Digital literacy interventions can boost humans in discerning deepfakes

– Preliminary analysis —

Dominique Geissler*[1,2], Claire Robertson[3,4], and Stefan Feuerriegel[1,2]

[1]LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning (MCML), Munich, Germany

[3]New York University, New York, NY, U.S.

[4]Rotman School of Management, University of Toronto, Toronto, Canada

*Corresponding author: d.geissler@lmu.de

Abstract

Deepfakes, i.e., images generated by artificial intelligence (AI), can erode trust in institutions and compromise election outcomes, as people often struggle to discern real images from deepfakes. Improving digital literacy can help address these challenges, yet scalable and effective approaches remain largely unexplored. Here, we compare the efficacy of five digital literacy interventions to boost people's ability to discern deepfakes: (1) textual guidance on common indicators of deepfakes; (2) visual demonstrations of these indicators; (3) a gamified exercise for identifying deepfakes; (4) implicit learning through repeated exposure and feedback; and (5) explanations of how deepfakes are generated with the help of AI. We conducted an experiment with $N = 1,200$ participants from the United States to test the immediate and long-term effectiveness of our interventions. Our results show that our interventions can boost deepfake discernment by up to 13 percentage points while maintaining trust in real images. Altogether, our approach is scalable, suitable for diverse populations, and highly effective for boosting deepfake detection while maintaining trust in truthful information.

1 Introduction

Artificial intelligence (AI) is increasingly used to generate highly realistic content across various domains, including entertainment, advertising, and education [1]. AI-generated images – commonly referred to as "deepfakes" – are not inherently harmful [2]; in fact, they offer a range of beneficial applications, such as animating sign language interpreters [3] or animating historical figures for documentaries. However, when used to spread disinformation, defined here as false or misleading content deliberately intended to deceive [4], deepfakes pose serious risks [5–9]. Notable examples of such disinformation that have previously gone viral include a fabricated image showing the arrest of U.S. President Donald Trump, which fooled millions of viewers online [10]; a deepfake of Pope Francis in a fashionable white puffer jacket [11]; and a deepfake of Ukraine's president Zelensky falsely announcing surrender to Russia [12]. Research shows that deepfakes are often highly persuasive [5, 13, 14], and many people struggle to discern real images from deepfakes [15–22]. Here, we thus propose multiple behavioral interventions aimed at boosting people's ability to discern between real images and deepfakes.

AI-generated disinformation, especially through increasingly realistic deepfakes, has several characteristics with serious societal implications. First, the production and dissemination of such content have become increasingly more accessible, as, unlike earlier tools such as Photoshop, modern AI image generators do not require advanced technical skills [4, 23]. Second, the realism of AI-generated content undermines trust in both interpersonal communication and public institutions, as individuals may come to believe that *any* image, video, or audio recording could be fabricated [24, 25]. Third, AI-generated disinformation can be strategically deployed to manipulate public opinion, influence electoral outcomes, or provoke social instability [23, 26–28]. Hence, there is a growing need for effective strategies to counter people's susceptibility to falling for deepfakes.

Countering AI-generated disinformation in the form of deepfakes is notoriously difficult. One

proposed solution is the use of automated detection tools, for example in the form of automated flagging of deepfakes on social media or through browser extensions. However, these tools, while widely researched, are often inaccurate [29–32] and lag behind the rapid advances in deepfake generators. Moreover, many deepfakes lack watermarks or sufficient metadata, which makes it more challenging for automated detection tools to flag them automatically [23]. Another challenge arises when automated flagging of deepfakes is not available. Then, even if automated detection tools were accurate, users would need to manually upload images for analysis, which is impractical and time-consuming and because of which the scalability is often limited [33]. In light of these limitations, strategies that operate independently of automated detection tools are scarce. To address this gap, we propose a set of digital literacy interventions aimed at boosting people's ability to discern deepfakes.

Digital literacy, i.e., the ability to analyze and evaluate online information, can help individuals in assessing the veracity of deepfakes. Recent studies show that digital literacy interventions can increase the resilience to disinformation [34–46], improve human discernment of disinformation [47–49], and reduce the spread of disinformation online [50]. However, nearly all of these interventions focus on textual misinformation. Few interventions have targeted the detection of deepfake images directly. In the context of deepfakes, early research suggests that explaining how an image was generated and why it is fake leads people to agree less with the false information presented in the image [51]. Other digital literacy interventions educate people on verification strategies, such as reverse image search, to help detect deepfakes [52, 53]. However, such verification strategies only work once multiple versions of an image are published online and are further limited by time lags. Moreover, verification strategies require substantial cognitive effort, which can lead to fatigue and reduced motivation to apply them consistently. As a result, verification strategies are generally not scalable. In contrast, we introduce digital literacy interventions that boost people's ability to discern deepfakes but without relying on external tools or sources, which makes our approach minimally disruptive, broadly applicable, and scalable.

Here, we propose, test, and compare five digital literacy interventions designed to boost people's ability to discern between real images and deepfakes. The interventions take the form of media literacy tips, which provide people with strategies to help them identify deepfakes more effectively (see [34] for a general overview of media literacy tips). Our interventions are as follows (Figure 1): (1) a textual description of typical errors found in deepfakes (*Textual*); (2) the same textual description but paired with an illustrative example image (*Visual*). (3) a gamified version in which users are asked to identify the errors in example deepfake images (*Gamified*); (4) a task involving 10 rounds of image discernment with feedback on each assessment (*Feedback*); and (5) an explanation of how deepfakes are generated (*Knowledge*). The first three interventions intentionally use a similar stimulus, i.e., presenting typical errors found in deepfakes, but where we compare different formats (i.e., text, visual, and gamified). This comparison is grounded in evidence that identical information can have different effects on behavior depending on how it is presented [37, 54]. It also allows us to test how the mode of presentation can influence learning, and, further, these formats vary in how they can be deployed across communications channels, such as private messages, social media posts, and digital ads. The fourth intervention instead draws upon the concept of implicit learning through repeated exposure [55], which further allows us to better understand the role of learning effects from repeated exposure effects. Finally, the fifth intervention is inspired by prior research [56] and tests whether a conceptual understanding of the underlying technologies of deepfake generation helps users generalize their detection skills to new deepfakes.

We evaluate the effectiveness of our five interventions in boosting individuals' ability to discern between real images and deepfakes through a between-subject experiment. For this study, we recruited $N = 1,200$ people from the United States (U.S.). Participants were randomly assigned to one of the intervention conditions or the control condition. Following the intervention, all participants completed an image discernment task, in which they are asked to classify a set of 15 images (real images and deepfakes) presented in a random order as either real or fake. We thus

collected a total of 37,980 image-level responses in our study. To assess the long-term effects of our interventions, we repeated the image discernment task with a different set of images after two weeks using a counter-balanced design. Figure 1 provides an overview of the experimental flow of our study. Our hypotheses are stated in the design table in Table 1.

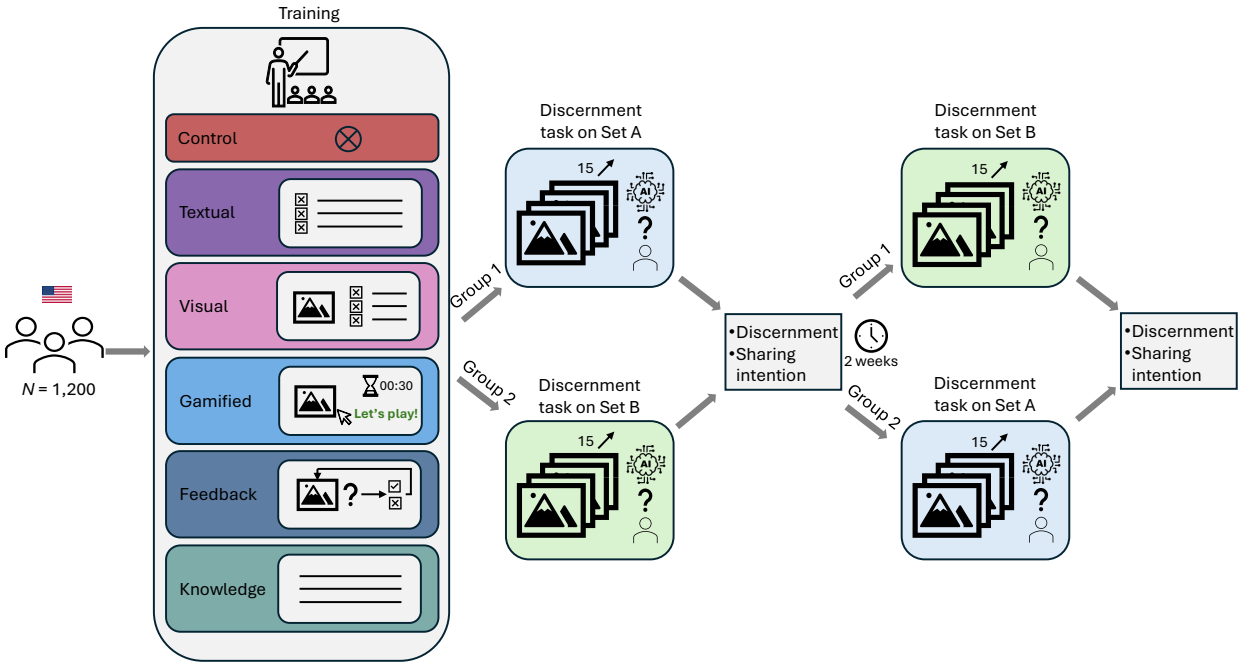


Figure 1: **Experimental flow.** We recruited participants from the U.S. for our online experiments. Participants were randomly assigned to either a control condition or one of five intervention conditions: (1) a textual description of ways to spot errors commonly found in deepfakes (*Textual*); (2) the same description with illustrative deepfakes (*Visual*); (3) a gamified exercise to spot errors in deepfakes (*Gamified*); (4) a discernment task with feedback across multiple rounds (*Feedback*); and (5) an explanation of how deepfakes are generated (*Knowledge*). Afterward, participants completed an image discernment task following a counter-balanced design (i.e., Group 1 sees image set A; Group 2 sees image set B). In the image discernment task, participants were asked to (a) classify 15 images as real or fake (using a four-point Likert scale) and (b) indicate whether they would share them on social media (with answer options "Yes", "No", "Don't know"). From these responses, we assess how the interventions affect the discernment, i.e., the accuracy in discerning real images and deepfakes, and the sharing intention for each participant. After two weeks, participants repeated the image discernment task with the alternate image set (i.e., Group 1 now sees image set B; and Group 2 sees image set A) to assess the long-term effects of the interventions.

Our digital literacy interventions offer several key benefits for addressing the growing threat

of deepfakes in everyday online settings. First, our interventions provide a *scalable* solution to counter the spread of disinformation by boosting people's ability to discern deepfakes. This is in contrast to automated detection tools or technical methods such as watermarking, which are often difficult to scale and can be easily circumvented by malicious actors. Second, our interventions aim to enhance users' ability to detect deepfakes but *without decreasing trust in real images*. This addresses a common limitation of some behavioral interventions, such as fact checking, which may unintentionally foster general skepticism and thus reduce trust in accurate content [57–59], To this end, we explicitly test whether the perceived veracity of real images is preserved (see H2 in Table 1). Third, our interventions are *minimally disruptive*. They support different formats and thus can be deployed through various communication channels such as private messages, social media posts, or digital ads. As such, they allow for repeated use to reinforce and sustain learning effects over time. In sum, boosting digital literacy skills can help individuals judge the veracity of online content and mitigate the societal risks posed by deepfakes, ultimately contributing to a more resilient society in the long term.

Hypothesis	**Sampling plan**	**Analysis plan**	**Interpretation given to different outcomes**
H1: People who receive a digital literacy intervention will have better discernment of deepfakes.	A power analysis suggested that a sample size of $N = 1,200$ participants will have sufficient power to achieve 80% power to detect an effect size of 0.25, which is considered a small effect [60].	We plan to conduct multiple Mann-Whitney U tests to test for significant differences in deepfake discernment (as measured by accuracy in the discernment task) between the control condition and the intervention conditions.	A significant test is interpreted as evidence that the respective intervention leads to a significantly better discernment ability.
H2: People who receive a digital literacy intervention will not be more skeptical of real images.	A power analysis suggested that a sample size of $N = 1,200$ participants will have sufficient power to achieve 80% power to detect an effect size of 0.25, which is considered a small effect [60].	We plan to conduct multiple equivalence tests [61] to evaluate whether the discernment of real images (as measured by accuracy in the discernment task) between the control condition and the intervention conditions is equivalent or not meaningfully different. We will test for equivalence against an interval of $(-0.05, 0.05)$.	If our observed confidence interval is fully contained in the testing interval, we will consider this as evidence for a null effect; otherwise, we will consider the results inconclusive with respect to the null.
H3: People who receive a digital literacy intervention will have better long-term discernment of deepfakes.	To account for possible attrition [62], our sample size of $N = 1,200$ participants includes a buffer of 30% participants to achieve 80% power to detect an effect size of 0.25, which is considered a small effect [60].	We plan to conduct multiple Mann-Whitney U tests to test for significant differences in deepfake discernment (as measured by accuracy in the discernment task) between the control condition and the intervention conditions in the follow-up.	A significant test is interpreted as evidence that the respective intervention leads to significantly better discernment in the follow-up after two weeks.

