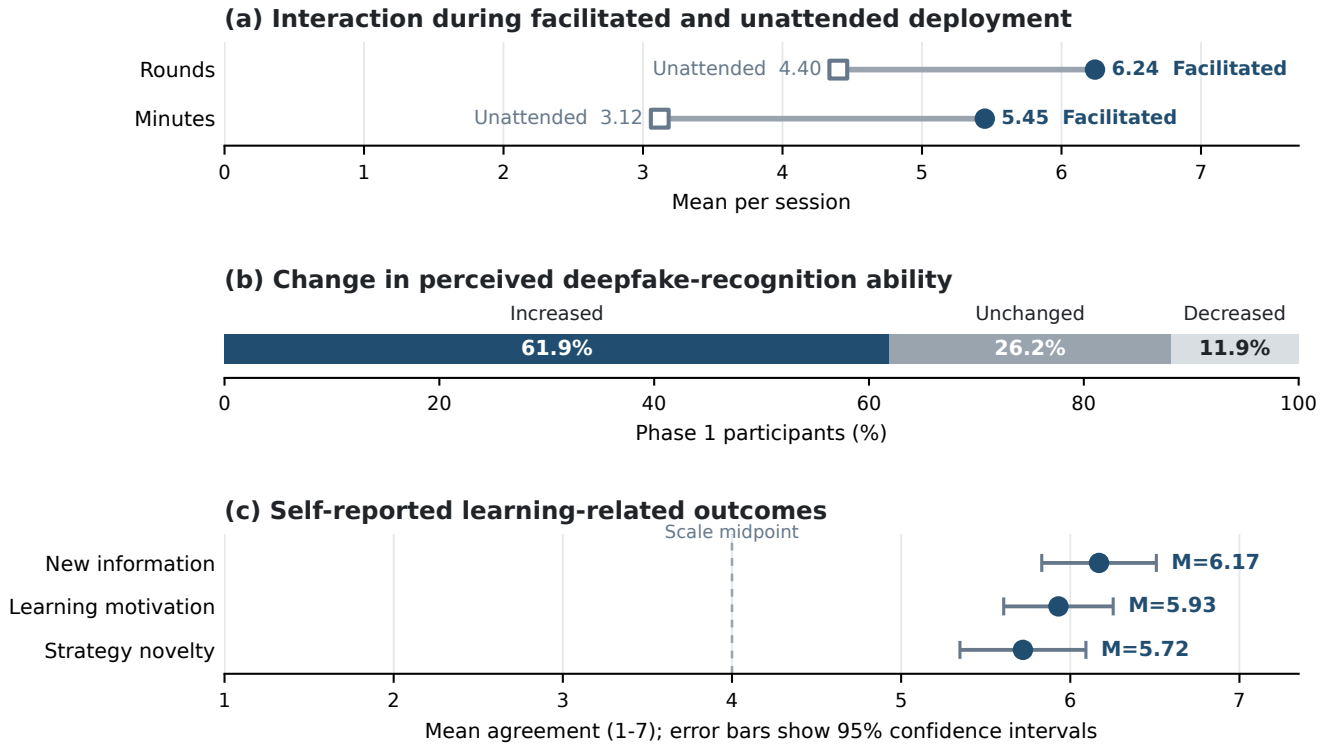


Figure 2: Overview of behavioral and self-reported outcomes. Panel (a) compares the mean number of rounds and interaction duration between the facilitated Phase 1 and unattended Phase 2 deployments. Panel (b) shows the direction of change in participants' self-assessed deepfake-recognition ability from before to after the Phase 1 interaction. Panel (c) presents mean agreement ratings for perceived information gain and learning motivation among Phase 1 participants ($N = 42$), and strategy novelty among participants who accessed the strategies ($n = 36$). Error bars in Panel (c) represent 95% confidence intervals; the dashed line marks the midpoint of the seven-point response scale.

Table 1: Reported reasons for engaging with and ending the Phase 1 interaction. Multiple responses were permitted ($N = 42$).

