Journalism & Mass Communication Quarterly
2025, Vol. 102(4) 1135–1156

DOI: 10.1177/10776990251373088
http://journals.sagepub.com/home/jmq


The Short-Term Impact of an On-Site Literacy Intervention on Discerning Deepfake Videos Based on Visual Features

Daniel Vogler[1], Adrian Rauchfleisch[2], and Gabriele de Seta[3]

Abstract

This study examines how well individuals in Switzerland can distinguish high-quality deepfakes from real videos and whether a brief literacy intervention improves detection. In an online experiment with 1,361 participants, we tested the deepfake detection skills and how prior exposure and experience with deepfakes and various forms of media literacy relate to performance. Participants struggled to identify deepfakes when attending to visual features of 10-s clips and the literacy intervention showed no direct effect. However, prior deepfake experience and media literacy moderated the intervention's impact. Findings highlight the need for comprehensive digital literacy strategies to address deepfake-related challenges.

Keywords

deepfake, digital literacy, media literacy, synthetic media, disinformation

Encouraging individuals to embrace new technologies while also educating them about potential risks is a key challenge for modern societies (Huessy, 1979). While technological literacy is typically developed over time through schooling or professional education, new technologies often emerge and integrate into daily life at a much faster pace (McCosker, 2022). Consequently, there is an ongoing debate about how

[1]University of Zurich, Zurich, Switzerland
[2]National Taiwan University, Taipei City, Taiwan
[3]University of Bergen, Bergen, Norway

Corresponding Author:
Daniel Vogler, Research Center for the Public Sphere & Society (fög), University of Zurich, Andreasstrasse 15, Zurich 8050, Switzerland.
Email: daniel.vogler@foeg.uzh.ch

best to educate individuals on technology's opportunities and risks. With the rapid pace of digitization and advancements in artificial intelligence (AI), there is an urgent need to build literacy around new information and communication technologies, including social media, mobile phones, and, more recently, deepfakes (Appel & Prietzel, 2022; Hwang et al., 2021; Koltay, 2011).

Deepfakes are a relatively new phenomenon on the internet (Westerlund, 2019) and belong to the broader category of synthetic media, which includes text, images, video, and audio content generated by various machine-learning models (de Seta, 2024). While deepfake technology offers promising applications in education and entertainment, it also poses potential threats to democratic societies (Ahmed, 2021; Fallis, 2020; Hameleers et al., 2022; Vaccari & Chadwick, 2020). Deepfake technology can be used to manipulate existing media or generate entirely new content to deceive viewers, for example, a fake video of a politician or celebrity appearing to say or do something they never did. As a result, the ability to recognize and differentiate synthetic or manipulated media from real content is becoming an essential skill, complementing general internet and media literacy (Hwang et al., 2021; McCosker, 2022; Shin & Lee, 2022).

With the rapid advancement of AI and deepfake technology, individuals are increasingly confronted with new applications that present both opportunities and risks (Jungherr et al., 2025; Rauchfleisch et al., 2025). Many individuals are still unfamiliar with the term deepfake, have little firsthand experience creating deepfakes, and possess limited knowledge of the technology (Vogler & Rauchfleisch, 2024; Sippy et al., 2024). To address these shortcomings, literacy initiatives and short-term interventions have been proposed, similar to measures against disinformation in online media environments (Guess et al., 2020; Hameleers et al., 2020; Pennycook et al., 2021; Van Der Meer et al., 2023). However, there is still limited research on people's ability to recognize synthetic media, the effectiveness of literacy interventions, and the role of (digital) media literacy and internet skills in this process. In this study, we address this gap in research by analyzing deepfake discernment based on the visual dimension.

In our online experiment, we investigate the capability of the Swiss population to discern deepfakes from real videos based on visual features and whether a short onsite literacy intervention, consisting of basic image-related cues, affects people's ability to recognize deepfakes. We also investigate if news media literacy, social media literacy, and internet skills, as adjacent skills, as well as prior experience and exposure to deepfakes, moderate the intervention's effectiveness. Drawing on transfer learning theory (Perkins & Salomon, 1992), we propose that existing knowledge and experience can be more effectively connected to the new information provided in the literacy intervention. For our study, we used a wide range of existing deepfakes and real videos of well-known politicians and celebrities, which were of good quality (i.e., no cheapfakes, cf. Hameleers, 2024). The deepfakes were tested from a larger pool of videos with a prestudy and grouped into three difficulty levels. Overall, our study gives a comprehensive picture of people's ability to discern deepfakes, the efficacy of an onsite literacy intervention, and the influence of different types of literacy.

Literature Review

Recent advances in AI have made it possible to create high-quality deepfakes that are nearly indistinguishable from real images. As this technology evolves, discerning real from synthetically generated media will become increasingly difficult (Fallis, 2020; Godulla et al., 2021). While deepfake technology has beneficial applications, such as in the film and gaming industries or for personal recreation, it also poses significant risks, particularly in spreading disinformation and causing harm to individuals through deepfake pornography (Birrer & Just, 2024; Hameleers et al., 2022; Vaccari & Chadwick, 2020). To address the risks associated with deepfakes, three key measures are discussed in the literature: state regulation through restrictions or bans, technological detection methods, and (digital) media literacy (Birrer & Just, 2024; McCosker, 2022). To successfully tackle the threat posed by the harmful use of deepfakes, a combination of regulation, technological measures, and literacy is essential. However, large-scale regulations and bans are complex to implement legally, may create economic disadvantages for individual countries, and are often viewed as restrictions on freedom of speech (Bareis & Katzenbach, 2022; Birrer & Just, 2024). Technological detection tools can help, but remain underdeveloped and, like fact-checking for disinformation, often lag behind advances in deepfake production (Bray et al., 2023; Sharma et al., 2024).

This makes literacy measures particularly important to address the risks of emerging AI technologies and deepfakes for individuals and society (McCosker, 2022). Media and information literacy are shown to benefit not only individuals but also society as a whole (Leaning, 2019) and increasingly contain individual skills and social practices for navigating new media environments (McCosker, 2022). Such skills are not static and cannot be learned at a single point in time, such as in school. In dynamic digital media environments, individuals must continuously develop and adapt their media literacy skills in response to new technological innovations, such as deepfake technology or emerging social media platforms (Cho et al., 2024; Leaning, 2019). As a result, digital media literacy is a complex, dynamic, and multi-dimensional concept. Therefore, the literature has begun to explore deepfake literacy, with initial frameworks emerging (Ali et al., 2021; McCosker, 2022; Shin & Lee, 2022; Wagner & Blewer, 2019).