Table 1: **Design table.** A significance level of $\alpha = 0.05$ will be used for all statistical comparisons.

2 Methods

2.1 Ethics information

This research complies with all relevant ethical regulations. Ethical approval was obtained from the Institutional Review Board (IRB) of the Faculty of Mathematics, Informatics, and Statistics at LMU Munich (EK-MIS-2024-319). All participants gave informed consent at the beginning of the experiment. At the end of the first experiment, participants were debriefed that some of the images they viewed were deepfakes. At the end of the follow-up, they were informed which specific images were artificially generated.

2.2 Overview

We employed a between-subjects experiment with a counter-balanced design to analyze the effect of digital literacy interventions in boosting individuals' ability to discern between real images and deepfakes. Participants were randomly assigned to either the control condition or one of five intervention conditions: (1) *Textual*, (2) *Visual*, (3) *Gamified*, (4) *Feedback*, or (5) *Knowledge*, as described below. Participants in the intervention groups received the corresponding training, while those in the control group received no training. Following this, all participants were given a discernment task in which they classified images as real or fake. After a period of two weeks, we conducted a follow-up where the image discernment task was repeated to analyze the long-term effects of our interventions. We conducted the experiment with a convenience sample of people from the U.S.

2.3 Sampling plan

We recruited a convenience sample of $N = 1,200$ participants from the U.S. via the online platform Prolific (prolific.com). Participants can take part in the experiment if they reside in the U.S.

and are over 18 years old. Participants received compensation equivalent to 12$/hour, paid through Prolific. We collected the unique IDs of the participants on the recruitment platform to invite them to the follow-up conducted two weeks later. Data was collected through an online survey hosted on Qualtrics.

To determine the sample size, we performed a power analysis for a one-tailed t-test for equally allocated groups. For the effect size, we find that, on average, the effect size of media literacy training studies is $d = 0.37$ [63], which would be a medium effect according to common interpretations of Cohen's d [60]. We choose an α error probability of 0.05 and aim for a power of 0.8 [62]. This gives a minimum sample size of $N = 96$ per condition for equally allocated groups. Following best practices to account for potential attrition over the two-week interval [62], we increased the minimum sample size by 30% to $N = 125$ participants per condition. Given this minimum and accounting for data filtering, we recruited $N = 200$ participants per condition, which is in line with previous work on literacy interventions [63]. This gives a final sample size of $N = 1,200$. With this sample size, we are able to detect effect sizes above $d = 0.25$ with a power of 0.8 according to a post-hoc sensitivity analysis.

2.4 Interventions

We propose five different digital literacy interventions aimed at boosting people's ability to discern between real images and deepfakes, as detailed in the following.

(1) *Textual*: We presented participants with short, textual descriptions of typical errors found in deepfakes (see Table 2 for the content of the intervention). Prior research has categorized such errors found in deepfakes into five different types: anatomical implausibilities (e.g., distorted hands or unlikely alignment of teeth), stylistic artifacts (e.g., unnaturally smooth skin or hyper-real details), functional implausibilities (e.g., dysfunctional objects or incomprehensible text), violations of physics (e.g., missing shadows of objects and people or inconsistent reflections), and sociocultural implausibilities (e.g., clothing or symbols that are culturally out of place) [64]. We teach

these to participants using so-called media literacy tips, i.e., tips that give people a list of strategies for identifying deepfakes [34]. We selected a textual format for this intervention as it can be easily deployed across various channels, such as private messages, social media posts, or news articles. By presenting only text without images, the intervention may further avoid biasing participants toward specific visual cues and instead foster a more general understanding of how to detect the characteristics of deepfakes. To improve attention, each error was presented on a separate page, where the order was randomized to mitigate potential order effects.

***Hint**: Fake images often have anatomical errors. For example, people often have missing, extra, or merged fingers, nonexistent fingernails, and unlikely hand proportions.*
***Hint**: Fake images often look a bit too perfect. People may have glossy, shiny skin, windswept hair, and they may look like a photo in a magazine or a scene from a movie. Sometimes, parts of an image have a different level of detail or vibrancy compared to the rest.*
***Hint**: Fake images often have functional errors. Image generators do not have a structural understanding of how objects in the world work and interact with each other. This can cause composition errors, dysfunctional objects, and atypical designs. Looking closely may also reveal distorted and unresolved details or incomprehensible text and logos.*
***Hint**: Fake images can produce subtle artifacts that are inconsistent with the laws of physics. These include inaccurate shadows, reflections of alternative realities, and depth and perspective issues.*
***Hint**: Fake images may have sociocultural errors. For example, the images may show scenarios that are inappropriate, unlikely to be seen in the real world, violate subtle rules specific to different cultures, or are historically inaccurate.*

Table 2: The intervention *Textual* presents common errors found in deepfakes through short, textual descriptions. Each hint explains one specific error.

(2) *Visual*: We presented participants with the same textual descriptions of common errors found in deepfakes as in the *Textual* intervention, but each description was accompanied by an illustrative deepfake image that exemplifies the respective error. For the sociocultural implausibilities, we used the deepfake image of Pope Francis in a puffer jacket. For the remaining errors, we generated example images using DALL-E 3 [65] via the ChatGPT website to show examples of the errors (see Supplementary Table S16 for the prompts and sources). Each image was supplemented with a short explanatory sentence that guides participants to the relevant error in the image (see Table 3 for the exact intervention content). We chose to show the textual information together with an example image to make the error more tangible and easier to recognize. Compared to the *Textual* intervention, the visual format helps illustrate how these errors manifest in practice, which may

support learning. This format is also well-suited for deployment in visually-driven environments, such as social media posts or digital ads. As in the previous intervention, each error was shown on a separate page in randomized order.

Image	Error description
	***Hint**: Fake images often have anatomical errors. For example, people often have missing, extra, or merged fingers, nonexistent fingernails, and unlikely hand proportions.* *In the image below, you can see that the hands are merged together and that the hand on the right should have been a left hand, so the thumb should not be visible.*
	***Hint**: Fake images images often look a bit too perfect. People may have glossy, shiny skin, windswept hair, and they may look like a photo in a magazine or a scene from a movie. Sometimes, parts of an image have a different level of detail or vibrancy compared to the rest.* *In the image below, you can see that the face of the woman is not realistic.*
	***Hint**: Fake images often have functional errors. Image generators do not have a structural understanding of how objects in the world work and interact with each other. This can cause composition errors, dysfunctional objects, and atypical designs. Looking closely may also reveal distorted and unresolved details or incomprehensible text and logos.* *In the image below, the saucer seems to be floating underneath the cup on the left with no hand holding it.*
	***Hint**: Fake images can produce subtle artifacts that are inconsistent with the laws of physics. These include inaccurate shadows, reflections of alternative realities, and depth and perspective issues.* *In the image below, the reflection of the man doesn't match the man in front of the mirror. The wrong arm is holding the toothbrush and the toothbrush should be facing a different direction.*
	***Hint**: Fake images may have sociocultural errors. For example, the images may show scenarios that are inappropriate, unlikely to be seen in the real world, violate subtle rules specific to different cultures, or are historically inaccurate.* *In the image below, the scenario is highly unlikely.*

Table 3: The *Visual* intervention shows common errors found in deepfakes through a combination of an example image and an explanatory text.

(3) *Gamified*: This intervention draws upon elements of gamification, which are known to

increase engagement, motivation, and learning [66]. Here, we first presented participants with brief instructions on how to play the game and showed them the same example images and textual descriptions of how to recognize deepfakes as in the *Visual* intervention. Then, the participants played a game in which the goal was to locate the error in each deepfake image by clicking on the corresponding position in the image. We employ various elements that are common in gamification [66], namely, a timer, rewards in the form of points, and real-time feedback. Participants had 30 seconds per image and were rewarded 10 points for each error they found. We showed their overall score throughout the game. To further motivate participants, participants received immediate feedback when clicking on an image: participants that identify the error correctly saw the error highlighted with a green circle and received a message in green font (*"Correct! You found the error. You earned 10 points!"*); for all other clicks, the participants received a message in red font (*"Not quite! Try again."*). After finishing the game, participants saw their total score and a message that ranks their achievement (0–10 points: *"Thanks for participating! Spotting errors in fake images can be challenging. With more exposure, you'll likely get better at it."*; 20–30 points: *"Good eye! You've shown a solid ability to spot errors in fake images."*; 40–50 points: *"Impressive results! You've demonstrated a keen ability to spot errors in fake images."*). An example is shown in Table 4. As in the other interventions, each error and image was presented on a separate page in randomized order.

Points: 10 Time left: 20 seconds 	***Hint**: Fake images often have anatomical errors. For example, people often have missing, extra, or merged fingers, nonexistent fingernails, and unlikely hand proportions.* *Can you spot the error in the image? Click on it to earn points!* ***Not quite! Try again.***
Points: 10 Time left: 20 seconds 	***Hint**: Fake images often have anatomical errors. For example, people often have missing, extra, or merged fingers, nonexistent fingernails, and unlikely hand proportions.* *Can you spot the error in the image? Click on it to earn points!* ***Correct! You found the error. You earned 10 points!***

Table 4: The *Gamified* intervention shows participants examples of deepfake images and explanatory texts in a gamified setting, which includes a timer, point-based rewards, and real-time feedback.

(4) *Feedback*: We presented participants with five real images and five deepfakes and asked them to complete ten rounds of image discernment with immediate feedback after each assessment (see Table 5 for the images). The deepfakes in this intervention are analogous to those in the *Visual* and *Gamified* intervention. Unlike the other interventions, no descriptions of errors are provided; this choice is informed by prior research suggesting that repeated exposure alone can support generalization to new stimuli [67]. For each image, we asked participants whether they think the image is real or fake. After answering, they received immediate feedback indicating whether their answer was correct. If the participant misidentifies an image (e.g., by labeling a deepfake as real), a red-colored text box appears stating that their response is incorrect and that the image is fake. In addition, the answer button briefly "shakes" when an incorrect answer is given to reinforce the feedback. If the answer is correct, a green-colored text informs that the response was correct. Each image was again shown on a separate page in randomized order.

Real images				
				
Deepfakes				
				

Table 5: The *Feedback* intervention presents five real images and five deepfakes to the participants to discern, but without explanation. Instead, participants receive immediate feedback indicating whether their assessment of each image is correct.

(5) *Knowledge*: We presented participants with a textual explanation of how AI technologies are used in image generators to create deepfakes. The explanation is written in simple language and presented as a brief paragraph (see Table 6 for the exact content). This intervention tests whether increased conceptual knowledge about AI image generation influences participants' ability to discern deepfakes. This is inspired by prior research which suggests that familiarity with a topic can affect discernment performance [56].

AI can learn how to generate realistic images by looking at millions of example pictures. After inputting the example images into the AI, they are transformed into representations that only computers can understand and that don't mean anything to humans. From those, AI then learns patterns of how the world, such as objects, people, and places look like. After training, humans can ask AI models to generate images through simple input text, this is called prompting. The possibilities for AI-generated images are endless, for example, AI can generate things that have never been seen before such as made-up scenarios or even attempt to imitate real people or places.

Table 6: The *Knowledge* intervention provides a short explanations of how deepfakes are generated with the help of AI.

2.5 Task and materials

Participants completed an image discernment task in which they are shown a set of 15 images, presented one at a time in randomized order. The 15 images include five images that are real and 10 images that are deepfakes. For each image, we asked the participants (a) whether they think the image is real (i.e., the image is shot by a human) or fake (i.e., AI-generated) using a four-point Likert scale (i.e., "Definitely fake", "Probably fake", "Probably real", "Definitely real"); and (2) whether they would consider sharing the image on their social media ("Yes", "No", "Don't know"). We use the simplified terms "real" and "fake" as they are more accessible to a general audience. A screenshot of the task interface is provided in the Supplementary Materials B.

Additionally, participants were given the option to tick a checkbox if they have seen the image before or if the image failed to load properly (e.g., due to bandwidth issues with their Internet). All responses for images where a participant checked one of the options (i.e., reported they had seen the image or that the image had not loaded properly) were excluded from the analysis. Later, we provide a robustness check with regard to people who have seen the images in Supplementary Section E.2.

We presented participants with both real images and deepfakes that cover a broad range of themes, including everyday activities, potential disinformation scenarios, and portraits. In total, we collected 10 real images and 20 deepfakes, which we randomly split into two sets (Group A and B), each containing 15 images (see Supplementary Section G for an overview of all images, including

the source links and the collection date). Due to our counter-balanced design, participants were randomly assigned to two groups: one group received image set A in the initial session and set B in the 14-day follow-up, while the other group received the sets in reverse order, which allows us to rule out ordering effects when assessing the long-term effect. Later, we provide a robustness check where we test for differences in discernment ability taking into account the image set that participants were assigned in Supplementary Section E.10.