Reason for engagement	*n*	%	Reason for ending interaction	*n*	%
Spare time	24	57.1	Limited time	17	40.5
Striking animation	10	23.8	No particular reason	13	31.0
Playful format	10	23.8	Viewed all videos	9	21.4
Interest in deepfakes	9	21.4	Lost interest	3	7.1
Approached by another person	8	19.0	Other	3	7.1
Waiting for someone	2	4.8			
Other	2	4.8			

The differences between phases should not be interpreted as an experimental effect of facilitation. Phase 1 participants were actively recruited and invited to complete a survey, whereas Phase 2 sessions were initiated without researcher involvement. However, anonymous logging prevented us from determining whether Phase 2 sessions involved first-time users, repeat visitors, or Phase 1 participants returning with other people. Shorter sessions and fewer rounds may therefore reflect several factors, including facilitation, differences in the people who encountered the display, prior exposure to the application, or group use. Phase 2 consequently demonstrates interaction without active recruitment, but not necessarily independent first-time use.

The save function was used in only one session during each phase, and no subsequent session resumed saved progress. This establishes that the feature was rarely used in this deployment, but it does not explain why. Users may not have noticed or understood the feature, may not have perceived sufficient value in returning, or may not have expected to encounter the installation again. The observed pattern therefore motivates the hypothesis that self-contained encounters may be particularly important in transient public-display settings; it does not demonstrate that continuity features are generally unnecessary. Future deployments should examine the discoverability and perceived value of such features in longer studies and in locations with recurring audiences.

RQ2: Perceptions of the Game Format and Detection Strategies. Participants generally reported that the application provided new information and increased their motivation to learn about deepfakes. Most also accessed the detection strategies and perceived them as novel. These findings support the relevance of combining an experiential classification task with explanatory content. In contrast to feedback that communicates only whether a decision was correct, the strategy cards offered a vocabulary for reflecting on perceptual and contextual cues. This corresponds with work on conceptual scaffolding, in which guidance directs attention and supports interpretation during learning activities [7], and with game-based misinformation interventions that combine active judgment with feedback and reflection [24, 29].

However, the study cannot determine whether the strategies improved participants' subsequent decisions. Strategy access was voluntary, almost all participants viewed the strategies, and the small group that did not view them differed in prior deepfake experience. The reported information gain and novelty ratings should therefore be understood as participants' perceptions of the material rather than evidence of knowledge transfer.

Participants' self-assessed recognition ability was higher after interacting with the application, and larger increases were moderately associated with higher overall in-game accuracy. This pattern suggests that participants' confidence was related to their experience within the game. One possible explanation is that the application's immediate performance feedback helped participants adjust their self-assessment in relation to their decisions. However, the direction of this relationship cannot be determined: stronger in-game performance may have increased confidence, changes in confidence may have been related to how participants approached the task, or both may reflect other individual differences.

The association does not demonstrate improved detection ability. Participants provided their post-interaction self-assessment after receiving feedback on their decisions, in-game accuracy was measured only on the presented videos, and the study did not include an objective pre–post assessment using unseen material. Moreover, the aggregate logs did not allow us to determine whether individual accuracy improved across successive rounds. Although changes in perceived ability were moderately associated with in-game accuracy, confidence is not equivalent to competence, and the two can diverge in perceptual judgments [43]. The findings should therefore be interpreted as an association between perceived ability and in-game performance, not as evidence of learning or transfer.

This distinction is particularly important because many visual artifacts associated with earlier deepfakes may become less visible as generation techniques improve. Detection strategies should consequently be presented as fallible prompts for investigation rather than as deterministic rules. Combining perceptual cues with contextual practices, such as source verification and corroboration, may provide a more durable basis for media literacy than teaching visual artifacts alone.

RQ3: Opportunities and Challenges of the Public-Display Medium. During the facilitated phase, spare time was the most frequently reported reason for engaging with the installation, while animation, playful presentation, and pre-existing interest were selected less frequently. This finding underscores the situated character of public-display interaction: engagement depends not only on interface design but also on whether passersby have an appropriate moment to stop. Similarly, limited time was the most common reason for ending an interaction.

The unattended deployment yielded 30 analyzed sessions, showing that the installation was used without active recruitment. However, the study did not record the total number of passersby or people who noticed but did not approach the display. It therefore cannot establish how effectively the installation attracted the broader corridor population. Only three corresponding survey responses were submitted. This low response count should not necessarily be interpreted as a lack of interest: some sessions may have involved groups in which no individual assumed responsibility for completing the optional survey, and some users may previously have participated during Phase 1 or returned to show the application to others. Because individual identity and group composition were not recorded, these explanations cannot be distinguished. Phase 2 therefore provides evidence about interaction behavior but little insight into who used the application, why they engaged, or how they perceived it.