Although challenging to implement (Wagner & Blewer, 2019), deepfake literacy equips individuals with the skills to navigate the risks and opportunities of deepfakes. Strengthening deepfake-related literacy enhances resilience against the misuse of deepfakes for disinformation (Shin & Lee, 2022) and promotes responsible use of the technology. But what does it mean to be deepfake literate? From an operational perspective, this refers to the individual ability to discern false from real content. However, this operational aspect is embedded within a broader context, one that involves understanding and making sense of deepfakes, as well as the digital and cultural environments in which they are produced and circulated (McCosker, 2022). Therefore, deepfake literacy includes critically evaluating information, verifying sources and contextual details, knowledge of sociocultural aspects of new media, and

understanding the potential, limitations, and production processes of deepfakes. Key competencies, such as knowledge of algorithms and AI, align with broader internet and new media skills (Hargittai & Hsieh, 2012; Hargittai et al., 2019; Koc & Barut, 2016; Tandoc et al., 2021).

Literacy Intervention

Building up literacy takes time and often collides with the fast development of technology. Therefore, related to the potential threats of deepfake on-site literacy interventions are being discussed as short-term measures to educate individuals and strengthen their literacy (Hwang et al., 2021; Iacobucci et al., 2021). Such interventions include simply raising awareness of the existence and risks of deepfake technology (Iacobucci et al., 2021; Ternovski et al., 2022), explaining how the technology works (Hwang et al., 2021), or providing strategies and specific hints on how to discern deepfakes from real videos (Bray et al., 2023). A key advantage of these interventions is their ability to reach a wide audience and be implemented on platforms where users commonly encounter deepfakes, such as social media or video-sharing sites. However, their short- and long-term effectiveness is debated, with concerns about potential backfire or downstream effects. First, the dynamic and fast-paced development of deepfake technology has raised doubts about the effectiveness of interventions that rely on specific visual cues (McCosker, 2022). Image-based strategies for detecting deepfakes can quickly become outdated, potentially giving individuals a false sense of security. Thus, timely, adaptable interventions that can be regularly updated may offer an advantage over fixed, once-learned approaches. Second, individuals exposed to literacy interventions may become overly skeptical of visual content online, leading them to misidentify real videos as deepfakes (Ternovski et al., 2022). Indeed, research has shown that warnings, such as labels, generally lower trust in information, which can negatively affect the credibility of true content as well (Freeze et al., 2021; Van Der Meer et al., 2023). Ternovski et al. (2022) tested the effect of such warnings in the context of deepfake videos, finding that they did not improve detection ability and instead increased distrust in real videos. Similarly, de Seta (in press) described how consistent platform warnings on synthetic media do not seem to impact the critical evaluation of content, as most users do not pay attention to them or actively seek AI-generated content for enjoyment. Most critically, Hameleers and Van der Meer (2023) demonstrated that general misinformation literacy interventions could, under certain circumstances, backfire.

Despite skepticism about the effectiveness of interventions and warnings, research on disinformation suggests that even brief literacy interventions can significantly improve recognition skills (Guess et al., 2020). It is also crucial to distinguish between different types of interventions, as Van Der Meer et al. (2023) also show that literacy interventions, unlike general misinformation warnings, do not produce adverse spillover effects.

To test the effect of a literacy intervention, we conducted a survey experiment to assess whether a short on-site literacy intervention (cf. Potter, 2013) enhances deepfake detection abilities. We hypothesize that individuals who receive brief instructions on recognizing deepfakes based on visual clues will be better at identifying them than individuals who did not receive such an intervention (Hwang et al., 2021):

H1: A deepfake literacy intervention will increase discernment of deepfake versus real videos, more so than in the absence of the intervention[1].

As a competing alternative, the intervention may instead trigger generalized skepticism and lead to misclassifications. As already mentioned, recent debates in the literature indicate that some forms of interventions might backfire (Guay et al., 2023; Hameleers & Van der Meer, 2023). The most prominent concern is that interventions will increase skepticism and lead to real content being mistakenly classified as false. In the worst case, a literacy intervention triggers generalized doubt: people become more likely to identify deepfakes at the expense of increased suspicion toward authentic material (Guay et al., 2023). Therefore, we also test the following competing hypothesis:

H2: A deep fake literacy intervention will increase the likelihood of incorrectly classifying real videos as deepfakes.

As we gave one group of participants instructions on identifying deepfake videos, we were interested in how participants explained their decision to categorize the videos as real or false. Hameleers et al. (2025) and Jin et al. (2025) show that qualitative and quantitative content analysis of open-ended questions delivers in-depth insights into people's strategies and reasoning when confronted with assessing the plausibility and authenticity of deepfake videos. We therefore ask,

RQ1: What reasons given by participants best explain the correct or incorrect discernment of deepfake and real videos?

Prior News Media Literacy, Social Media Literacy, and Internet Skills

Literacy related to deepfakes is a multi-dimensional concept encompassing the general understanding of media, data, algorithms, and internet skills (McCosker, 2022). Most exposure to and ongoing learning about deepfake technology occurs online, in social and news media environments, where people can gain experience and develop knowledge of the technology (Hameleers et al., 2022). Drawing on Perkins and Salomon's (1992) high-road transfer framework, we anticipate that participants will draw on their broader literacies when tackling deepfake detection. However, we do not expect this to unfold as an automatic, low-road transfer of existing skills. Instead, our targeted literacy intervention should activate high-road transfer. We expect that those with stronger adjacent literacies, namely, news media literacy (Ashley et al., 2013),

social media literacy (Tandoc et al., 2021), and general internet skills (Hargittai et al., 2019; Hargittai & Hsieh, 2012), will benefit more from the intervention than those with weaker foundational skills. Indeed, Hwang et al. (2021) show that general disinformation literacy education "can be as effective as and, in some cases, even more beneficial than deepfake-specific literacy education" (p. 191). We assume that individuals who already possess high literacy skills related to media can better connect and implement the provided deepfake literacy intervention in the experiment. Therefore, we investigate if there are interaction effects between deepfake discernment and participants' prior news media literacy, social media literacy, and internet skills:

RQ2: Do different forms of literacy, such as news media literacy, social media literacy, and internet skills, strengthen the effect of the literacy intervention on the likelihood of correctly discerning deepfakes from real videos?

Exposure and Individual Experience

As an additional exploratory analysis, we investigate the role of prior exposure and experience with deepfakes. Many people have little to no experience with deepfake technology or may not even know it exists (Vogler & Rauchfleisch, 2024; Sippy et al., 2024). Therefore, the first step in adapting to new technology and developing literacy skills is understanding how the technology works and how to use it. Research indicates that knowledge of deepfakes and prior exposure to them enhance people's ability to recognize such content (Shin & Lee, 2022). Furthermore, we anticipate that individuals with firsthand experience of deepfakes, and thus greater familiarity with the technology, will benefit more from the literacy intervention due to transfer learning (Perkins & Salomon, 1992). However, as there is no clear evidence on the interaction between prior experience or exposure and literacy interventions when it comes to deepfake discernment, we pose the following research question:

RQ3: Does prior experience with and exposure to deepfakes strengthen the effect of the literacy intervention on discerning deepfakes from real videos?[2]