To account for the potential heterogeneity of deepfakes, we included images from two sources: half of the images have gone viral and the other half have not. We thereby aim to account for that viral deepfakes may possess certain features that contribute to their widespread dissemination (e.g., such as being more convincing, more emotionally engaging, or crafted with greater technical sophistication). Including both types allows us to examine a broader and more representative spectrum of deepfake content. The viral deepfakes were sourced from global news outlets reporting on widely shared deepfakes. The non-viral deepfakes were collected from Midjourney's Discord channel [68], self-generated using DALL-E 3 [65] as provided by the ChatGPT website, or sourced from the Center for Countering Digital Hate [27]. Real images were selected to match the themes of the deepfakes and were obtained from the public photo hosting service Flickr (www.flickr.com). We provide a robustness check in Supplementary Section E.11, where we test for differences in discernment abilities for viral and non-viral deepfakes.

2.6 Procedure

Participants first provided informed consent and completed an attention check. They were then randomly assigned to either the control condition or one of five intervention conditions. Only participants in the intervention conditions received a digital literacy intervention. Further, participants were randomly assigned to one of two groups as part of a counter-balanced design. Group 1 completed image set A immediately after the intervention and image set B in the two-week follow-up; Group 2 received the sets in reverse order. This design controls for potential order and image-

specific effects, ensuring that differences in the discernment task are attributable to the interventions rather than the sequence or content of the image sets.

Next, we collected covariates such as the participants' sociodemographics, social media use, and digital literacy levels. Participants were also asked to complete a cognitive reflection test (CRT) [69] (see Supplementary Table S2 for details on questions and scales). At the end of the study, we also performed a second attention check and two honesty checks (i.e., asking whether participants responded randomly at any point or whether they searched for any of the images online such as, e.g., via Google). Participants who failed both of the attention checks and/or both honesty checks were excluded from the analysis.

After two weeks, we invited participants to complete a follow-up. They were asked to repeat the image discernment task but with the previously unseen set of 15 images (i.e., set A or B, depending on their group assignment). Then, participants answered the same set of questions about their digital literacy abilities. The follow-up remained open for one week, after which it was closed for submissions. Finally, participants were debriefed and informed on which images were real and which were deepfakes.

2.7 Measures

We collected the following dependent variables from our image discernment task, namely, (a) *Discernment*, and (b) *SharingIntention* as follows. After finishing the task, we also asked participants once for their overall (c) *Confidence* during the discernment task.

(a) *Discernment*: We asked participants to rate 15 images as real or fake on a four-point Likert scale ("Definitely fake", "Probably fake", "Probably real", "Definitely real"). Each response was coded as correct or incorrect based on the image's actual label, based on which we then computed two discernment measures for each participant, namely, their accuracy in detecting real images and in detecting deepfakes. The accuracy for real images is calculated as TP/P, where TP denotes the true positives and P the total number of real images. The accuracy for

deepfakes is calculated as TN/N, where TN denotes true negatives and N the total number of deepfakes.

(b) *SharingIntention*: For each image, we asked participants whether they would share it on their social media (answer options: "Yes", "No", "Don't know"). A "Yes" is counted as an intention to share. Again, we computed two measures for each participant: The sharing intention for real images, computed as TP/P, where TP denotes the real images the participant is willing to share and P the total number of real images; and the sharing intention for deepfakes, calculated as TN/N, where TN denotes the deepfakes the participant is willing to share and N the total number of deepfakes.

(c) *Confidence*: After completing the image discernment task, we asked participants about their confidence in identifying images as real and fake on a scale of 0 (not confident at all) to 100 (very confident).

Details are in Supplementary Table S1.

To account for between-participant heterogeneity in our analyses, we further collected four sets of covariates. (1) We asked participants about their sociodemographics such as age, ethnicity, gender, level of education, religion, political orientation, income, and subjective social status. This is motivated by prior research showing that, for example, age is a strong predictor of susceptibility to misinformation, with older adults particularly struggling to distinguish real from fake content [70, 71], and that tailored interventions can help improve their discernment abilities [48]. (2) We asked participants about their social media use, including the platforms they use, their sharing habits, and the amount of time they spend online. This helps capture the media environment participants are typically exposed to, which may influence their familiarity with manipulated content. (3) We collected information about the participants' digital literacy, such as their knowledge of deepfakes, their experience in detecting deepfakes, as well as their experience with search engines, reverse image search, and generative AI. (4) We collected information on the participants'

analytical thinking skills through the use of CRT [69]. This is informed by previous studies showing that psychological factors, such as partisanship [72] and analytical thinking [36, 73, 74], are important predictors of digital literacy. Details on all are provided in Supplementary Table S2.

2.8 Analysis plan

First, we filtered the collected data as follows. Participants who failed both attention checks and/or both honesty checks were excluded from the analysis. We also removed participants whose completion time fell outside the 95% interval of the sample distribution. In addition, we removed participants who gave the same answer for every image in the discernment task. Lastly, if a participant indicated that they had previously seen the image or experienced loading issues, we set these responses to missing values and computed the discernment and sharing intention without those responses.

We conducted multiple descriptive as well as exploratory analyses. To evaluate the effectiveness of our digital literacy interventions on deepfake discernment, we compared each intervention condition to the control group using Mann-Whitney U tests. We chose to use the Mann-Whitney U test as a non-parametric test to account for potential non-normality in our data (for details, see the detailed normality check in Supplementary Section C) . To assess whether our intervention made participants unintentionally more skeptical of real images, we conducted equivalence tests where we compared each intervention condition against the control condition. To analyze the long-term effect of our interventions on deepfake discernment, we again use Mann-Whitney U tests to assess a difference in means.

All analyses are conducted using Python (3.8) with *numpy* (1.24.4) , *pandas* (2.0.3) , *scipy* (1.10.1) , and *statsmodels* (0.14.1) . Data visualizations were created with *matplotlib* (3.7.5) , and *seaborn* (0.13.2) . All analysis scripts will be made publicly available.

2.9 Robustness checks

We conducted a series of robustness checks to assess the validity and reliability of our findings. First, we repeated our analyses without excluding participants who failed both attention checks and/or honesty checks, to examine whether our effects remain robust. Second, we repeated our analysis without removing image-level responses flagged by participants as previously seen or not properly loaded. For both, we compared the difference in discernment ability for filtered and non-filtered responses using Mann-Whitney U tests.

Third, to account for individual-level heterogeneity, we estimated the effect of our interventions on participants' ability to discern deepfakes using an ordinary least squares (OLS) regression model via

$$Y_i = \beta_0 + \beta_1 \textit{Condition}_i + \beta_2 X_i + \varepsilon_i, \tag{1}$$

where Y_i is the observed discernment ability for participant i, $\textit{Condition}_i$ is a binary variable equal to 1 for intervention and 0 for control, and X_i is a vector of participant-level control variables.

Fourth, we accounted for possible interaction effects with participant-specific control variables in our regression model via

$$Y_i = \beta_0 + \beta_1 \textit{Condition}_i + \beta_2 X_i + \beta_3 (\textit{Condition}_i \odot X_i) + \varepsilon_i, \tag{2}$$

where $\textit{Condition}_i \odot X_i$ is a two-way interaction term.

Fifth, to test for non-random attrition, we compared participants who completed the follow-up with those who dropped out using Mann-Whitney U tests for differences in discernment abilities and sociodemographics.

Sixth, we controlled for the image set that participants saw in the image discernment task via

$$Y_i = \beta_0 + \beta_1 \textit{Condition}_i + \beta_2 Set_i + \beta_3 (\textit{Condition}_i \odot Set_i) + \varepsilon_i, \tag{3}$$

where Set_i is a binary variable which equals 0 if participant i saw image set A in the first session and 1 if participant i saw image set B, and where $\odot$ denotes element-wise multiplication.

Seventh, we assessed whether the interventions are differentially effective for viral versus non-viral deepfakes using a linear mixed-effects model

$$Y_{ij} = \beta_0 + \beta_1 \textit{Condition}_i + \beta_2 \textit{Source}_j + \beta_3(\textit{Condition}_i \odot \textit{Source}_j) + u_i + \varepsilon_{ij}, \tag{4}$$

where Y_{ij} is the observed discernment ability for participant i for deepfakes set j, $\textit{Condition}_i$ is a categorical variable indicating the intervention condition for participant i (with control as reference category), $\textit{Source}_j$ is a binary variable which equals 1 for deepfakes sets consisting of viral images and 0 for non-viral deepfakes, $\textit{Condition}_i \odot \textit{Source}_j$ is the interaction term, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ are random intercepts accounting for individual differences, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual error.

Eight, we conducted a robustness check where we evaluated the long-term effectiveness of our interventions using a linear mixed effects regression with participant-level random effects. Here, we estimate

$$Y_{it} = \beta_0 + \beta_1 \textit{Condition}_i + \beta_2 \textit{TimePoint}_t + \beta_3(\textit{Condition}_i \odot \textit{TimePoint}_t) + u_i + \varepsilon_{it}, \tag{5}$$

where Y_{it} is the deepfake discernment ability of participant i at time point t, $\textit{Condition}_i$ is a binary variable indicating the intervention condition for participant i (with control as reference category), $\textit{TimePoint}_t$ is a binary variable which equals 0 for the first session and 1 for the follow-up time, $\textit{Condition}_i \odot \textit{TimePoint}_t$ is the interaction term, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ are random intercepts for individual differences, and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual error.

Lastly, we conducted a robustness check to test for possible learning effects from the image discernment task itself (see Supplements F for more information). We found that mere exposure to the discernment task did not create a learning effect for the control condition, suggesting that improvements in discernment performance cannot be explained by exposure to the task but must

be attributed to our interventions.

3 Results

We conducted our experiment with a sample of $N = 1,200$ U.S. participants. Applying the above filtering resulted in a final sample $N = 1,112$ participants. People were filtered roughly equally from each condition, with 187 people remaining in the *Control* condition, 188 people in the *Textual* condition, 183 in the *Visual* condition, 184 in the *Gamified* condition, 185 people in the *Feedback* condition, and 185 people in the *Knowledge* condition. The response rate for the follow-up was: 72.19% for *Control* condition (135 out of 187 participants), 76.60% for *Textual* (144 out of 188 participants), 70.49% for *Visual* (129 out of 183 participants), 69.57% for *Gamified* (128 out of 184 participants), 76.76% for *Feedback* (142 out of 185 participants), and 71.51% for *Knowledge* (133 out of 185 participants). Hence, the number of participants in the follow-up was relatively balanced across conditions. In the following, we present our results. A significance level of $\alpha = 0.05$ was used for all statistical comparisons. The sociodemographics of our participants can be found in Supplementary Section D, and the results of our robustness checks are in Supplementary Section E.

3.1 Effects of interventions on discernment abilities

We analyze the effect of our digital literacy interventions on participants' abilities to discern between real images and deepfakes compared to the control condition. We first assess in Section 3.1.1 whether our interventions boosted participants' ability to correctly identify deepfakes, both immediately after the intervention and at follow-up (Hypothesis H1 and H3, respectively; see Design Table 1). In Section 3.1.2, we analyze whether our interventions affected participants' trust in real images and made them more skeptical (Hypothesis H2 in Design Table 1).

3.1.1 Discerning deepfakes

In our main session (Hypothesis H1; see Design Table 1), participants identified deepfakes with a mean accuracy of 61.3% (see Figure 2**a**). Deepfake discernment was boosted by 7.5 percentage points for participants in the *Textual* condition ($\mu = 68.8\%$, Mann-Whitney $U = 14302.0$, $p = 0.001$), and by 13 percentage points for participants in the *Visual* intervention ($\mu = 74.3\%$, Mann-Whitney $U = 11665.0$, $p < 0.001$). Discernment for deepfakes also improved for interventions *Gamified* ($+4.4$ percentage points, $\mu = 65.7\%$, Mann-Whitney $U = 15292.5$, $p = 0.062$) and *Knowledge* ($+3.2$ percentage points, $\mu = 64.5\%$, Mann-Whitney $U = 16061.0$, $p = 0.199$), although not significantly. Finally, participants in the *Feedback* condition showed no improvement over the control condition ($\mu = 60.0\%$, Mann-Whitney $U = 17872.5$, $p = 0.577$).

To analyze long-term effects (Hypothesis H3; see Design Table 1), we conducted a follow-up two weeks after the initial intervention (see Figure 2**b**). We compared each intervention against the control condition using Mann-Whitney U tests but found no significant difference in deepfakes discernment (*Textual*: Mann-Whitney $U = 8635.5$, $p = 0.105$; *Visual*: $U = 7724.0$, $p = 0.110$; *Gamified*; $U = 9132.5$, $p = 0.421$; *Feedback*: $U = 9654.5$, $p = 0.917$; and *Knowledge*: $U = 8915.5$, $p = 0.922$). The accuracy in detecting deepfakes in the control group was comparable and non-significantly different for the main session and the follow-up (Wilcoxon signed rank test

$W = 2816.0$, $p = 0.244$).