Informal observations during deployment suggested that some people interacted in pairs or discussed the content with bystanders. Because group size and social interaction were not recorded systematically, these observations cannot support conclusions about collaborative learning. They nevertheless identify a relevant question for future public-display research: whether visibility and co-located use can extend deepfake reflection beyond the person directly operating the interface.

6.2 Design and Deployment Implications

Design for Complete but Extendable Micro-Interactions. The brief interactions observed in both phases suggest that public deepfake-awareness interventions should provide meaningful content within a single encounter. Each round can offer an understandable challenge, immediate feedback, and a concise reflective explanation without requiring users to complete a longer sequence.

At the same time, the limited use of the save function does not establish that support for continued interaction is unnecessary. Progressive disclosure can provide additional depth for users who remain longer, while save-and-resume functionality may be more valuable in locations with recurring audiences or when its purpose and benefits are made more visible. The relative value of these alternatives remains a question for future investigation.

Implication 1: Self-Contained Micro-Interactions

Design public AI-literacy interventions as short, self-contained challenge–feedback–reflection cycles that offer a meaningful opportunity for reflection within a single encounter. Allow users to extend the engagement through optional explanations and additional challenges without making these necessary for completing the core experience.

Frame Detection Strategies as Evolving Heuristics. Participants frequently accessed the detection strategies, perceived them as novel, and often identified them as a motivating element of the application. These findings support including concise explanatory

content alongside classification feedback. However, our evaluation does not establish which strategies improve subsequent judgments or whether perceptual and contextual strategies differ in their effectiveness.

The recommendation to present detection cues as fallible and evolving is therefore grounded in prior deepfake research rather than directly in our evaluation. Prior work shows that observable artifacts vary across generation methods and may become less reliable as these methods improve [1, 6, 14]. Drawing on this literature, designers should avoid presenting individual visual cues as definitive indicators and should complement perceptual guidance with source verification and contextual reasoning. Such content should also be reviewed and updated as synthetic-media techniques evolve.

Implication 2: Reflective Detection Guidance

Pair classification feedback with concise explanations that encourage reflection beyond correct and incorrect choices. In deepfake-awareness applications, complement perceptual cues with source verification and contextual reasoning, while regularly updating examples.

Match Facilitation and Data Collection to Deployment Goals. The contrast between phases highlights a trade-off between naturalistic deployment and data completeness. Active recruitment produced longer sessions and substantially higher survey participation, whereas the unattended installation captured autonomous use but yielded little self-report data.

Researchers should determine whether their primary objective is broad public exposure, detailed evaluation, or a combination of both. For unattended deployments, feedback collection should be integrated into the display and kept extremely brief rather than relying exclusively on an optional external survey. Where the topic is complex or sensitive, intermittent facilitation or clear contextual signage may help users interpret the intervention without requiring continuous supervision.

Implication 3: Balancing Naturalistic Deployment and Evaluation

Align the level of facilitation and feedback mechanisms with the goals of the public-display deployment. Facilitated deployments can support participant-level assessments, while unattended deployments provide opportunities to examine voluntary interaction but may yield limited self-report data. Incorporate lightweight, in-situ feedback mechanisms when evaluating unattended installations.

6.3 Limitations

Several limitations constrain the conclusions that can be drawn. First, the study used an exploratory field design without a control or comparison condition. Changes in self-assessed ability cannot be attributed specifically to the game format, detection strategies, feedback, or exposure to deepfake examples. The study also did not include an objective pre–post assessment using unseen videos or a delayed measurement; it therefore cannot establish skill acquisition, transfer, or retention.

Second, the Phase 1 sample was small, predominantly female, and recruited in a university environment. Participants' educational backgrounds and prior experience may differ from those of the broader population envisioned for public AI-literacy interventions. Phase 2 provided no usable demographic or perception data, and anonymous logging prevented us from identifying repeated participation or overlap between phases. Because overall corridor traffic and non-interactions were not recorded, attraction and conversion rates cannot be calculated.