Method

Prestudy

We first conducted a prestudy in June 2023 ($n = 753$) with a Swiss online panel from Respondi-Bilendi, which was also used for the main study. In the prestudy, we assessed the difficulty of correctly discerning each video by calculating how far the average rating (on a scale from 1 = *clearly real* to 7 = *clearly a deepfake*) was from the correct end of the scale. Videos with mean scores closer to the correct endpoint were considered as easier (see Supplemental Appendix D for more information). We selected a deepfake and a real video of politicians and celebrities. We used short videos of Tom

Cruise, Elon Musk, Barack Obama, Hillary Clinton, Volodymyr Zelenskyy, and Vladimir Putin. For each person, we identified one deepfake and one real video that was approximately 10seconds long. The videos featured individuals in neutral speaking settings, either positioned behind a desk or lectern (in the case of politicians) or recorded as selfie videos. All clips showed only the upper body and included no significant gestures beyond natural, coordinated speaking movements. The videos were implemented as Graphics Interchange Format (GIF) without any audio features. We exported all GIFs at a standard-definition size (around 640 × 440) that preserves key facial and contextual details while keeping file sizes manageable. This format was chosen to foreground purely visual features of authenticity, making it more likely that people could participate in the study as the GIF format ensured quick loading and broad compatibility across devices and browsers. In the pretest, participants had to rate with the same scale as in the main study if a video was a deepfake or a real video (1 = *clearly real*; 7 = *clearly a deepfake*). We ran the pretest with these videos at the end of another study, which allowed us to use a large sample size of 753 participants and around 376 observations for each video. We selected six videos for our main study based on the pretest results.

Each participant in the pretest was randomly assigned either the deepfake video or the real video for each of the six people. We selected videos for our main study based on three criteria. First, we chose clips spanning varying difficulty levels (see Supplemental Appendix D for more information). Second, we wanted to select three fake and three real videos, each featuring one of six different celebrities or politicians. Third, we measured familiarity (on a 7-point scale) and asked whether people believe they know the video already (binary). Each participant then viewed one randomly assigned video per public figure. We used regression models to identify and exclude videos where familiarity was a significant predictor. We then added prior exposure as a predictor: in almost all cases, it was nonsignificant, and where significant, the effect was in the opposite direction. Thus, prior knowledge did not confound our test set.

Main Study Sample

Our main study was approved by the Ethics Committee of the Faculty of Arts and Social Sciences of the University of Zurich and preregistered (https://aspredicted.org/9z6s-gfx9.pdf). For our online experiment, we recruited participants from the online panel (Respondi-Bilendi). Participants are people living in Switzerland who are above 16 years. Participants from the pretest sample were not invited to this study. The sample includes participants from both the French and German language regions. The surveys were both programmed using Unipark software. As we sampled without quotas, specific age brackets are overrepresented by female participants. We thus calculated survey weights based on the Swiss population and estimated the models using these weights. The results are identical to those without survey weights. More importantly, all of our strata are saturated, which is the most crucial factor when examining correlational relationships (see Supplemental Appendix C.3 for complete age distribution and models with survey weights).

Before we started our main study, we estimated the required sample size to have enough statistical power to test our hypotheses and research questions. Our power simulations indicated a power of 97% for the more complex hypotheses and 100% for the main hypotheses with a sample size of 1,200 (see Supplemental Appendix B). The main study took place in September 2023. Our final sample size was 1,361.[3] Initially, people were randomly assigned to either the literacy intervention or the control group. We then asked them basic socio-demographic questions about their general social media use. After that, we measured the different literacy and skill scales. Then we asked about prior experience and exposure to deepfakes. We also included two attention checks in the first part of the questionnaire before the treatment. The first, placed at the beginning, tested whether participants could view GIFs on their device using a multiple-choice question. The second was an instructional manipulation check that asked participants to select a specific response. Inattentive participants were directly filtered out when they failed one of them. Participants who were assigned to the literacy intervention group then received the literacy intervention before starting to rate the six videos, which were always shown in randomized order to each participant. Participants had to stay 20seconds on the page with the literacy intervention before they could continue with the test. Participants in the control group proceeded to the videos without receiving the intervention. As in the pretest, each of the six test clips was shown as a controlled, ten-second visual excerpt in GIF format without any audio features, standardized in size and resolution, so participants' judgments relied on visual markers alone. In contrast to the pretest, the participants rated all six videos.

Analytical Strategy

The literature presents various approaches to testing interventions. Guay et al. (2023) showed that it is crucial to include both true and false headlines in the same model as an interaction term with the treatment, rather than evaluating only one type (e.g., fake news). This approach is still rare in communication science. However, it allows us to see whether an intervention improves discernment or merely heightens skepticism at the expense of correctly identifying real headlines. For example, Pennycook et al. (2021) applied this design when testing the impact of accuracy prompts on fake-news sharing.

Independent Variables

The literacy intervention provided guidance on identifying deepfake videos based on visual features, emphasizing movements, depicted details, and contextual clues as the three key factors for detection. Drawing on the MIT Media Lab's (2023) deepfake detection guidelines, we structured our intervention around three visual-dimension factors—movement, details, and context. This tip-driven approach mirrors the style of Guess et al. (2020) and Hameleers (2022) and aligns closely with how literacy interventions are offered today (AI for Education, 2024). For each factor, we provide two specific cues ("fixed eye gaze" and "floating facial features" for

movements, "unrealistic fine details" and "blurred backgrounds" for depicted details, and "unfamiliar appearance" and "unexpected behavior" for context). To maintain a real setting, the instructions did not include specific hints directly related to any of the deepfake videos used in our experiment (see Supplemental Appendix A.2 for the exact wording of the literacy intervention).

For **RQ1**, we asked participants whether there was a specific reason for the rating decision ("Was there a specific reason for your decision on whether the video is a deepfake or real?"). We then used ChatGPT (for more details, see Supplemental Appendix A.3) to classify the answers. We used the four categories: context (refers to the situation in which a person is shown or aspects of the background), movement (relates to any motion or movement of the person), details about the person (includes any specific feature or characteristic of the person), and no answer/not applicable. We used these categories as they broadly connect with the aspects discussed in the literacy intervention. Two authors validated the classifier manually with a random sample of 50 unique answers. The classifier reached satisfactory inter-coder reliability (Krippendorff's α of .84 for the two human coders and the automatic classification).

Media literacy as variable for **RQ2** was measured with six items that we took from Ahsley et al. (2013) and combined into a mean index. The items covered the dimensions "authors and audiences" (e.g., "the owner of a media company influences the content that is produced"), "messages and meanings" (e.g., "people are influenced by news whether they realize it or not"), and "representations of reality" (e.g., "news makes things more dramatic than they really are"). Social media literacy (**RQ2**) was also measured with six items that were combined to a mean index. We used the items from Tandoc et al. (2021) that covered technical (e.g., "I know how to post content, such as photos, on my social media"), informational (e.g., "I know how to verify whether what is shared on social media is correct"), and privacy-related (e.g., "social media sites such as Facebook control what I see on social media") competences. For **RQ2**, we also measured internet skills with six items (level of understanding of six internet-related terms, e.g., advanced search, phishing) that we selected from an established scale by Hargittai and Hsieh (2012).