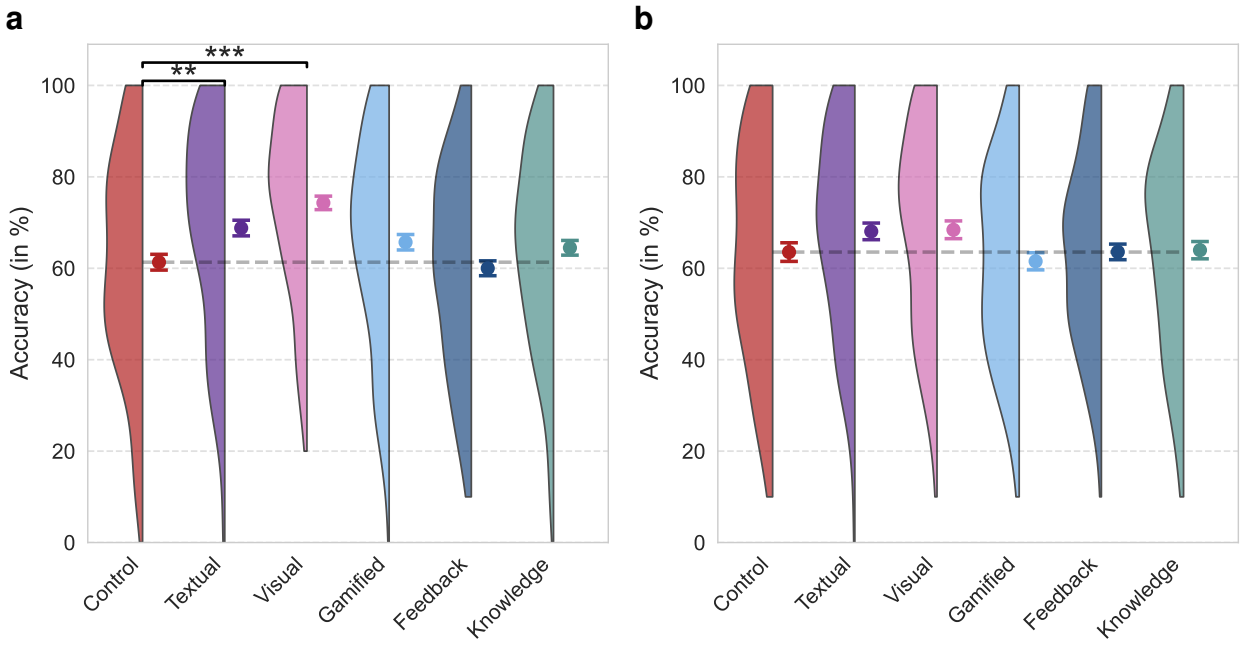


Figure 2: **Discernment accuracy for deepfakes across conditions.** Shown is the participant-level accuracy in discerning deepfakes: (**a**) immediately after the intervention (time point T1) and (**b**) two weeks after the initial intervention (time point T2). The sample size was $N = 1,112$ for time point T1 and $N = 764$ for time point T2 after attrition. We hypothesized that participants who received a digital literacy interventions would show a better accuracy in discerning deepfakes compared to the control group. Evidently, at T1, the discernment accuracy was significantly higher for the *Textual* and *Visual* interventions. Each violin plot shows the distribution of the participant-level accuracy within each condition. Dots indicate mean values; whiskers indicate standard errors. Significance levels for Mann-Whitney U tests: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

3.1.2 Discerning real images

Next, we examine whether our interventions had any unintended effects on participants' ability to detect real images (Hypothesis H2; see Design Table 1). On average, participants correctly identified real images with a mean accuracy of 83.9% (see Figure 3**a**). We performed an equivalence test [61] to see whether the difference between the control condition and the intervention conditions can be dismissed as null. The equivalence test involves defining a threshold for the smallest meaningful effect (here: -0.05 to $+0.05$) and determining whether the effect of interest falls within that threshold. We hypothesized that our intervention did not make participants more skeptical of real images. We find that the equivalence tests for each intervention fall in undecided range (*Textual*: $p = 0.669$, 95% CI: $[-3.192, 5.149]$; *Visual*: $p = 0.958$, 95% CI: $[-0.453, 8.404]$; *Gamified*: $p = 0.875$, 95% CI: $[-1.749, 6.964]$; *Feedback*: $p = 0.547$, 95% CI: $[-4.371, 3.781]$; *Knowledge*: $p = 0.496$, 95% CI: $[-3.986, 3.918]$). This suggests that the results are statistically inconclusive with respect to a null effect.

Similarly, we compare the discernment of real images between the control condition and the intervention conditions in the follow-up (see Figure 3**b**). For this, we conducted equivalence tests [61] with threshold $[-0.05, +0.05]$ for the smallest meaningful effect. We again find that the equivalence tests were undecided and hence inconclusive with respect to the null effect (*Textual*: $p = 0.604$, 95% CI: $[-4.275, 5.720]$; *Visual*: $p = 0.709$, 95% CI: $[-3.702, 6.718]$; *Gamified*: $p = 0.626$, 95% CI: $[-6.196, 4.365]$; *Feedback*: $p = 0.900$, 95% CI: $[-7.983, 1.605]$; *Knowledge*: $p = 0.546$, 95% CI: $[-4.757, 5.457]$). Overall, the results indicate that none of the interventions decreased participant's ability to detect real images, neither immediately after the intervention nor at the follow-up.

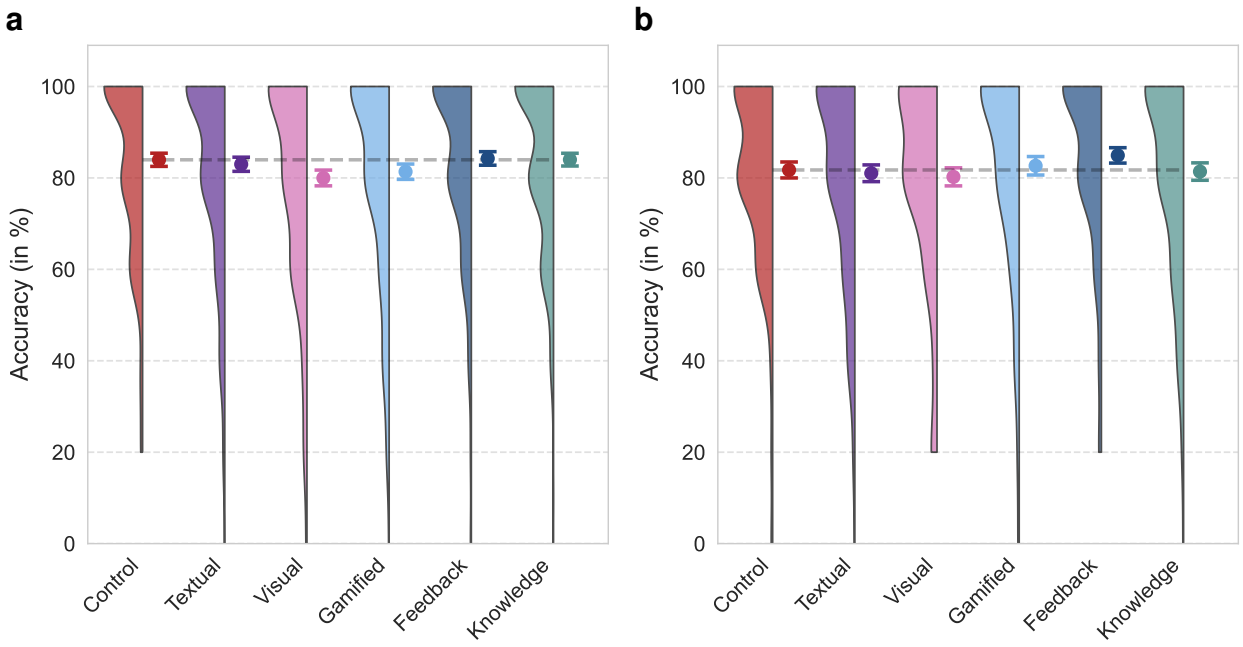


Figure 3: **Discernment accuracy for real images across conditions.** Shown is the participant-level accuracy in discerning real images: (**a**) immediately after the intervention (time point T1) and (**b**) two weeks after the initial intervention (time point T2). The sample size was $N = 1,112$ for time point T1 and $N = 764$ for time point T2 after attrition. We hypothesized that participants who receive a digital literacy interventions would not become more skeptical of real images. Across both time points, the accuracy remained statistically indistinguishable from the control group. Each violin plot shows the distribution of the participant-level accuracy within each condition. Dots indicate mean values; whiskers indicate standard errors.

3.2 Effect of interventions on sharing intention

Next, we examine the effect of our interventions on the sharing intentions for real images and deepfakes.

3.2.1 Sharing deepfakes

Participants were willing to share deepfakes at an average rate of 15.58% (Figure 4**a**). The *Visual* intervention significantly reduced the sharing intention of participants by 5.18 percentage points ($\mu = 10.4\%$, Mann-Whitney $U = 19224.5$, $p = 0.023$). The other interventions had no significant effect on the sharing intention for deepfakes (*Textual*: $\mu = 13.5\%$, Mann-Whitney $U = 18679.0$, $p = 0.252$; *Gamified*: $\mu = 14.0\%$, $U = 18404.0$, $p = 0.204$; *Feedback*: $\mu = 16.8\%$, $U = 18138.0$, $p = 0.377$; and *Knowledge*: $\mu = 16.1\%$, $U = 18322.0$, $p = 0.329$).

In the two-week follow-up, the sharing intention of participants for deepfakes were not different from the control condition (*Textual*: Mann-Whitney $U = 9991.5$, $p = 0.654$; *Visual*: $U = 9759.5$, $p = 0.054$; *Gamified*: $U = 8589.5$, $p = 0.928$; *Feedback*: $U = 9816.0$, $p = 0.700$; and *Knowledge*: $U = 9314.0$, $p = 0.555$; see Figure 4**b**).

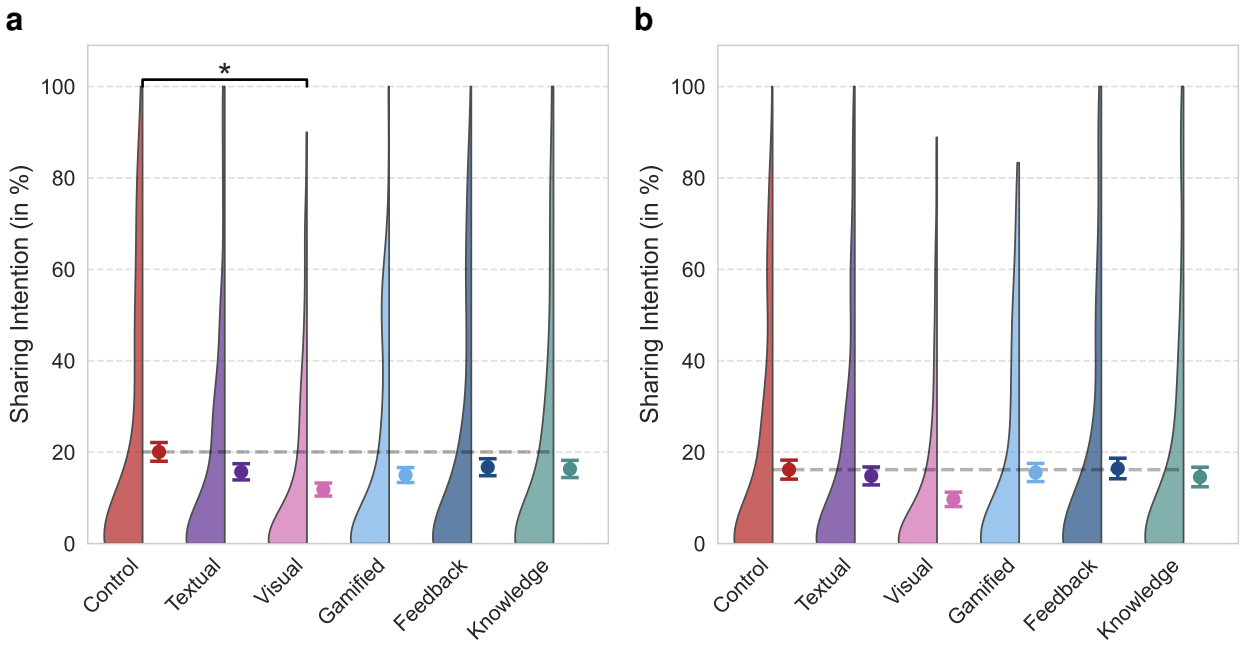


Figure 4: **Sharing intention for deepfakes across conditions.** Shown is the participant-level sharing rate of deepfakes: (**a**) immediately after the intervention (time point T1) and (**b**) two weeks after the initial intervention (time point T2). The sample size was $N = 1,112$ for time point T1 and $N = 764$ for time point T2 after attrition. Each violin plot shows the distribution of the participant-level sharing rate within each condition. Dots indicate mean values; whiskers indicate standard errors. Significance levels for Mann-Whitney U tets: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

3.2.2 Sharing real images

Participants were willing to share real images with a mean rate of 21.27% (see Figure 5**a**). As expected, participants in the control condition were, on average, less willing to share deepfakes than real image. Since our interventions aim not to increase skepticism toward real images, we tested for equivalence in the sharing intention for real images between the control condition and each of our intervention conditions using a threshold of $[-0.05, +0.05]$. All equivalence tests for each intervention were undecided (***Textual***: $p = 0.658$, 95% CI: $[-8.671, 5.618]$; ***Visual***: $p = 0.872$, 95% CI: $[-2.839, 10.849]$; ***Gamified***: $p = 0.859$, 95% CI: $[-3.087, 10.779]$; ***Feedback***: $p = 0.853$, 95% CI: $[-3.166, 10.668]$; and ***Knowledge***: $p = 0.884$, 95% CI: $[-2.660, 11.189]$). This suggests that our results are inconclusive with respect to the null effect.