Third, the evaluation used eleven video pairs from Celeb-DF (v2). The clips were not formally calibrated by difficulty, and familiarity with the depicted public figures may have affected engagement or classification. The dataset focuses on facial video manipulation and does not represent other synthetic-media forms, such as generated audio, images, or fully generated video. The detection strategies were derived from prior literature but were not validated systematically against each selected clip. A few crashes in the comparison view required affected records to be merged or excluded and may have influenced individual experiences.

Finally, social interaction was observed only informally. Without structured observation, recordings, or group-level logs, the study cannot determine how collaboration or bystander presence affected engagement, decision-making, or survey participation. In particular, group interaction may have reduced survey completion if no individual felt responsible for responding.

6.4 Future Work

Future evaluations should compare the complete application with meaningful alternatives, such as exposure to the same videos without game elements or feedback without detection strategies. Objective assessments using unseen material are needed to examine whether participants improve beyond the practiced examples and whether any effects persist over time. Larger deployments across more diverse locations and populations could assess how context influences uptake. Future systems should also include broader and regularly updated synthetic-media examples, video-level difficulty assessment, and privacy-preserving measures of group interaction and passerby conversion. Together, these extensions would allow stronger conclusions about when and how public displays can contribute to AI and media literacy.

7 Conclusion

This paper presented *Face the Fake*, a game-based public display application designed to support brief, situated engagement with deepfake awareness content. Through an exploratory two-phase field deployment, we characterized how people interacted with the application under facilitated and unattended conditions and how participants perceived its game format and detection strategies. Participants engaged with multiple classification rounds, frequently accessed the explanatory strategies, and reported increased motivation, information gain, and perceived recognition ability. These self-reported outcomes indicate interest in the intervention but do not establish objective improvements in deepfake-detection skills.

The deployment further illustrates both the potential and challenges of bringing AI-literacy interventions into everyday public settings. Public displays can provide opportunities for low-barrier encounters with synthetic-media content, but engagement depends strongly on situational availability, and unattended deployments

make sustained participation and feedback collection difficult. Future work should compare the application with alternative intervention formats, assess transfer using unseen material, examine longer-term outcomes, and evaluate deployments across more diverse populations and public contexts.

Acknowledgments

This research is part of the *Voice of Wisdom* project and received funding from dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] Fakhar Abbas and Araz Taeihagh. 2024. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications* 252 (2024), 124260. doi:10.1016/j.eswa.2024.124260

[2] Samer Al-khazraji, Hassan Hadi Saleh, Adil Ibrahim Khalid, and Israa Adnan Mishkhal. 2023. Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics* (2023). https://api.semanticscholar.org/CorpusID:264186108

[3] Safinah Ali, Daniella DiPaola, Irene Lee, Jenna Hong, and Cynthia Breazeal. 2021. Exploring Generative Models with Middle School Students. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445226

[4] Florian Alt, Daniel Buschek, David Heuss, and Jörg Müller. 2021. Orbuculum - Predicting When Users Intend to Leave Large Public Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 46 (March 2021), 16 pages. doi:10.1145/3448075

[5] Gilbert Beyer, Florian Alt, Jörg Müller, Albrecht Schmidt, Karsten Isakovic, Stefan Klose, Manuel Schiewe, and Ivo Haulsen. 2011. Audience Behavior Around Large Interactive Cylindrical Screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 1021–1030. doi:10.1145/1978942.1979095

[6] Lokendra Birla, Trishna Saikia, and Puneet Gupta. 2025. AVENUE: A Novel Deepfake Detection Method Based on Temporal Convolutional Network and rPPG Information. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 17 (Jan. 2025), 16 pages. doi:10.1145/3702232

[7] Joan Bliss, Mike Askew, and Sheila Macrae. 1996. Effective Teaching and Learning: scaffolding revisited. *Oxford Review of Education* 22, 1 (1996), 37–61. arXiv:https://doi.org/10.1080/0305498960220103 doi:10.1080/0305498960220103

[8] Harry Brignull and Yvonne Rogers. 2003. Enticing People to Interact with Large Public Displays in Public Spaces. In *IFIP TC13 International Conference on Human-Computer Interaction*. https://api.semanticscholar.org/CorpusID:45271459