Prior experience with deepfakes as variable for **RQ3** was measured with a sum index (0–4) based on four categories that participants could select (e.g., "I have heard about deepfakes" or "I have created deepfakes"). Prior exposure to deepfakes (**RQ3**) was measured with three items asking how often people have encountered deepfakes on social media platforms, messenger apps, video platforms, and news websites (see Table 1 for an overview of all variables).

We added gender (male vs. other), age, educational attainment (university degree vs. other), and language region (French vs. German) as covariates to all of our models.

Dependent Variable

As the outcome variable for all hypotheses and research questions, we used participants' evaluations of each video, asking them to rate it on a scale from 1 (*clearly real*)

Table 1. Descriptive Statistics of All Variables.

Variable	*M* (*SD*)	*n*
H1-H2/RQ2/RQ3 Treatment	49.22%	1,361
RQ3: Prior experience with deepfakes? (sum index)	1.11 (0.56)	1,361
RQ3: Prior exposure to deepfakes (3 items, $\alpha = 0.8$)	3.81 (1.54)	1,361
RQ2: Media literacy (6 items, $\alpha = 0.74$)	5.83 (0.77)	1,361
RQ2: Social media literacy (6 items, $\alpha = 0.8$)	5.36 (1.17)	1,361
RQ2: Internet skills (6 items, $\alpha = 0.85$)	4.35 (1.47)	1,361
Outcome variable: Tom Cruise (fake)	4.01 (2.09)	1,361
Outcome variable: Barack Obama (fake)	4.06 (2.03)	1,361
Outcome variable: Hillary Clinton (real)	3.16 (1.89)	1,361
Outcome variable: Volodymyr Zelenskyy (fake)	4.98 (1.94)	1,361
Outcome variable: Elon Musk (real)	4.00 (1.95)	1,361
Outcome variable: Vladimir Putin (real)	3.68 (2.01)	1,361
University degree	27.63%	1,361
Gender male	36.00%	1,361
Region (French)	33.28%	1,361
Age	43.24 (16.28)	1,361

to 7 (*clearly a deepfake*). To assess the effect of the literacy intervention on discernment, the extent to which individuals distinguish between real and deepfake videos, we included an interaction between a dummy variable indicating if someone has received the literacy intervention and a dummy variable indicating video type (real vs. deepfake). The coefficient on this interaction term then reflects participants' ability to discriminate between real and deepfake videos. As our data are nested, we used a linear mixed-effects model with varying intercepts for participants and videos. The discernment ability was then tested as an interaction term with video type (real or deepfake). For all models except the one for **H2**, which includes only true videos, all variables were added as interaction terms with video type. As described above, the key values of interest are the estimates for these interaction terms, as done in prior discernment studies (Guay et al., 2023; Pennycook et al., 2021).[4]

Results

Overall, our test with the real and the deepfake videos showed that participants had difficulties distinguishing real from deepfake videos. With an average rating of 4.00 ($SD = 1.95$), the most difficult real video (of Elon Musk) had almost the same rating as the most difficult deepfake (of Tom Cruise), with an average rating of 4.01 ($SD = 2.09$). The other videos received average ratings between 3.16 and 4.98. While these scores indicate that videos were, on average, not clearly identified, there is an overall difference, t-Welch $(8,147.8) = 16.39$, $p < .001$, between the average rating for real ($M = 3.61$, $SD = 1.98$) and fake videos ($M = 4.35$, $SD = 2.07$).

To test our first hypothesis, we used a multilevel model with varying intercepts for videos and participants (see Supplemental Appendix C.1 for the complete models

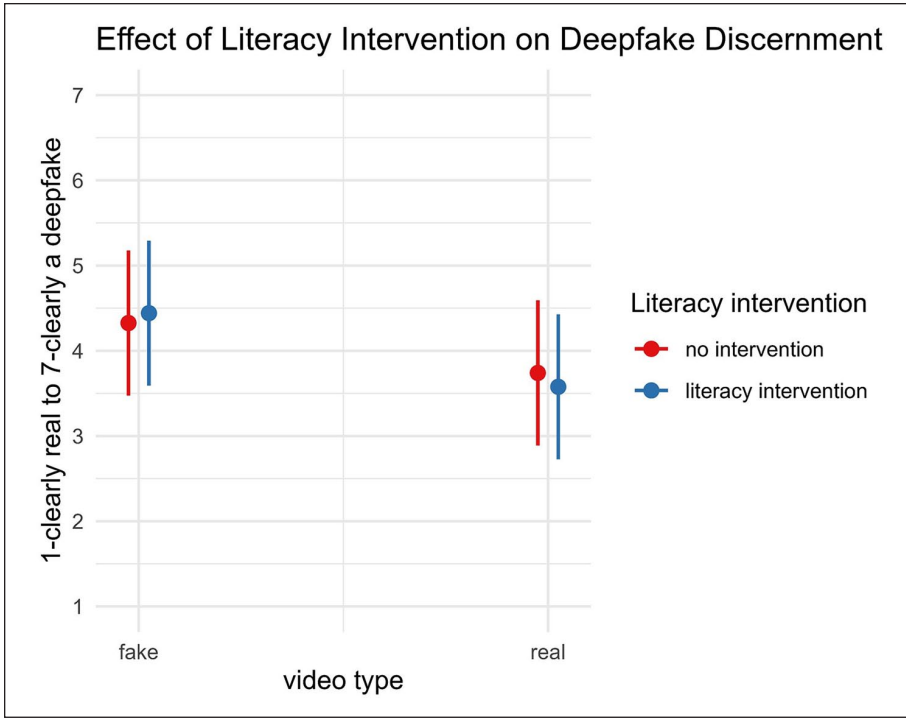

Figure 1. Interaction Between Video Type and Literacy Intervention. The Interaction Effect Is Not Significant.

reported in the main paper).[5] Our data show no effect of the literacy intervention ($b = –0.06$, 95% confidence interval [CI] = [–0.22, 0.10], $p = .452$). Respondents who received the intervention did not perform better in discerning deepfakes from real videos than those who did not (see Figure 1). Thus, **H1** is not supported.[6]

For **H2**, the competing hypothesis to **H1**, we tested only true videos to evaluate whether the intervention might lead to more wrong evaluations as people become overly skeptical. Our data show no such effect ($b = –0.06$, 95% CI = [–0.21, 0.08], $p = .390$). The intervention, therefore, did not lead to a hypercritical assessment of the videos. People who received the literacy intervention were not more likely to rate real videos as deepfakes than people who did not receive a literacy intervention. Thus, our data do not support **H2**.