We repeated this analysis for the two-week follow-up. Again, all equivalence tests were inconclusive, suggesting no clear evidence for or against a null effect (***Textual***: $p = 0.553$, 95% CI: $[-8.236, 7.087]$; ***Visual***: $p = 0.752$, 95% CI: $[-4.779, 9.992]$; ***Gamified***: $p = 0.542$, 95% CI: $[-7.215, 8.135]$; ***Feedback***: $p = 0.736$, 95% CI: $[-10.454, 5.305]$; and ***Knowledge***: $p = 0.522$, 95% CI: $[-7.387, 7.919]$; see Figure 5**b**).

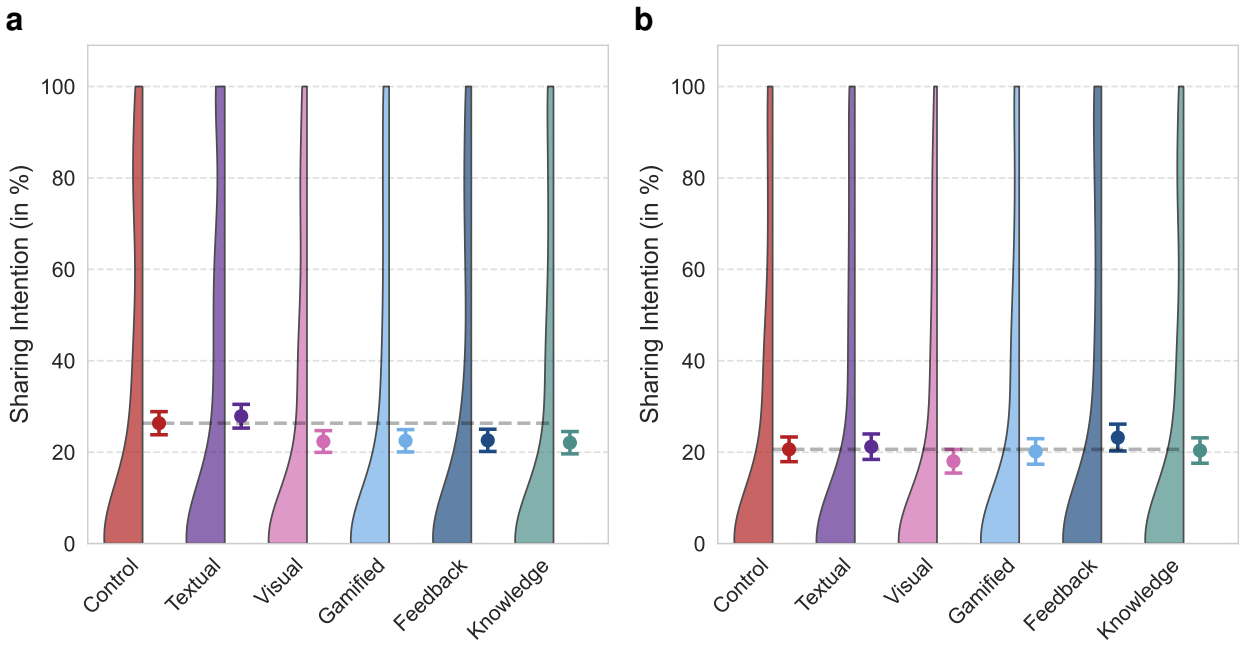


Figure 5: **Sharing intention for real images across conditions.** Shown is the participant-level sharing rate of real images: (**a**) immediately after the intervention (time point T1) and (**b**) two weeks after the initial intervention (time point T2). The sample size was $N = 1,112$ for time point T1 and $N = 764$ for time point T2 after attrition. Each violin plot shows the distribution of the participant-level accuracy within each condition. Dots indicate mean values; whiskers indicate standard errors.

4 Discussion

We conducted an experiment with $N = 1,200$ participants and found that the *Textual* and *Visual* interventions significantly boosted deepfake discernment by 7.5 percentage points and 13 percentage points, respectively (both $p < 0.001$). At the same time, both interventions did not increase participants' skepticism toward real images. In the two-week follow-up, participants in the *Textual* and *Visual* conditions continued to perform better than those in the control group, although the differences were no longer statistically significant. Overall, our results suggest that our lightweight, scalable digital literacy interventions can boost deepfake discernment without reducing trust in real images.

Our interventions offer several practical advantages that support their real-world applicability. First, our toolbox of digital literacy interventions includes a variety of formats (i.e., textual, visual, gamified, etc.). Having such a broad toolbox is beneficial because it allows the interventions to be flexibly deployed and format-matched across different communication channels, such as private messages, social media posts, and advertising campaigns. Second, our interventions are scalable and require no specialized technical infrastructure or expertise for deployment. In particular, the skills acquired through our interventions can be applied even when online platforms do not remove deepfakes from their platforms or cooperate in detection efforts. Third, unlike other behavioral interventions, such as fact checking or media coverage of misinformation [57], our interventions improve deepfake detection capabilities without undermining trust in real content. This also distinguishes our interventions from other media literacy tips [58, 59] that often suffer from the unintended consequence of reducing overall trust in authentic content. In contrast, our deepfake-specific training maintains participants' ability to trust real images while boosting their ability to identify deepfakes.

Our study design offers several methodological strengths that contribute to the validity and generalizability of our findings. First, our comparative design allows us to systematically evaluate

different intervention formats, which enables us to isolate the effects of presentation modality on learning outcomes [37, 54]. Second, our longitudinal follow-up design provides rare evidence on the persistence of digital literacy gains, addressing a critical gap in understanding whether intervention effects maintain over time [37].

Our research remains highly relevant, even as AI tools continue to advance and deepfakes become more realistic. First, the cognitive skills targeted by our interventions, such as visual discrimination, analytical judgement, and contextual reasoning, are not tied to any specific generation technique. Instead, they directly mitigate the potential for deception, which means that these skills remain applicable even as surface-level flaws in deepfakes diminish [23, 75]. Second, many disinformation campaigns do not strive for perfect realism; instead, in practice, malicious actors often use lower-quality models that are faster and cheaper to mislead in low-attention or emotionally charged environments [76]. Third, there are growing regulatory efforts aimed at restricting access to advanced generative models or limiting the capabilities of these models [77]. If such policies succeed, many real-world deepfakes may continue to be produced with mid-tier tools where human detection remains effective. Fourth, the intervention effects we study are not only useful for visual content; they may transfer to other forms of synthetic media, such as AI-generated audio or video, which are increasingly part of disinformation campaigns. Taken together, these points suggest that digital literacy, rather than automated detection tools alone, will play a crucial part in the long-term response to deepfakes.

As with any other study, ours is also subject to limitations. First, the experimental setup in our study does not fully replicate real-world online environments, where users are embedded in fast-paced and emotionally-charged contexts such as those of social media platforms. While such controlled experimental setups are common in research testing behavioral interventions designed for online settings [37, 78], future research could implement and evaluate our digital literacy interventions in field experiments. Second, our interventions are intentionally designed to be lightweight, minimally-disruptive, and scalable, which may have limited their long-term effectiveness. Re-

peated exposure or multi-session training as in [37] may lead to stronger or more persistent effects. Third, while we assess the risk of increased skepticism toward real images, future research should more deeply examine possible downstream effects of deepfake literacy training on trust in institutions, media sources, or interpersonal communication.

AI-generated disinformation poses a serious threat to information integrity, such as eroding public trust, fueling disinformation campaigns, and enabling political manipulation [75]. In response, our study demonstrates that simple, scalable digital literacy interventions can boost people's ability to discern deepfakes without diminishing their trust in authentic content. Our findings offer actionable insights for policy-makers, educators, and social media platforms seeking to counter deepfakes through evidence-based behavioral interventions. As generative AI technology continues to advance, empowering users with practical, transferable detection skills will be crucial for safeguarding public discourse and fostering resilient societies.

Acknowledgments

Funding by the German Research Foundation (Grant: 543018872) is acknowledged.

Competing interests

The authors declare no competing interests.

References

[1] Baidoo-Anu, D. & Ansah, L. O. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* **7**, 52–62 (2023).

[2] de Ruiter, A. The distinct wrong of deepfakes. *Philosophy & Technology* **34**, 1311–1332 (2021).

[3] Naeem, S. *et al.* Generation and detection of sign language deepfakes - A linguistic and visual analysis. *arXiv:2404.01438* (2024).

[4] Feuerriegel, S. *et al.* Research can help to tackle AI-generated disinformation. *Nature Human Behaviour* **7**, 1818–1821 (2023).

[5] Spitale, G., Biller-Andorno, N. & Germani, F. AI model GPT-3 (dis)informs us better than humans. *Science Advances* **9**, eadh1850 (2023).

[6] Zhou, J. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. *CHI Conference on Human Factors in Computing Systems* (2023).

[7] Kupferschmidt, K. A field's dilemmas: Misinformation research has exploded. But scientists are still grappling with fundamental challenges. *Science* **386**, 478–482 (2024).

[8] Kreps, S., McCain, R. M. & Brundage, M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* **9**, 104–117 (2022).

[9] DiResta, R. & Goldstein, J. A. How spammers and scammers leverage AI-generated images on Facebook for audience growth. *Harvard Kennedy School Misinformation Review* **5**, 4 (2024).

[10] BBC. Fake Trump arrest photos: How to spot an AI-generated image (2023). URL https://www.bbc.com/news/world-us-canada-65069316.

[11] Forbes. Why did 'Balenciaga Pope' go viral? (2023). URL https://www.forbes.com/sites/danidiplacido/2023/03/27/why-did-balenciaga-pope-go-viral/.

[12] BBC. Deepfake presidents used in Russia-Ukraine war (2022). URL https://www.bbc.com/news/technology-60780142.

[13] Goldstein, J. A., Chao, J., Grossman, S., Stamos, A. & Tomz, M. How persuasive is AI-generated propaganda? *PNAS Nexus* **3**, pgae034 (2024).

[14] Drolsbach, C. & Pröllochs, N. Characterizing AI-generated misinformation on social media. *arXiv:2505.10266* (2025).

[15] Clark, E. *et al.* All that's human is not gold: Evaluating human evaluation of generated text. *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing* (2021).

[16] Jakesch, M., Hancock, J. T. & Naaman, M. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* **120**, e2208839120 (2023).

[17] Köbis, N. & Mossink, L. D. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior* **114**, 106553 (2021).

[18] Kasra, M., Shen, C. & O'Brien, A. F. Seeing is believing: How people fail to identify fake images on the web. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2018).

[19] Nightingale, S. J., Wade, K. A. & Watson, D. G. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications* **2**, 30 (2017).

[20] Frank, J. *et al.* A representative study on human detection of artificially generated media across countries. *IEEE Symposium on Security and Privacy* 55–73 (2024).

[21] Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* **119**, e2120481119 (2022).

[22] Groh, M. *et al.* Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications* **15**, 7629 (2024).

[23] Goldstein, J. A. *et al.* Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv:2301.04246* (2023).

[24] Vaccari, C. & Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* **6**, 205630512090340 (2020).

[25] Hancock, J. T. & Bailenson, J. N. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking* **24**, 149–152 (2021).

[26] Dobber, T., Metoui, N., Trilling, D., Helberger, N. & de Vreese, C. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics* **26**, 69–91 (2019).

[27] Center for Countering Digital Hate. Fake image factories: How AI image generators threaten election integrity and democracy. URL https://counterhate.com/research/fake-image-factories/.

[28] Ternovski, J., Kalla, J. & Aronow, P. Negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety* **1**, 2 (2022).