[9] Sherry X. Chen, Alex Christopher Lim, Yimeng Liu, Pradeep Sen, and Misha Sra. 2025. SiCo: An Interactive Size-Controllable Virtual Try-On Approach for Informed Decision-Making. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1815–1825. doi:10.1145/3715336.3735693

[10] Victor Cheung, Diane Watson, Jo Vermeulen, Mark Hancock, and Stacey Scott. 2014. Overcoming Interaction Barriers in Large Public Displays Using Personal Devices. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (Dresden, Germany) *(ITS '14)*. Association for Computing Machinery, New York, NY, USA, 375–380. doi:10.1145/2669485.2669549

[11] Nigel Davies, Sarah Clinch, and Florian Alt. 2014. *Pervasive Displays: Understanding the Future of Digital Signage* (1st ed.). Morgan & Claypool Publishers. doi:10.1007/978-3-031-02484-9

[12] Nigel Davies, Marc Langheinrich, Sarah Clinch, Ivan Elhart, Adrian Friday, Thomas Kubitza, and Bholanathsingh Surajbali. 2014. Personalisation and privacy in future pervasive display networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2357–2366. doi:10.1145/2556288.2557287

[13] Casimer DeCusatis, Erin Alvarico, and Omar Dirahoui. 2022. Gamification of cybersecurity training. In *Proceedings of the 1st International Workshop on Gamification of Software Development, Verification, and Validation* (Singapore, Singapore) *(Gamify 2022)*. Association for Computing Machinery, New York, NY, USA, 10–13. doi:10.1145/3548771.3561409

[14] Alexander Diel, Tania Lalgi, Isabel Carolin Schröter, Karl F. MacDorman, Martin Teufel, and Alexander Bäuerle. 2024. Human Performance in Detecting Deepfakes: A Systematic Review and Meta-Analysis of 56 Papers. *Computers in Human Behavior Reports* 16 (Dec. 2024), 100538. doi:10.1016/j.chbr.2024.100538

[15] Dilrukshi Gamage, Kazutoshi Sasahara, and Jiayu Chen. 2021. The Emergence of Deepfakes and its Societal Implications: A Systematic Review. In *Conference for Truth and Trust Online*. https://api.semanticscholar.org/CorpusID:244716544

[16] Ángel Fernández Gambín, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud, and Youcef Djenouri. 2024. Deepfakes: Current and Future Trends. *Artificial Intelligence Review* 57, 3 (Feb. 2024), 64. doi:10.1007/s10462-023-10679-x

[17] Saul Greenberg, Nicolai Marquardt, Till Ballendat, Rob Diaz-Marino, and Miaosen Wang. 2011. Proxemic Interactions: The New Ubicomp? *Interactions* 18, 1 (Jan. 2011), 42–50. doi:10.1145/1897239.1897250

[18] Keith Raymond Harris. 2021. Video on Demand: What Deepfakes Do and How They Harm. *Synthese* 199, 5 (Dec. 2021), 13373–13391. doi:10.1007/s11229-021-03379-y

[19] Hendrik Heuer. 2025. The Phase Model of Misinformation Interventions. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW190 (May 2025), 28 pages. doi:10.1145/3711088

[20] Uta Hinrichs, Sheelagh Carpendale, Nina Valkanova, Kai Kuikkaniemi, Giulio Jacucci, and Andrew Vande Moere. 2013. Interactive Public Displays. *IEEE Computer Graphics and Applications* 33, 2 (March 2013), 25–27. doi:10.1109/MCG.2013.28

[21] Micol Iannuzzi, Glynis Bogaard, and Jennifer M. Schell-Leugers. 2026. AI-Generated Media and the Erosion of Trust in Legal Settings: A Review on Impact and Distortion of Belief. *AI & SOCIETY* 41, 4 (April 2026), 4045–4055. doi:10.1007/s00146-025-02755-3

[22] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. 2022. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. arXiv:2103.00218 [cs] doi:10.48550/arXiv.2103.00218

[23] Catherine Kerner and Mathias Risse. 2021. Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics* 8, 1 (April 2021), 81–108. doi:10.1515/mopp-2020-0024