For **RQ1**, we focused on the optional comment field that we provided for each video, where we asked participants whether there was a specific reason for their decision to classify the video as a deepfake or real video. Overall, respondents who received the literacy intervention were more likely to provide a specific reason for both real and fake videos instead of providing no answer or something not applicable, $\chi^2(3, N = 8{,}166) = 89.66$, $p < .001$. Comments mentioning the person's movement in the video and specific details about the person were more often mentioned by participants who received the literacy intervention (see Figure 2). Context, however, was

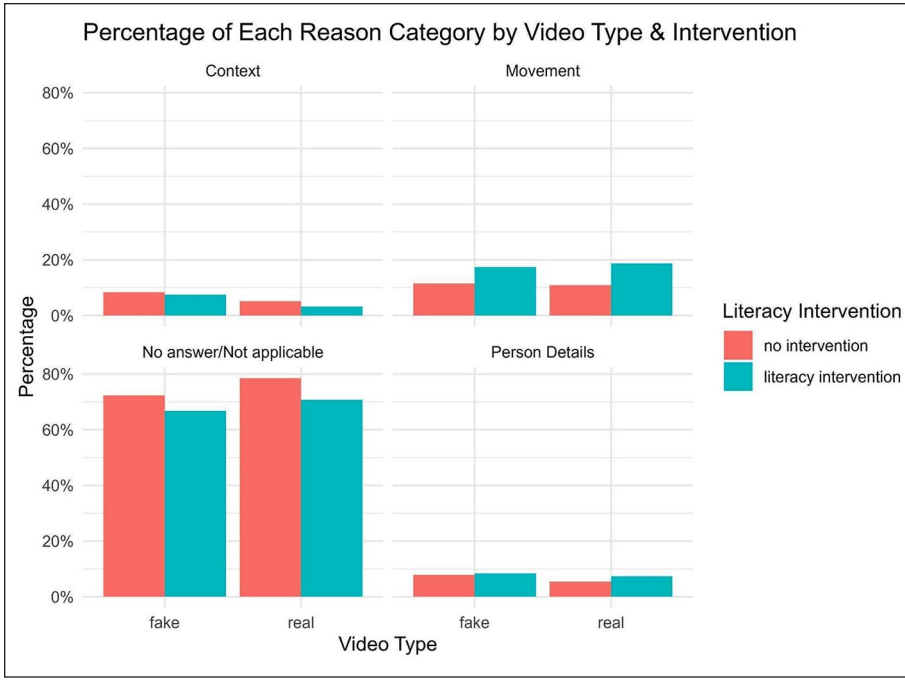

Figure 2. Comparing Video Type and Intervention With How Often No Answer or Nothing Applicable Was Provided.
Note. Percentages are calculated within each intervention and video type group. Individual chi-square tests are all significant (Context: $p < .01$; Movement: $p < .001$; No answer/Not applicable: $p < .001$; Person Details: $p < .05$).

slightly more often mentioned by people who had not received the literacy intervention. This analysis further indicates that the intervention prompted participants to apply the guidance actively: for instance, among comments about movement, 27.5% in the intervention group noted eye movements, compared with only 13.1% in the control group.

The Role of Different Literacies in Combination With the Literacy Intervention

As stated in **RQ2**, we also wanted to test whether the three literacy variables strengthen the effect of the literacy intervention. For this analysis, we estimated an additional model with three-way interactions between the literacy variables, the literacy intervention, and the video type. Our results indicate a significant result for news media literacy ($b = –0.24$, 95% CI = [–0.46, –0.03], $p = .027$), but not for social media literacy ($b = 0.01$, 95% CI = [–0.15, 0.16], $p = .919$) or internet skills ($b = 0.13$, 95% CI = [–0.00, 0.25], $p = .055$). The interaction effect with media literacy shows that people

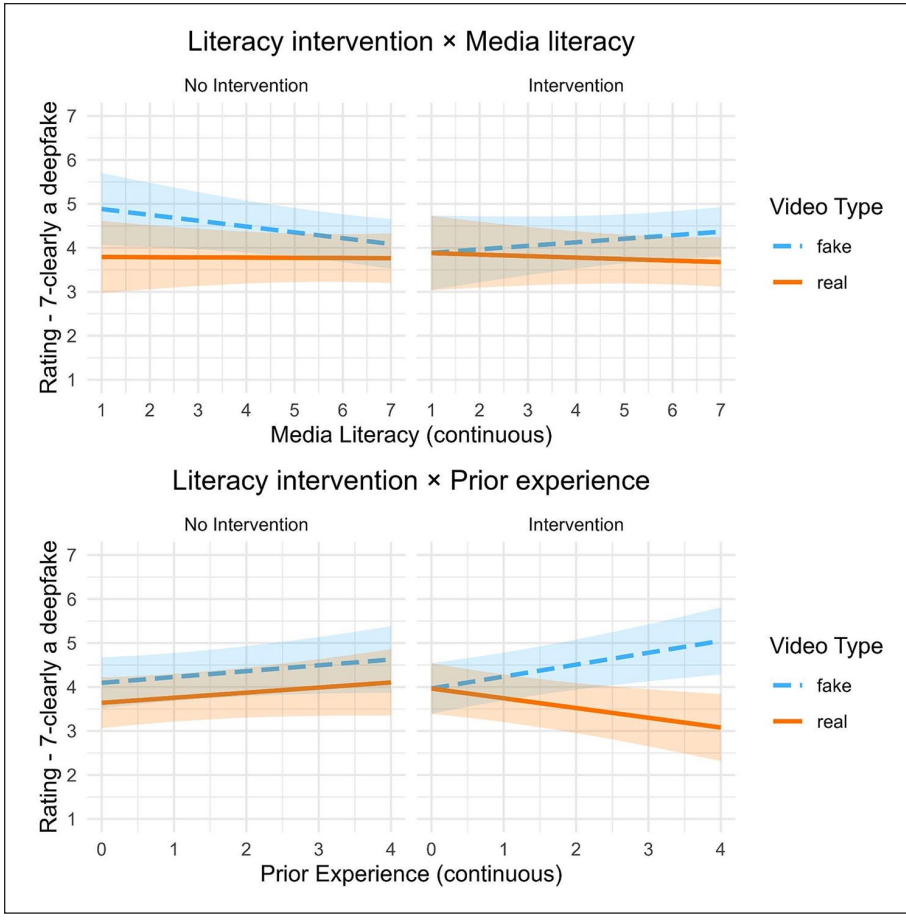

Figure 3. Interaction Between Literacy Intervention, Video Type, and News Media Literacy (Top) and Prior Experience (Bottom).
Note. Both interactions are significant.

with higher news media literacy benefited more from the intervention on discerning deepfake from real videos (see Figure 3).

The Role of Prior Experience and Exposure in Combination With the Literacy Intervention

In **RQ3**, we examined whether prior experience with deepfakes or prior exposure to them moderates the effect of the literacy intervention on discernment. The results show no significant interaction between prior exposure and the intervention (b = –0.04, 95% CI = [–0.15, 0.07], p = .461). In contrast, there is a significant interaction

between prior experience and the intervention ($b = –0.48$, 95% CI = [–0.77, –0.18], $p = .002$), indicating that participants with more firsthand experience of deepfakes benefited more from the literacy intervention (see Figure 3).