[29] Mirsky, Y. & Lee, W. The creation and detection of deepfakes: A survey. *ACM Computing Surveys* **54**, 1–41 (2021).

[30] Shiohara, K. & Yamasaki, T. Detecting deepfakes with self-blended images. *Conference on Computer Vision and Pattern Recognition* (2022).

[31] Korshunov, P. & Marcel, S. Deepfakes: A new threat to face recognition? Assessment and detection. *arXiv:1812.08685* (2018).

[32] Almars, A. M. Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications* **9**, 20–35 (2021).

[33] Zhong, W., Tucker, J. A. & Sanderson, Z. The effect of AI labeling on perceptions of images. *osf.io/zjhd4* (2024).

[34] Kozyreva, A. *et al.* Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* **8**, 1044–1052 (2024).

[35] Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, 6714 (2024).

[36] Arechar, A. A. *et al.* Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* **7**, 1502–1513 (2023).

[37] Maertens, R. *et al.* Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications* **16**, 2062 (2025).

[38] Ali, A. & Qazi, I. A. Countering misinformation on social media through educational interventions: Evidence from a randomized experiment in Pakistan. *Journal of Development Economics* **163**, 103108 (2023).

[39] Apuke, O. D., Omar, B. & Asude Tunca, E. Literacy concepts as an intervention strategy for improving fake news knowledge, detection skills, and curtailing the tendency to share fake news in Nigeria. *Child & Youth Services* **44**, 88–103 (2023).

[40] Badrinathan, S. Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review* **115**, 1325–1341 (2021).

[41] Blair, R. A. *et al.* Interventions to counter misinformation: Lessons from the global north and applications to the Global South. *Current Opinion in Psychology* **55**, 101732 (2024).

[42] Roozenbeek, J., Culloty, E. & Suiter, J. Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist* **28**, 189–205 (2023).

[43] Epstein, Z. *et al.* Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review* (2021).

[44] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science* **31**, 770–780 (2020).

[45] Lee, A. Y., Moore, R. C. & Hancock, J. T. Building resilience to misinformation in communities of color: Results from two studies of tailored digital media literacy interventions. *New Media & Society* **27**, 3545–3576 (2024).

[46] Basol, M., Roozenbeek, J. & van der Linden, S. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition* **3**, 2 (2020).

[47] Guess, A. M. *et al.* 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* **117**, 15536–15545 (2020).

[48] Moore, R. C. & Hancock, J. T. A digital media literacy intervention for older adults improves resilience to fake news. *Scientific Reports* **12**, 6008 (2022).

[49] Panizza, F. *et al.* Lateral reading and monetary incentives to spot disinformation about science. *Scientific Reports* **12**, 5678 (2022).

[50] Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications* **13**, 2333 (2022).

[51] Ruffin, M., Wang, G. & Levchenko, K. Explaining why fake photos are fake. *Proceedings of the ACM on Human-Computer Interaction* **7**, 1–22 (2023).

[52] Aprin, F., Chounta, I. A. & Hoppe, H. U. "See the image in different contexts": Using reverse image search to support the identification of fake news in Instagram-like social media. *Intelligent Tutoring Systems International Conference* **13284** (2022).

[53] Qian, S., Shen, C. & Zhang, J. Fighting cheapfakes: Using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication* **28**, zmac024 (2022).

[54] Yang, T., Allen, R. J., Yu, Q. & Chan, R. C. K. The influence of input and output modality on following instructions in working memory. *Scientific Reports* **5**, 17657 (2015).

[55] Reber, A. S. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* **6**, 855–863 (1967).

[56] Chein, J. M., Martinez, S. A. & Barone, A. R. Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports* **14**, 25989 (2024).

[57] Hoes, E., Aitken, B., Zhang, J., Gackowski, T. & Wojcieszak, M. Prominent misinformation interventions reduce misperceptions but increase skepticism. *Nature Human Behaviour* **8**, 1545–1553 (2024).

[58] Altay, S., de Angelis, A. & Hoes, E. Media literacy tips promoting reliable news improve discernment and enhance trust in traditional media. *Communications Psychology* **2**, 74 (2024).

[59] Guo, S., Swire-Thompson, B. & Hu, X. Specific media literacy tips decrease belief in AI-generated visual misinformation. *osf.io/phr79* (2024).

[60] Cohen, J. *Statistical power analysis for the behavioral sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988), 2nd edn.

[61] Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* **1**, 259–269 (2018).

[62] Haile, Z. T. Power analysis and exploratory research. *Journal of Human Lactation* **39**, 579–583 (2023).

[63] Jeong, S.-H., Cho, H. & Hwang, Y. Media literacy interventions: A meta-analytic review. *Journal of Communication* **62**, 454–472 (2012).

[64] Kamali, N., Nakamura, K., Chatzimparmpas, A., Hullman, J. & Groh, M. How to distinguish AI-generated images from authentic photographs. *arXiv:2406.08651* (2024).

[65] OpenAI. DALL-E 3: Generative image model (2023). URL https://openai.com/dall-e.

[66] Kiryakova, G., Angelova, N. & Yordanova, L. Gamification in education. *International Balkan Education and Science Conference* (2014).

[67] Gordon, P. C. & Holyoak, K. J. Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology* **45**, 492–500 (1983).

[68] Midjourney. Discord channel. URL https://discord.com/channels/662267976984297473/999550150705954856.

[69] Frederick, S. Cognitive reflection and decision making. *Journal of Economic Perspective* **19**, 25–42 (2005).

[70] Bashardoust, A., Feuerriegel, S. & Shrestha, Y. R. Comparing the willingness to share for human-generated vs. AI-generated fake news. *Computer Supported Cooperative Work* (2024).

[71] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. M. J. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).

[72] Roozenbeek, J. *et al.* Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making* **17**, 547–573 (2022).

[73] Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).

[74] Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends in cognitive sciences* **25**, 388–402 (2021).

[75] Diakopoulos, N. & Johnson, D. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society* **23**, 2072–2098 (2021).

[76] Łabuz, M. & Nehring, C. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science* **23**, 454–473 (2024).

[77] Romero Moreno, F. Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology* **38**, 297–326 (2024).

[78] Spampatti, T., Hahnel, U. J. J., Trutnevyte, E. & Brosch, T. Psychological inoculation strategies to fight climate disinformation across 12 countries. *Nature Human Behaviour* **8**, 380–398 (2024).

[79] Bennett, J. Improving the way we categorize family income (2021). URL https://www.pewresearch.org/decoded/2021/06/04/improving-the-way-wecategorize-family-income/.

[80] O'brien, R. M. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* **41**, 673–690 (2007).

A Variables

Variable	Question	Options
Discernment†	*"Is this image real or fake?"*	definitely fake / probably fake / probably real / definitely real
SharingIntention	*"Would you share this image on your social media?"*	yes (1) / no / don't know (0)
Confidence	*"How confident are you about your fake image detection choices?"*	not confident at all (0) / ... / very confident (100)

†Encoded as discernment accuracy for each participant in the analysis

Table S1: Dependent variables.

Variable	Question	Options
Sociodemographics		
Age	*"How old are you? (in years)"*	Continuous variable
Ethnicity*	*"Select all that apply to you"*	White (0) / Black or African American (1) / Hispanic, Latino or Spanish origin (2) / Asian (3) / American Indian or Alaska Native / Middle Eastern or North African / Native Hawaiian or Other Pacific Islander / some other race, ethnicity, or origin (4)
Gender**	*"What is your gender?"*	male (0) / female (1) / non-binary / third gender / other / prefer not to say
Education***	*"How many years of formal education have you completed?"*	7–12 (up to high school) (0) / 13–16 (college / undergraduate university / certificate training) (1) / more than 17 years (doctorate degree, medical degree, etc.) (2) / 0–6 (up to primary school) / prefer not to say
Religion†	*"What describes you best?"*	a religious person (0) / not a religious person (1) / prefer not to say
Social orientation‡	*"What is your political orientation for social issues? For example in health care, education, etc* *Here: 'liberal' means classically left-wing; 'conservative' means classically right-wing"*	very liberal (1) / liberal (2) / somewhat liberal (3) / neutral (4) / somewhat conservative (5) / conservative (6) / very conservative (7)
Economic orientation‡	*"What is your political orientation for economic issues? For example in taxes, etc* *Here: 'liberal' means classically left-wing; 'conservative' means classically right-wing"*	very liberal (1) / liberal (2) / somewhat liberal (3) / neutral (4) / somewhat conservative (5) / conservative (6) / very conservative (7)
Income§	*"What is your total yearly family/household income?"*	[$10,000, $19,999] / [$20,000, $29,999] / [$30,000, $39,999] / [$40,000, $49,999] / [$50,000, $59,999] / [$60,000, $69,999] / [$70,000, $79,999] / [$80,000, $89,999] / [$90,000, $99,999] / [$100,000, $109,999] / [$110,000, $119,999] / [$120,000, $129,999] / [$130,000, $139,999] / [$140,000, $149,999] / [$150,000 or more] encoded as [$10,000, $29,999]: low (0) / [$30,000, $79,999]: middle (1) / $\geq$ $80,000: high (2)
Status	*"Think of this ladder as representing where people stand in your country. Where would you place yourself on this ladder? (10 equals highest, 1 equals lowest)"*	1 (0) / . . . / 10 (9)
Social media use		
PlatformCount	*"Which social media platforms do you use (if any)?"*	Facebook / Twitter/X / Snapchat / Instagram / WhatsApp / TikTok / other / none; encoded as count
SharingCount	*"Which type of content would you consider sharing on social media (if any)?"*	political news / sports news / celebrity news / science/technology news / business news / personal content / other / none; encoded as count
TimeOnline	*"How much time (in hours) on average do you spend on social media daily?"*	Continuous variable
Digital literacy		
KnowledgeDeepfakes	*"How do you rate your knowledge about deepfakes?"*	no knowledge at all (0) / . . . / very knowledgeable (7)
ExpDetecting	*"How do you rate your experience with detecting deepfakes?"*	no experience at all (0) / . . . / very experienced (7)
ExpSearchEngines	*"How do you rate your experience with using online search engines? For example with Google, Bing, etc"*	no experience at all (0) / . . . / very experienced (7)
ExpImageSearch	*"How do you rate your experience with using reverse image search?"*	no experience at all (0) / . . . / very experienced (7)
ExpGenAI	*"How do you rate your experience with using AI image generators. For example with DALL-E, Midjourney, etc"*	no experience at all (0) / . . . / very experienced (7)
CRT	*"The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How old is Adam?"* *"If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take 50 printers to print out 50 pages of paper?"* *"On a loaf of bread, there is a patch of mold. Every day, the patch doubles in size. If it takes 40 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?"*	all incorrect (0) / . . . / all correct (3)

* We grouped ethnicities with lower counts into one group.

** In our study, only 10 subjects identified themselves as being different from male or female and 1 subject preferred not to answer. Therefore, in our analysis, we encoded Gender as a binary variable and removed other observations from the analysis due to the low statistical power.

*** In our study, only 6 subjects did not answer and 1 subject had less than 7 years of formal education. Therefore, in our analysis, we removed these observations.

† We encoded religion as a binary variable and removed observations where subjects preferred not to answer.

‡ The variables were averaged into a single score that represents political leaning encoded as very liberal (1) / liberal (2) / somewhat liberal (3) / neutral (4) / somewhat conservative (5) / conservative (6) / very conservative (7).

§ We transform our income values into three categories of incomes: low, middle, and high. This coding is based on [79], and we use $80,000 as the threshold between middle and high. We removed observations where subjects preferred not to answer.

Table S2: Covariates.

B Example discernment task

For the image discernment task, we show 15 images to each participant and ask them about the image's veracity as well as their willingness to share the image. In addition, we include an optional question where participants can tick a checkbox if they had trouble seeing the image or if they have seen the image before. A screenshot of the question setup can be found in Supplementary Figure S1.