[24] Kristian Kiili, Juho Siuko, and Manuel Ninaus. 2024. Tackling Misinformation with Games: A Systematic Literature Review. *Interactive Learning Environments* 32, 10 (2024), 7086–7101. doi:10.1080/10494820.2023.2299999

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 3207–3216. doi:10.1109/CVPR42600.2020.00328

[26] Johan Lundberg and Peter Mozelius. 2025. The Potential Effects of Deepfakes on News Media and Entertainment. *AI & Society* 40 (2025), 2159–2170. doi:10.1007/s00146-024-02072-1

[27] Clara Maathuis, Samuel Derous, and Ebrahim Rahimi. 2024. Design Insights for Developing Deepfakes Awareness Games for Adolescents. *Proceedings of the European Conference on Games-based Learning* 18, 1 (2024), 561–570. doi:10.34190/ecgbl.18.1.2691

[28] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. 2011. Using F-formations to Analyse Spatial Patterns of Interaction in Physical Environments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. ACM, New York, NY, USA, 445–454. doi:10.1145/1958824.1958893

[29] Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. 2021. Fakey: A Game Intervention to Improve News Literacy on Social Media. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 6 (April 2021), 27 pages. doi:10.1145/3449080

[30] Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and design space for interactive public displays. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1285–1294. doi:10.1145/1873951.1874203

[31] Jörg Müller, Robert Walter, Gilles Bailly, Michael Nischt, and Florian Alt. 2012. Looking Glass: A Field Study on Noticing Interactivity of a Shop Window. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 297–306. doi:10.1145/2207676.2207718

[32] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

[33] Doruntina Murtezaj, Viktorija Paneva, Verena Distler, and Florian Alt. 2025. Public Security User Interfaces: Supporting Spontaneous Engagement with IT Security. In *Proceedings of the New Security Paradigms Workshop (NSPW '24)*. Association for Computing Machinery, New York, NY, USA, 56–70. doi:10.1145/3703465.3703470

[34] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research* 154 (2023), 113368. doi:10.1016/j.jbusres.2022.113368

[35] Callum Parker, Martin Tomitsch, and Judy Kay. 2018. Does the Public Still Look at Public Displays? A Field Observation of Public Displays in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (July 2018), 73:1–73:24. doi:10.1145/3214276

[36] Maria Pawelec. 2024. Politische Manipulation und Desinformation. https://www.bpb.de/lernen/bewegtbild-und-politische-bildung/556305/politische-manipulation-und-desinformation/
[37] Thorsten Prante, Carsten Röcker, Norbert Streitz, Richard Stenzel, Carsten Magerkurth, Daniel Van Alphen, and Daniela Plewe. 2003. Hello.Wall–beyond ambient displays. In *Adjunct Proceedings of Ubicomp*. Citeseer, 277–278.
[38] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. 2022. Deepfake Detection: A Systematic Literature Review. *IEEE Access* 10 (2022), 25494–25513. doi:10.1109/ACCESS.2022.3154404
[39] Maurice Ten Koppel, Gilles Bailly, Jörg Müller, and Robert Walter. 2012. Chained Displays: Configurations of Public Displays Can Be Used to Influence Actor-, Audience-, and Passer-by Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, New York, NY, USA, 317–326. doi:10.1145/2207676.2207720
[40] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (2020), 2056305120903408. arXiv:https://doi.org/10.1177/2056305120903408 doi:10.1177/2056305120903408
[41] Robert Walter, Gilles Bailly, and Jörg Müller. 2013. StrikeAPose: Revealing Mid-Air Gestures on Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 841–850. https://doi.org/10.1145/2470654.2470774
[42] Ning Wang, Boxi Fu, Betul Dincer, Omkar Masur, David Faizi, Harshul Ravindran, Julia Wang, Devashish Lai, and Chirag Merchant. 2025. Becoming Fei: An Educational Game for AI and Data Science Education for Novice Learners. In *Learning and Collaboration Technologies*, Brian K. Smith and Marcela Borge (Eds.). Springer Nature Switzerland, Cham, 69–79. doi:10.1007/978-3-031-93746-0_6
[43] Gary L. Wells and Elizabeth A. Olson. 2003. Eyewitness Testimony. *Annual Review of Psychology* 54 (2003), 277–295. doi:10.1146/annurev.psych.54.101601.145028
[44] Mika Westerlund. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9, 11 (2019), 40–53. doi:10.22215/timreview/1282
[45] Niels Wouters, John Downs, Mitchell Harrop, Travis Cox, Eduardo Oliveira, Sarah Webber, Frank Vetere, and Andrew Vande Moere. 2016. Uncovering the Honeypot Effect: How Audiences Engage with Public Interactive Systems. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. ACM, New York, NY, USA, 5–16. doi:10.1145/2901790.2901796
[46] Valdemar Švábenský, Jan Vykopal, Pavel Čeleda, and Ján Dovjak. 2024. Automated Feedback for Participants of Hands-on Cybersecurity Training. *Education and Information Technologies* 29, 9 (2024), 11555–11584. doi:10.1007/s10639-023-12265-8