Additional Analyses With Knowledge of Video and Familiarity With the Person in the Video

We also conducted two non-preregistered follow-up analyses (for the model with these additional analyses, see Supplemental Appendix C.1.2). We asked participants whether they had seen the video before the study. For 8.9% of the observations, people indicated that they knew the video. The observations for videos people believed they knew are equally distributed for false and true videos, $\chi^2(1, N = 8{,}166) = 0.256$, $p = .613$. Therefore, we tested whether people who indicated they knew the video could correctly identify it. Knowing the video did not influence the discernment ability ($b = –0.04$, 95% CI = [–0.33, 0.24], $p = .778$). While some people might have seen one of the deepfake or real videos before, and thus more likely correctly rated it, our results indicate that it is also very likely that many of the 8.9% observations, where people indicated they knew the video, was a misperception.

As a second additional analysis that we have not preregistered, we added familiarity with the depicted person as an additional predictor to the model. Our analysis shows that with higher familiarity, the discernment ability increases. Thus, fake and real videos are more likely to be correctly classified ($b = –0.06$, 95% CI = [–0.10, –0.01], $p = .016$).

Discussion

Using an online experiment, this study investigated the ability of people in Switzerland to discern deepfake videos and tested whether a situational literacy intervention could improve this ability. Furthermore, the study examined whether people's news media literacy, social media literacy, and internet skills, as well as their prior exposure and experience of deepfake technology, affect the effect of the literacy intervention. Our experimental study shows that people in Switzerland can hardly distinguish high quality deepfake videos from real videos. Similar results were found in other studies, for instance, by Bray et al. (2023), who concluded that "overall detection accuracy was only just above chance" (p. 1). Age was the only covariate affecting discerning ability, with younger participants being better at correctly identifying real and deepfake videos. As we used well-made deepfakes for our experiment and not low-quality videos or so-called cheapfakes (Hameleers, 2024), our task for participants was difficult. In addition, the stimuli in our experimental setup excluded audio features and focused only on the persons' upper body. Therefore, the participants only had access to a limited set of visual cues both for the fake and real videos, making them very similar. Overall, the results indicate that with the right quality, a deepfake is hardly distinguishable from a real video, highlighting the need for new strategies to educate users on navigating this emerging internet phenomenon.

Our literacy intervention, which consisted of a brief assistance for recognizing deepfakes based on visual cues given to participants immediately before showing the videos, had no effect on discernment skills. This finding aligns with the study by Bray et al. (2023), who did not find any effect of a literacy intervention similar to ours, that is, providing specific hints related to abnormalities, asymmetries, and the background of the videos. In contrast to other studies (e.g., Ternovski et al., 2022), we found no evidence of a negative downstream effect of the intervention. The literacy intervention did not cause respondents to become overly critical or mistakenly identify real videos as deepfakes. This may be due to the different focus of our intervention, which emphasized image-specific aspects without including initial warnings (priming) about the dangers of deepfakes, as in Ternovski et al. (2022) or the general warning literacy intervention by Hameleers and Van der Meer (2023). Therefore, future research could explore how different types of interventions influence deepfake discernment. In addition, the explorative analysis of the open-ended question on participants' reasons for classifying a video as real or fake shows that a literacy intervention can at least direct people's attention to specific details, thereby also showing that people received our intervention and tried to apply it for their decision-making.

Our study further shows that participants' self-reported firsthand experience with deepfakes (i.e., in our case, heard of, saw, shared, or created a deepfake) and people with high prior news media literacy benefited more from the intervention. At least for some people, short literacy interventions could thus be helpful for deepfake discernment, possibly through a form of transfer learning (Perkins & Salomon, 1992) where existing knowledge could be connected with the new information provided in the literacy intervention. This idea is also supported by the finding that prior experience increased the effect of the literacy intervention. As we did not find any negative downstream effects, literacy interventions might still be worthwhile pursuing, even if they only work for specific groups. However, only **RQ2**, focusing on the different forms of literacies, was preregistered, and **RQ3** is more exploratory. Thus, this finding should be taken as first evidence, and future research could help to clarify this with more specific designs that isolate this potential connection. Given the positive role of firsthand experience with deepfakes in combination with literacy interventions, future research could explore whether repeated exposure to deepfakes in a controlled setting enhances individuals' ability to discern them. Such inoculation effects were demonstrated by Basol et al. (2020), who used an online game as an anti-misinformation intervention. The authors showed that playing *Bad News*, an online game through which users learn about techniques for identifying misinformation, improves participants' ability to spot misinformation.

The interaction effects of media literacy and prior experience reveal noteworthy patterns in the control group (see Figure 3). For prior experience, we observe an increase in ratings for both true and false videos as experience grows. Thus, participants with higher experience are more inclined to rate any video as deepfake, though their ability to distinguish between real and fake remains constant across experience levels. Looking at media literacy, the detection of fake videos decreases as media literacy increases in the control group, resulting in poorer overall discernment. The most

plausible explanation is that participants with high media literacy skills focused heavily on contextual cues without guidance from the literacy intervention. However, context was not a reliable cue for identifying fakes in most of the videos used for this study. This may have led to misjudgments as participants applied their media literacy skills via low-road transfer. This tendency only reversed into a positive effect when supplemented by the literacy intervention.

Overall, these results suggest that an intervention to improve deepfake discernment focused solely on visual cues may be insufficient; a more holistic approach that enhances adjacent literacies, such as general media literacy, could foster greater skepticism without producing negative downstream effects. A recent study by Jin et al. (2025) highlights the role of specific visual literacy, which consists of the ability to interpret and create visual content, for deepfake discernment. Future studies could investigate the role of visual literacy in combination with literacy interventions like the one applied in this study.

Our findings are not surprising, as they conform to longer cycles of panic, skepticism, and integration described by existing studies of technology. Accounts of the momentous transformations and restructuring in people's media consumption practices—from the printing press (McLuhan, 1962) to television (Meyrowitz, 1985) and the internet (Gitelman, 2008)—all include early stages of adoption in which new literacies are acquired organically and pre-existing literacies complement one another. People mainly learn to establish or at least question the veracity and authenticity of content in an organic way through exposure and ongoing engagement, so it is expected that a new form of deceptive media (Natale, 2024) will trigger widespread panics and skeptical reactions which in turn support dystopian journalistic narratives (Yadlin-Segal & Oppenheim, 2021) before people acquire the necessary literacy to make sense of them.

In comparison with previous cycles of deceptive media, such as the popularization of digital image manipulation, which could be mainly addressed through editorial policies and authorial responsibility (Pantti & Sirén, 2015), deepfakes and synthetic content are likely to become so pervasive and easy to create that verification requires high cognitive and fact-checking costs. In this scenario, people's expectations of authenticity are likely to change, and content is assumed to be synthetic by default, with authenticity becoming the exception (Bajohr, 2023). The present study has examined Swiss attitudes in the early historical stages of deepfake circulation, and future studies could analyze the diminishing value of authenticity as synthetic media becomes more and more widespread.