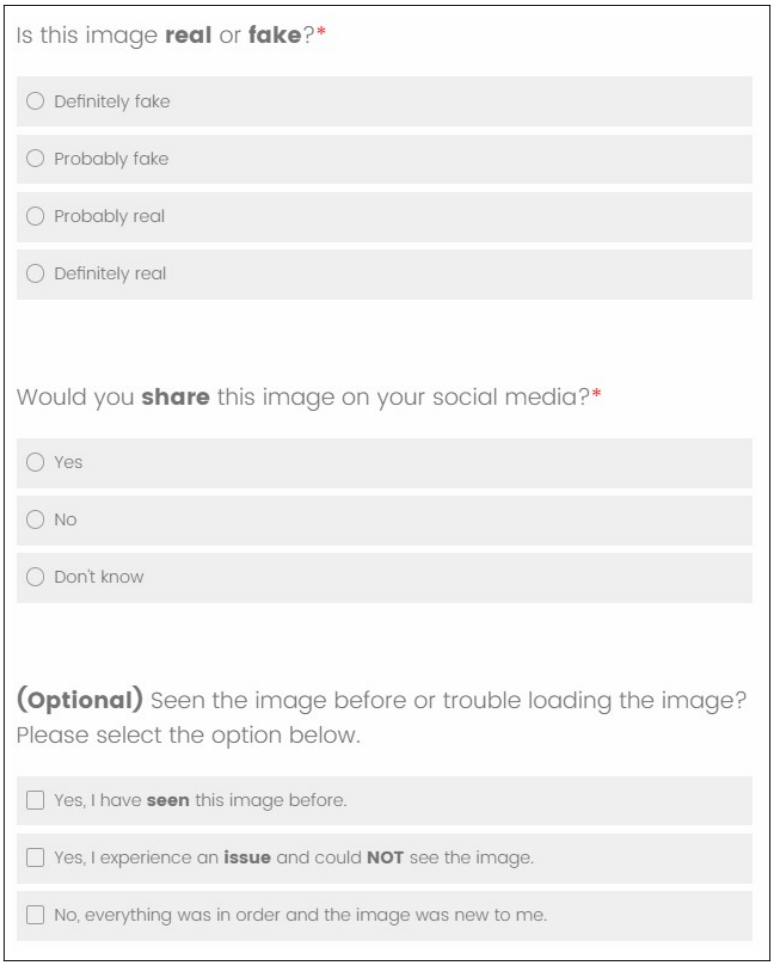


Figure S1: **Screenshot of the question setup for each image rating task.**

C Normality check

To assess whether parametric statistical tests are appropriate, we conducted Shapiro-Wilk tests for normality on all outcome variables across experimental conditions. Table S3 shows that none of the outcome measures (real image discernment, deepfake discernment, real image sharing intention, deepfake sharing intention) were normally distributed in any condition (all $p < 0.001$). Specifically, this was the case for the control condition as well as all intervention conditions: *Textual*, *Visual*, *Gamified*, *Feedback*, and *Knowledge*. Given these violations of the normality assumption, we used the Mann-Whitney U test as a non-parametric alternative to the independent samples t-test in order to compare differences between conditions.

Condition	**Variable**	W**-statistic**	p**-value**
Control	Real image discernment	0.7756	< 0.001
Textual	Real image discernment	0.7756	< 0.001
Visual	Real image discernment	0.8038	< 0.001
Gamified	Real image discernment	0.7887	< 0.001
Feedback	Real image discernment	0.7583	< 0.001
Knowledge	Real image discernment	0.7863	< 0.001
Control	Deepfake discernment	0.9666	< 0.001
Textual	Deepfake discernment	0.9368	< 0.001
Visual	Deepfake discernment	0.9280	< 0.001
Gamified	Deepfake discernment	0.9540	< 0.001
Feedback	Deepfake discernment	0.9667	< 0.001
Knowledge	Deepfake discernment	0.9641	< 0.001
Control	Real image sharing intention	0.7472	< 0.001
Textual	Real image sharing intention	0.7551	< 0.001
Visual	Real image sharing intention	0.7144	< 0.001
Gamified	Real image sharing intention	0.7011	< 0.001
Feedback	Real image sharing intention	0.7022	< 0.001
Knowledge	Real image sharing intention	0.6932	< 0.001
Control	Deepfake sharing intention	0.7416	< 0.001
Textual	Deepfake sharing intention	0.6958	< 0.001
Visual	Deepfake sharing intention	0.6778	< 0.001
Gamified	Deepfake sharing intention	0.7187	< 0.001
Feedback	Deepfake sharing intention	0.7065	< 0.001
Knowledge	Deepfake sharing intention	0.6923	< 0.001

Table S3: Shapiro-Wilk normality test results for all outcome variables across conditions.

D Sociodemographics of participants

Sociodemographics	N	%
Gender		
Female	586	52.7
Male	515	46.3
Non-binary/Third gender/Other	10	0.9
Prefer not to say	1	0.1
Age		
18–34 years	410	36.9
35–54 years	506	45.5
55+ years	194	17.5
Mean (SD)	41.7 (13.3)	
Range	18-82	
Ethnicity		
White	786	70.6
Black or African American	192	17.3
Asian	39	3.5
Hispanic, Latino or Spanish origin	26	2.3
Other	69	6.2
Education		
0-6 (up to primary School)	1	0.1
Up to high school (7–12 years)	153	13.8
College/Undergraduate (13–16 years)	683	61.4
Graduate/Professional (17+ years)	269	24.2
Prefer not to say	6	0.5
Religious status		
Not religious	617	55.5
Religious	444	39.9
Prefer not to answer	51	4.6
Income group		
Low	164	14.8
Medium	422	38.1
High	503	45.2
Prefer not to say	23	2.1
Political orientation		
Very liberal (1–1.5)	135	12.1
Liberal (2–2.5)	210	18.9
Somewhat liberal (3–3.5)	117	10.5
Neutral (4)	200	18.0
Somewhat conservative (4.5–5)	130	11.7
Conservative (5.5–6)	205	18.5
Very conservative (6.5–7)	115	10.3

SD: standard deviation

Table S4: Sociodemographics of participants after filtering ($N = 1,112$).

E Robustness checks

E.1 Robustness check including participants failing attention and honesty checks

We repeated the analyses from Section 2.8 while including the participants who failed both attention and/or both honesty checks. For deepfake discernment, we find robust evidence of our findings. In particular, the effects of our *Textual* and *Visual* interventions remain significant (*Textual*: Mann-Whitney $U = 14306.5$, $p = 0.001^{**}$; *Visual*: Mann-Whitney $U = 11696.0$, $p < 0.001^{***}$; *Gamified*: Mann-Whitney $U = 15298.5$, $p = 0.052$; *Feedback*: Mann-Whitney $U = 17880.0$, $p = 0.636$; and *Knowledge*: Mann-Whitney $U = 15970.5$, $p = 0.170$).

We also repeated the equivalence tests to check for meaningful differences in the discernment ability for real images using a threshold of $[-0.05, +0.05]$. The equivalence tests for each intervention were undecided (*Textual*: $p = 0.683$, 95% CI: $[-3.100, 5.227]$; *Gamified*: $p = 0.883$, 95% CI: $[-1.657, 7.042]$; *Feedback*: $p = 0.530$, 95% CI: $[-4.280, 3.860]$; *Knowledge*: $p = 0.500$, 95% CI: $[-3.894, 3.997]$). Only, intervention *Visual* showed a significant difference ($p = 0.977$, 95% CI: $[0.120, 9.082]$). This suggests that our findings are largely robust.

To check for robustness of our follow-up results, we also repeat the Mann-Whitney U tests for deepfake discernment in the follow-up without excluding participants who failed the attention and honesty checks. We find robust results of our findings (*Textual*: Mann-Whitney $U = 8635.5$, $p = 0.105$; *Visual*: Mann-Whitney $U = 7724.0$, $p = 0.110$; *Gamified*: Mann-Whitney $U = 9132.5$, $p = 0.421$; *Feedback*: Mann-Whitney $U = 9654.5$, $p = 0.917$; and *Knowledge*: Mann-Whitney $U = 8915.5$, $p = 0.922$).

E.2 Discernment accuracy for previously seen images or loading issues

In the main analysis, we excluded all responses where participants indicated they had previously seen the image or experienced loading issues. Here, we repeated the analysis without excluding these responses and re-computed discernment accuracy for both real images and deepfakes. We again test for differences in discernment of deepfakes for each intervention against the control group using Mann-Whitney U tests. We find that the effects of our *Textual* and *Visual* interventions remain significant (*Textual*: Mann-Whitney $U = 58160.0$, $p < 0.001^{***}$; *Visual*: Mann-Whitney $U = 47543.5$, $p < 0.001^{***}$). Additionally, the effects of our intervention *Gamified* are now significant (Mann-Whitney $U = 62189.0$, $p = 0.013^{*}$). This might be the case because we now keep responses where participants had seen the image before, so participants might already know if the image is a deepfake from previous experience, which might affect their discernment accuracy. Further, our findings remain robust (*Feedback*: Mann-Whitney $U = 73192.5$, $p = 0.296$; and *Knowledge*: Mann-Whitney $U = 64471.0$, $p = 0.081$).

For the discernment of real images, we again repeat the equivalence tests using a threshold of $[-0.05, +0.05]$. Our findings are largely robust as we find that the equivalence tests for each intervention were undecided (*Textual*: $p = 0.734$, 95% CI: $[-1.919, 3.872]$; *Gamified*: $p = 0.944$, 95% CI: $[-0.513, 5.547]$; *Feedback*: $p = 0.523$, 95% CI: $[-2.971, 2.700]$; *Knowledge*: $p = -0.039$, 95% CI: $[-2.761, 2.683]$). Only, intervention *Visual* showed a significant difference ($p = 0.992$, 95% CI: $[0.774, 6.915]$).

We also repeat the Mann-Whitney U tests for the discernment of deepfakes in the follow-up while keeping the responses where participants indicted they had previously seen the image of experienced loading issues. We find that the effects of our *Textual* and *Visual* interventions were now significant (Mann-Whitney $U = 34853.5$, $p = 0.033^{*}$ and $U = 31064.0$,$p = 0.030^{*}$, respectively). The effects of our other interventions remain robust (*Gamified*: $U = 36761.5$,$p = 0.203$; *Feedback*: $U = 38855.5$,$p = 0.782$; and *Knowledge*: $U = 35870.0$,$p = 0.982$).

E.3 Regression analysis with sociodemographic controls

We now conduct a regression analysis where we control for sociodemographic covariates, namely, gender, age, ethnicity, education, religion, political orientation, and income (as defined in Supplementary Table S2). We removed observations with low-frequency categories or missing responses (e.g., because participants answered "prefer not to say"). That is, we removed 10 observations where participants identified as non-binary/Third gender/Other due to the limited sample size in the study and 1 participant who preferred not to report their gender. We also excluded observations with missing data on education (6 people) and those with less than 7 years of formal education (1 person). Similarly, we removed participants who preferred not to answer on their income (23 people) and religion (51 people), as well as 2 people who reported to be younger than 18 (see Supplementary Section D for more details on the distribution of sociodemographic in our experiment). After removing these observations, $N = 1034$ participants remained for the regression analysis.

Supplementary Table S5 reports the results from the OLS regression model (as defined in Equation 1) for the different interventions. Model (1) OVERALL estimates the effect of each of our interventions (all included in the model as a binary variable). In models (2) TEXTUAL, (3) VISUAL, (4) GAMIFIED, (5) FEEDBACK, and (6) KNOWLEDGE, we only include individuals from the respective intervention condition and the control condition to isolate the effect of our intervention on the outcome. Model (7) CONTROL estimates direct effects of sociodemographics on the outcome within the control condition. We find that the estimated coefficients for the interventions remain robust. We find that religion tends to play a significant role in the discernment ability of participants and that non-religious people tend to be better at discerning deepfakes.

	OVERALL	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE	CONTROL
Intercept	63.400***	67.954***	66.526***	55.851***	64.323***	67.370***	66.622***
	(3.866)	(6.761)	(6.124)	(6.845)	(6.215)	(6.279)	(9.254)
Textual	8.112***	8.089**					
	(2.423)	(2.597)					
Visual	13.491***		13.493***				
	(2.443)		(2.402)				
Gamified	5.670*			5.344*			
	(2.436)			(2.517)			
Feedback	–0.832				–1.245		
	(2.409)				(2.488)		
Knowledge	3.438					3.326	
	(2.436)					(2.480)	
Gender	–0.172	1.316	–4.571	–2.813	–0.533	–3.540	–4.495
	(1.418)	(2.645)	(2.423)	(2.558)	(2.549)	(2.503)	(3.742)
Age	–0.024	–0.128	–0.167	0.050	–0.154	–0.126	–0.224
	(0.054)	(0.101)	(0.093)	(0.096)	(0.092)	(0.087)	(0.132)
Ethnicity	0.243	–0.883	–2.249*	–1.335	–0.112	–0.559	–3.141
	(0.657)	(1.272)	(1.089)	(1.233)	(1.184)	(1.226)	(1.813)
Education	–1.178	0.075	1.828	2.206	0.647	2.836	4.415
	(1.228)	(2.197)	(2.044)	(2.160)	(2.068)	(2.189)	(3.020)
Religion	5.387***	5.005	6.141*	7.581**	6.292*	7.061*	8.021
	(1.561)	(2.954)	(2.646)	(2.897)	(2.796)	(2.757)	(4.148)
PoliticalOrientation	–0.186	–0.031	0.686	0.242	0.738	–0.017	1.217
	(0.410)	(0.788)	(0.685)	(0.750)	(0.743)	(0.727)	(1.089)
Income	–1.726	–3.446	–1.919	–1.338	–2.452	–3.944*	–4.103
	(1.047)	(1.924)	(1.755)	(1.865)	(1.820)	(1.880)	(2.664)
R^2	0.068	0.058	0.128	0.046	0.025	0.054	0.070
Adj. R^2	0.057	0.035	0.107	0.023	0.002	0.032	0.029
N	1034	345	338	339	348	340	169

Table S5: **Regression results for discernment of deepfakes while controlling for sociodemographics.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in the models did not exceed 10, indicating that multicollinearity is not a major concern [80] and that intervention effects remain reliable. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.4 Regression analyzing the varying effectiveness of interventions across sociodemographics)

We now control for sociodemographic covariates and their possible interaction effects with our interventions. Supplementary Table S6 shows the results of our OLS regression models with interaction terms for participant-specific sociodemographic controls as defined in Equation 2. We find that the estimated coefficients for the interventions remain robust. Overall, we find evidence that our interventions are again more effective for non-religious participants. Moreover, the *Textual* intervention was significantly more effective for females and less effective for participants with higher education. The *Gamified* intervention was significantly more effective for older participants. Both *Feedback* and *Knowledge* interventions are more effective for non-white participants.