A Survey

The post-interaction survey contained the following questions and statements. The response format is provided in square brackets.

(1) How would you rate your ability to recognize deepfake videos after playing the game? [5-point Likert scale]
(2) What prior experience have you had with deepfakes? [Multiple choice; "No prior experience with deepfakes" was mutually exclusive]
  (a) I have seen deepfakes (e.g., on social media or in the news)
  (b) I have created deepfakes myself
  (c) I have played games about recognizing deepfakes
  (d) I have no prior experience with deepfakes
  (e) Other [open text field]
(3) I gained new information about recognizing deepfakes through the game. [7-point Likert scale]
(4) Did you read the strategies for recognizing deepfakes presented in the game? [Single choice]
  (a) Yes
  (b) No
(5) If "Yes" was selected in the previous question:
  The strategies were new to me. [7-point Likert scale]
(6) If "No" was selected in Question 4:
  Why did you not read the presented strategies? [Multiple choice]
  (a) I did not have enough time
  (b) I thought I did not need any additional help
  (c) I overlooked or did not notice them
  (d) I wanted to try the game first
  (e) Other [open text field]
(7) Did you play the game more than once? [Single choice]
  (a) Yes
  (b) No
(8) If "Yes" was selected in Questions 4 and 7:
  The strategies I learned in previous rounds improved my ability to recognize deepfakes in later rounds. [7-point Likert scale]
(9) Have you ever played a game on a public display? [Single choice]
  (a) Yes
  (b) No
  (c) I am not sure
(10) The playful format increased my motivation to learn about recognizing deepfakes. [7-point Likert scale]
(11) Which features did you find particularly motivating? [Multiple choice]
  (a) Save-game function
  (b) Detection strategies
  (c) Statistics
  (d) Videos of public figures
  (e) Other features [open text field]
(12) What made you notice the game? [Multiple choice]
  (a) My interest in deepfakes
  (b) The eye-catching animation featuring the Pope or Trump
  (c) The playful format
  (d) I was waiting for someone
  (e) I had spare time
  (f) Other reasons [open text field]
(13) What prior experience have you had with public displays? [Multiple choice; "No prior experience with public displays" was mutually exclusive]
  (a) I have seen public displays (e.g., on public transport or in shopping malls)
  (b) I have interacted with public displays (e.g., for games or information)
  (c) I have no prior experience with public displays
  (d) Other [open text field]
(14) Why did you stop playing the game? [Multiple choice]
  (a) I lost interest in the game
  (b) I had limited time available
  (c) I watched all the videos
  (d) I cannot name a specific reason
  (e) Other reasons [open text field]
(15) How old are you? [Open numeric field]
(16) What is your gender? [Single choice]
  (a) Female
  (b) Male
  (c) Diverse
  (d) Other
(17) What is your highest educational qualification? [Single choice]
  (a) High school diploma or lower
  (b) Bachelor's degree or equivalent
  (c) Master's degree or equivalent

(d) Doctoral degree
(e) Other [open text field]

(18) What is your current employment status? [Single choice]
(a) Student
(b) Employed
(c) Self-employed
(d) Unemployed
(e) Retired
(f) Other [open text field]

(19) What is your field of study or profession? [Single choice]
(a) Computer science
(b) Health sciences
(c) Mathematics or physics
(d) Psychology
(e) Other [open text field]

(20) Do you have any additional feedback or suggestions about the game that you would like to share? [Open text field]

Note. For all statements measured on a 7-point Likert scale, the response options were: *Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree,* and *Strongly agree.*