Conclusion

This study advances research on deepfakes in three ways. First, it provides robust evidence that a brief, literacy intervention focusing on visual aspects does not improve overall discernment of high-quality deepfakes but also does not trigger generalized skepticism. Second, it identifies news media literacy and prior deepfake experience as factors that strengthen the impact of such interventions, highlighting the role of

adjacent literacies and transfer learning. Third, it shows that even without measurable gains in accuracy, interventions can shape the cues people attend to when evaluating video content.

However, our study also has some limitations, particularly regarding the stimuli used. This resulted from a trade-off between creating a realistic scenario and maintaining a controlled experimental setting with minimal influencing factors. We used already existing short portrait videos (deepfakes and real) of internationally known politicians or celebrities (none of them Swiss) in a neutral or common setting without much visual background information (e.g., politicians sitting at a desk or on stage in front of a microphone, celebrity in a casual setting at home). Therefore, the video's setting was very realistic and did not deliver much context, which was an advantage because it made the videos comparable and minimized the effects of prior knowledge about potentially well-known deepfakes. However, in reality, many deepfakes actually depict individuals in a highly controversial or surprising situation (e.g., Pope Francis wearing a Balenciaga-style white puffer jacket), providing context that may raise suspicion among users and improve the ability to discern deepfakes. The study also focused solely on the visual aspects of deepfakes and did not include sound. This approach helped minimize potential influencing factors in deepfake detection, ensuring a more controlled experimental setting. However, implausible or unrealistic statements, which often are at the core of deepfake disinformation (e.g., Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons), typically affect deepfake detection (Hameleers et al., 2024).

Furthermore, as previously mentioned, we used only very high-quality deepfakes. This choice again addresses a realistic scenario: As deepfake technology advances, producing high-quality deepfakes will become increasingly easier. However, many deepfakes on the internet are still of low quality and are often referred to as cheapfakes. As cheapfakes were perceived as more credible than deepfakes of good quality in an experimental study by Hameleers (2024), whether the effectiveness of a literacy intervention differs when applied to deepfakes of varying quality, including cheapfakes, is a question for further research.

In addition, the study was conducted in a single country using deepfakes that were not specifically related to the Swiss context. One reason for this choice was the lack of publicly available deepfakes of Swiss politicians or celebrities. However, deepfakes created for disinformation or other harmful purposes are often designed to exploit specific contexts and influence targeted (regional) audiences. Indeed, our study shows that familiarity with the person has a positive effect on deepfake recognition. Therefore, it would be valuable to investigate whether an individual's ability to detect (high-quality) deepfakes and the effect of a corresponding literacy intervention depend on whether local or foreign politicians or celebrities are depicted.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: D.V.'s research was funded by TA-SWISS, the Swiss Foundation for Technology Assessment. A.R.'s work was supported by the National Science and Technology Council, Taiwan (R.O.C.) (Grant No. 114-2628-H-002-007-) and by the Taiwan Social Resilience Research Center (Grant No. 114L9003) from the Higher Education Sprout Project by the Ministry of Education in Taiwan. G.d.S.'s work was supported by a Trond Mohn Foundation Starting Grant (TMS2024STG03).

ORCID iDs

Daniel Vogler https://orcid.org/0000-0002-0211-7574
Adrian Rauchfleisch https://orcid.org/0000-0003-1232-083X
Gabriele de Seta https://orcid.org/0000-0003-0497-2811

Supplemental Material

Supplemental material is available in the online version of the article.

Notes

1. We changed the wording to clearly indicate that we test for discernment. Furthermore, we added **RQ3** as a non-preregistered research question instead of some of the hypotheses.
2. We added this non-preregistered research question instead of using these variables without interacting with the literacy intervention. The original hypotheses were formulated for only the control group, which did not provide sufficient statistical power. We explain this decision in more detail in Supplemental Appendix C.2.
3. Our sample size is larger than the preregistered sample size of 1,200, which is an accidental deviation from our preregistration. Due to a miscommunication, we did not set an automatic stop, and the survey company allowed data collection to exceed this target, resulting in a larger final sample.
4. The code and data to replicate our analysis are available at https://osf.io/tdxgb/?view_only=286ab96e89244bdba0750db09bb8d9fd
5. We deviated from the preregistration by testing **H1** to **H4** and **H6** to **H8** together in a single model. We explain our decision (not enough power with only half the sample size) in Supplemental Appendix C.2 and also report additionally all models as described in the preregistration.
6. We also checked in non-preregistered exploratory analysis the effect of the literacy intervention for each video (for all models, see Supplemental Appendix C.4). Only for one video, the literacy intervention showed a significant effect (the real Hillary Clinton video). For the other videos, the literacy intervention was not significant. Still, in an overall model excluding the worst-performing fake (Tom Cruise) and real (Vladimir Putin) video, the literacy intervention is significant.

References

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, *57*, Article 101508. https://doi.org/10.1016/j.tele.2020.101508

AI for Education. (2024). *Uncovering deepfakes: Classroom guide + discussion questions*. https://www.aiforeducation.io/ai-resources/uncovering-deepfakes

Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers, and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence*, *2*, Article 100040. https://doi.org/10.1016/j.caeai.2021.100040

Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, *27*(4), Article zmac008. https://doi.org/10.1093/jcmc/zmac008

Ashley, S., Maksl, A., & Craft, S. (2013). Developing a news media literacy scale. *Journalism & Mass Communication Educator*, *68*(1), 7–21. https://doi.org/10.1177/1077695812469802

Bajohr, H. (2023). Artificial and post-artificial texts: On machine learning and the reading expectations towards literary and non-literary writing. *Basel Media Culture and Cultural Techniques Working Papers*, *7*, 1–31. https://doi.org/10.12685/bmcct.2023.007

Bareis, J., & Katzenbach, C. (2022). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values*, *47*(5), 855–881. https://doi.org/10.1177/01622439211030007

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *32*(1), Article 91. https://doi.org/10.5334/joc.91

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241253138

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

Cho, H., Cannon, J., Lopez, R., & Li, W. (2024). Social media literacy: A conceptual framework. *New Media & Society*, *26*(2), 941–960. https://doi.org/10.1177/14614448211068530

de Seta, G. (in press). Celebrity face-swaps and bikini beauties: The circulation of synthetic pornography in China. In S. Udupa & P. Hervik (Eds.), *Handbook on Anthropology and Artificial Intelligence*. Edward Elgar.

de Seta, G. (2024). Synthetic probes: A qualitative experiment in latent space exploration. *Sociologica*, *18*(2), 9–23. https://doi.org/10.6092/ISSN.1971-8853/19512

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, *34*, 623–643. https://doi.org/10.1007/s13347-020-00419-2

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Gitelman, L. (2008). *Always already new: Media, history and the data of culture* (1st paperback ed.). MIT Press.

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

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of*

Sciences of the United States of America, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Hameleers, M. (2022). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the U.S. and Netherlands. *Information, Communication & Society*, *25*(1), 110–126. https://doi.org/10.1080/1369118X.2020.1764603

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research*, *36*(1), Article edae004. https://doi.org/10.1093/ijpor/edae004

Hameleers, M., Meer, T. V., & der Dobber, T. (2025). How far can political deepfakes credibly deviate from reality? Responses to political deepfakes with varying degrees of deception. *International Journal of Communication*, *19*. https://ijoc.org/index.php/ijoc/article/view/22704

Hameleers, M., & van der Meer, T. (2023). Striking the balance between fake and real: under what conditions can media literacy messages that warn about misinformation maintain trust in accurate information? *Behaviour & Information Technology. Advance online publication*. https://doi.org/10.1080/0144929X.2023.2267700

Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

Hameleers, M., Van Der Meer, T. G. L. A., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society*, *8*(3), Article 211163. https://doi.org/10.1177/20563051221116346

Hameleers, M., Van Der Meer, T. G. L. A., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, *152*, Article 108096. https://doi.org/10.1016/j.chb.2023.108096

Hargittai, E., & Hsieh, Y. P. (2012). Succinct survey measures of web-use skills. *Social Science Computer Review*, *30*(1), 95–107. https://doi.org/10.1177/0894439310397146

Hargittai, E., Piper, A. M., & Morris, M. R. (2019). From internet access to internet skills: Digital inequality among older adults. *Universal Access in the Information Society*, *18*(4), 881–890. https://doi.org/10.1007/s10209-018-0617-5

Huessy, H. R. (1979). Dealing with the negative social consequences of technology. *Technology in Society*, *1*(3), 193–203. https://doi.org/10.1016/0160-791X(79)90022-8

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 194–202. https://doi.org/10.1089/cyber.2020.0149

Jin, X., Wang, G., Zhang, Z., Zhou, W., Yu, N., Gao, B., & Gao, S. (2025). Identifying individual differences in deepfake discernment: the effects of cognitive disposition and visual literacy. Information, Communication & Society. Advance online publication. https://doi.org/10.1080/1369118X.2025.249690

Jungherr, A., Rauchfleisch, A., & Wuttke, A. (2025). Artificial Intelligence in Election Campaigns: Perceptions, Penalties, and Implications. arXiv. https://doi.org/10.48550/ARXIV.2408.12613

Koc, M., & Barut, E. (2016). Development and validation of New Media Literacy Scale (NMLS) for university students. *Computers in Human Behavior*, *63*, 834–843. https://doi.org/10.1016/j.chb.2016.06.035

Koltay, T. (2011). The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture & Society*, *33*(2), 211–221. https://doi.org/10.1177/0163443710393382

Leaning, M. (2019). An approach to digital literacy through the integration of media and information literacy. *Media and Communication*, *7*(2), 4–13. https://doi.org/10.17645/mac.v7i2.1931

McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society*, *26*, 2786–2803. https://doi.org/10.1177/14614448221093943

McLuhan, M. (1962). *The Gutenberg galaxy: The making of typographic man*. University of Toronto Press.

Meyrowitz, J. (1985). *No sense of place: The impact of electronic media on social behavior*. Oxford University Press.

MIT Media Lab. (2023). Detect DeepFakes: How to counteract misinformation created by AI. https://web.archive.org/web/20230818153658/https://www.media.mit.edu/projects/detect-fakes/overview/

Natale, S. (2024). Digital media and the banalization of deception. *Convergence: The International Journal of Research into New Media Technologies*, *31*, 402–419. https://doi.org/10.1177/13548565241311780

Rauchfleisch, A., Vogler, D., & de Seta, G. (2025). Deepfakes or synthetic media? The effect of euphemisms for labeling technology on risk and benefit perceptions. Social Media + Society. Advance online publication. https://doi.org/10.1177/20563051251350975

Pantti, M., & Sirén, S. (2015). The fragility of photo-truth: Verification of amateur images in Finnish newsrooms. *Digital Journalism*, *3*(4), 495–512. https://doi.org/10.1080/21670811.2015.1034518

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén, & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 425–441). Pergamon.

Potter, W. J. (2013). Review of literature on media literacy: Media literacy. *Sociology Compass*, *7*(6), 417–435. https://doi.org/10.1111/soc4.12041

Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*, *84*, 22187–22229. https://doi.org/10.1007/s11042-024-19906-1

Shin, S. Y., & Lee, J. (2022). The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes. *Digital Journalism*, *10*(3), 412–432. https://doi.org/10.1080/21670811.2022.2026797

Sippy, T., Enock, F., Bright, J., & Margetts, H. Z. (2024). *Behind the deepfake: 8% create; 90% concerned. Surveying public exposure to and perceptions of deepfakes in the UK* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.05529

Tandoc, E. C. J., Yee, A. Z. H., Ong, J., Lee, J. C. B., Xu, D., Han, Z., Matthew, C. C. H., Ng, J. S. H. Y., Lim, C. M., Cheng, L. R. J., & Cayabyab, M. Y. (2021). Developing a perceived social media literacy scale: Evidence from Singapore. *International Journal of Communication*, *15*. https://ijoc.org/index.php/ijoc/article/view/16118

Ternovski, J., Kalla, J., & Aronow, P. (2022). Negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety*, *1*(2), Article 28. https://doi.org/10.54501/jots.v1i2.28

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1), 1–13. https://doi.org/10.1177/2056305120903408

Van Der Meer, T. G. L. A., Hameleers, M., & Ohme, J. (2023). Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies*, *24*(6), 803–823. https://doi.org/10.1080/1461670X.2023.2187652

Vogler, D., & Rauchfleisch, A. (2024). Deepfakes: Medienberichterstattung und Wahrnehmung in der Schweizer Bevölkerung. In fög – Forschungszentrum Öffentlichkeit und Gesellschaft / Universität Zürich (Ed.), *Jahrbuch Qualität der Medien* (pp. 63–72). Schwabe. https://doi.org/10.5167/uzh-264026

Wagner, T. L., & Blewer, A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of AI-altered video. *Open Information Science*, *3*(1), 32–46. https://doi.org/10.1515/opis-2019-0003

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence: The International Journal of Research into New Media Technologies*, *27*(1), 36–51. https://doi.org/10.1177/1354856520923963

Author Biographies

Daniel Vogler, PhD from the University of Zurich, serves as the Research Director of the Research Center for the Public Sphere and Society (fög) at the University of Zurich. Additionally, he is a Senior Research Associate at the Department of Communication and Media Research (IKMZ) at the same university. His research interests include crisis communication, public relations, journalism, online communication, and computational communication science.

Adrian Rauchfleisch, PhD from the University of Zurich, is an Associate Professor at the Graduate Institute of Journalism, National Taiwan University. His research focuses on the intersection of politics, technology, and journalism in Asia, Europe, and the United States. He is currently working on a project that examines the impact of Artificial Intelligence on society within various cultural contexts.

Gabriele de Seta, PhD from Hong Kong Polytechnic University, is a Researcher at the University of Bergen. He leads the ALGOFOLK project, which explores the relationship between algorithmic folklore and vernacular creativity. Using qualitative and ethnographic methods, his research delves into digital media practices, sociotechnical infrastructures, and vernacular creativity within the Chinese-speaking world.