	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE
Intercept	66.622***	66.622***	66.622***	66.622***	66.622***
	(9.247)	(8.695)	(9.029)	(9.025)	(8.914)
Intervention	11.926	13.528	–19.719	–6.676	5.287
	(13.374)	(12.110)	(13.524)	(12.193)	(12.402)
Gender	–4.495	–4.495	–4.495	–4.495	–4.495
	(3.739)	(3.516)	(3.651)	(3.650)	(3.605)
Intervention × *Gender*	11.982*	0.403	4.211	7.775	1.644
	(5.287)	(4.939)	(5.113)	(5.110)	(5.033)
Age	–0.224	–0.224	–0.224	–0.224	–0.224
	(0.132)	(0.124)	(0.129)	(0.129)	(0.127)
Intervention × *Age*	0.222	0.095	0.592**	0.149	0.198
	(0.210)	(0.192)	(0.193)	(0.183)	(0.175)
Ethnicity	–3.141	–3.141	–3.141	–3.141	–3.141
	(1.812)	(1.704)	(1.769)	(1.768)	(1.747)
Intervention × *Ethnicity*	4.735	1.802	3.219	5.785*	5.288*
	(2.519)	(2.234)	(2.455)	(2.378)	(2.454)
Education	4.415	4.415	4.415	4.415	4.415
	(3.017)	(2.837)	(2.946)	(2.945)	(2.909)
Intervention × *Education*	–11.181*	–6.000	–5.093	–8.258*	–4.681
	(4.391)	(4.176)	(4.327)	(4.138)	(4.452)
Religion	8.021	8.021*	8.021*	8.021*	8.021*
	(4.145)	(3.897)	(4.047)	(4.045)	(3.995)
Intervention × *Religion*	–6.210	–3.574	0.845	–2.596	–1.554
	(5.870)	(5.375)	(5.785)	(5.610)	(5.511)
PoliticalOrientation	1.217	1.217	1.217	1.217	1.217
	(1.089)	(1.024)	(1.063)	(1.062)	(1.049)
Intervention × *PoliticalOrientation*	–2.249	–0.844	–1.296	–0.499	–2.386
	(1.573)	(1.393)	(1.496)	(1.481)	(1.452)
Income	–4.103	–4.103	–4.103	–4.103	–4.103
	(2.662)	(2.503)	(2.599)	(2.598)	(2.566)
Intervention × *Income*	1.327	4.472	5.008	3.124	0.734
	(3.842)	(3.599)	(3.728)	(3.631)	(3.799)
R^2	0.107	0.138	0.086	0.060	0.082
Adj. R^2	0.067	0.098	0.043	0.018	0.039
N	345	338	339	348	340

Table S6: **Regression results for discernment of deepfakes while controlling for sociodemographics and their interactions with the interventions.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in the models did not exceed 10, indicating that multicollinearity is not a major concern [80] and that intervention effects remain reliable. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.5 Regression analysis controlling for self-reported digital literacy of participants

We also control for the digital literacy of participants in our robustness checks. For this, we collected the participants knowledge about deepfakes and their experience in detecting deepfakes, using online search engines, using reverse image search, and using AI image generators (see Supplementary Table S2 for details). All covariates were collected on a 7-point Likert scale.

Supplementary Table S7 shows the results of the OLS regression model as defined in Equation 1 where we control for digital literacy. Again, we find that the estimated coefficients for the interventions remain robust. We also find that previous experience with search engines (covariate *ExpSearchEngine*) has a significant direct effect on people's deepfake discernment abilities. This suggests that people who have more experience with search engines tend to do better at discerning deepfakes.

	OVERALL	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE	CONTROL
Intercept	50.600***	44.784***	48.190***	63.109***	54.784***	54.955***	56.434***
	(3.977)	(6.571)	(6.122)	(6.693)	(6.373)	(6.484)	(8.867)
Textual	7.270**	7.190**					
	(2.313)	(2.431)					
Visual	12.701***		12.982***				
	(2.340)		(2.311)				
Gamified	4.335			4.351			
	(2.326)			(2.442)			
Feedback	–1.438				–1.180		
	(2.334)				(2.389)		
Knowledge	2.854					3.046	
	(2.326)					(2.389)	
KnowledgeDeepfakes	–0.496	–0.168	0.180	–2.088	–2.836	–0.981	–2.122
	(0.872)	(1.664)	(1.597)	(1.475)	(1.483)	(1.537)	(2.246)
ExpDetecting	0.655	0.479	–0.322	1.193	1.095	–0.291	–0.067
	(0.794)	(1.521)	(1.464)	(1.339)	(1.431)	(1.389)	(2.120)
ExpSearchEngine	2.188**	3.030*	1.565	0.323	1.422	1.642	0.896
	(0.740)	(1.289)	(1.243)	(1.261)	(1.250)	(1.293)	(1.769)
ExpImageSearch	0.024	0.750	1.077	–0.448	1.086	0.026	0.978
	(0.476)	(0.857)	(0.808)	(0.841)	(0.844)	(0.818)	(1.193)
ExpGenAI	–0.305	–0.864	0.751	0.611	0.823	0.872	1.793
	(0.475)	(0.910)	(0.816)	(0.927)	(0.851)	(0.860)	(1.355)
R^2	0.053	0.047	0.104	0.015	0.020	0.013	0.024
Adj. R^2	0.044	0.031	0.089	–0.001	0.004	–0.004	–0.003
N	1112	375	370	371	372	372	187

Table S7: **Regression results for discernment of deepfakes while controlling for digital literacy.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in all models exceeded 10 for *KnowledgeDeepfakes*, *ExpDetecting*, and *ExpSearchEngine*, which means that we can only interpret the coefficients but not the standard errors or significance levels [80]. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.6 Regression analyzing the varying effectiveness of interventions across self-reported digital literacy

We also control for interactions between the digital literacy covariates and our interventions with an OLS regression model as defined in Equation 2. Supplementary Table S8 shows the results of models TEXTUAL, VISUAL, GAMIFIED, FEEDBACK, KNOWLEDGE, where we only include participants from the respective intervention condition and the control condition. We find that the estimated coefficients for the interventions remain robust. Moreover, we find that the *Textual* intervention were less effective for participants who are more experienced in AI image generation. This might be the explained by the fact that participants with experience in AI image generation already know about the typical errors found in deepfakes, which may have led to a ceiling effect, limiting their potential for further improvement. Another possibility is that they were overconfident in their expertise which might have lead to reduced attention to the intervention content.

	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE
Intercept	56.434***	56.434***	56.434***	56.434***	56.434***
	(8.673)	(8.158)	(8.730)	(8.599)	(8.571)
Intervention	–19.022	–3.892	18.568	–5.560	–1.501
	(12.925)	(12.272)	(13.355)	(12.876)	(13.119)
KnowledgeDeepfakes	–2.122	–2.122	–2.122	–2.122	–2.122
	(2.197)	(2.067)	(2.211)	(2.178)	(2.171)
Intervention × *KnowledgeDeepfakes*	4.324	5.327	0.693	–1.305	2.423
	(3.326)	(3.280)	(2.984)	(3.019)	(3.081)
ExpDetecting	–0.067	–0.067	–0.067	–0.067	–0.067
	(2.074)	(1.951)	(2.087)	(2.056)	(2.050)
Intervention × *ExpDetecting*	0.101	–1.537	1.737	2.089	–0.741
	(3.046)	(3.005)	(2.735)	(2.923)	(2.821)
ExpSearchEngine	0.896	0.896	0.896	0.896	0.896
	(1.730)	(1.627)	(1.741)	(1.715)	(1.710)
Intervention × *ExpSearchEngine*	4.887	1.624	–1.332	1.179	1.636
	(2.562)	(2.526)	(2.530)	(2.525)	(2.624)
ExpImageSearch	0.978	0.978	0.978	0.978	0.978
	(1.167)	(1.098)	(1.175)	(1.157)	(1.153)
Intervention × *ExpImageSearch*	–0.376	–0.112	–2.917	0.146	–1.798
	(1.696)	(1.634)	(1.678)	(1.717)	(1.657)
ExpGenAI	1.793	1.793	1.793	1.793	1.793
	(1.325)	(1.246)	(1.334)	(1.314)	(1.309)
Intervention × *ExpGenAI*	–4.926**	–1.796	–2.205	–1.644	–1.626
	(1.811)	(1.654)	(1.867)	(1.732)	(1.759)
R^2	0.084	0.117	0.036	0.024	0.024
Adj. R^2	0.056	0.090	0.007	–0.005	–0.006
N	375	370	371	372	372

Table S8: **Regression results for discernment of deepfakes while controlling for self-reported digital literacy and the interactions with the interventions.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in all models exceeded 10 for *KnowledgeDeepfakes*, *ExpDetecting*, and *ExpSearchEngine*, which means that we can only interpret the coefficients but not the standard errors or significance levels [80]. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.7 Regression analysis controlling for social media use of participants

Further, we control for the social media use of the participants when estimating the effects of our interventions on deepfake discernment. For this, we collected the number of platforms that participants use, what type of content they share, and how much time they spend on social media daily (see Supplementary Table S2 for details).

Supplementary Table S9 reports the result for the OLS regression model as defined in Equation 1. Again, we find robust results for the estimated coefficients. We also find indications that people who tend to share more content on social media and people who tend to spend more time online are less good at discerning deepfakes.

	OVERALL	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE	CONTROL
Intercept	64.708***	63.507***	62.633***	63.973***	65.825***	62.768***	62.326***
	(2.253)	(3.310)	(3.164)	(3.227)	(3.181)	(3.215)	(4.393)
Textual	7.702***	7.761**					
	(2.299)	(2.414)					
Visual	13.056***		13.136***				
	(2.315)		(2.285)				
Gamified	4.650*			4.537			
	(2.311)			(2.417)			
Feedback	–1.391				–1.443		
	(2.310)				(2.365)		
Knowledge	2.939					2.902	
	(2.308)					(2.356)	
PlatformCount	0.521	1.296	0.763	0.860	0.759	1.447	1.690
	(0.465)	(0.811)	(0.820)	(0.828)	(0.843)	(0.845)	(1.219)
SharingCount	–1.493**	–1.992*	–0.986	–2.132*	–2.106*	–1.579	–2.242
	(0.476)	(0.861)	(0.820)	(0.900)	(0.863)	(0.839)	(1.279)
TimeOnline	–0.496**	–0.603*	–0.520	–0.119	–0.639	–0.885	–0.468
	(0.170)	(0.296)	(0.361)	(0.319)	(0.367)	(0.580)	(0.792)
R^2	0.061	0.052	0.091	0.025	0.030	0.023	0.021
Adj. R^2	0.054	0.041	0.081	0.014	0.019	0.013	0.005
N	1112	375	370	371	372	372	187

Table S9: **Regression results for discernment of deepfakes while controlling for participants' social media use.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in the models did not exceed 10, indicating that multicollinearity is not a major concern [80] and that intervention effects remain reliable. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.8 Regression analyzing the varying effectiveness of interventions across participants social media use

To account for possible interaction effects, we run OLS regression models as defined in Equation 2 with social media covariates. Supplementary Tables S10 show the results. Here, the models TEXTUAL, VISUAL, GAMIFIED, FEEDBACK, KNOWLEDGE only include individuals from the respective intervention condition and the control condition to isolate the effect of our intervention on the outcome. We find that the estimated coefficients for the interventions remain robust.

	TEXTUAL	VISUAL	GAMIFIED	FEEDBACK	KNOWLEDGE
Intercept	62.326***	62.326***	62.326***	62.326***	62.326***
	(4.362)	(4.081)	(4.337)	(4.230)	(4.234)
Intervention	10.036	13.915*	8.031	5.703	3.952
	(6.257)	(5.909)	(6.112)	(5.974)	(5.970)
PlatformCount	1.690	1.690	1.690	1.690	1.690
	(1.211)	(1.133)	(1.204)	(1.174)	(1.175)
Intervention × *PlatformCount*	–0.783	–1.858	–1.529	–2.033	–0.370
	(1.650)	(1.653)	(1.674)	(1.697)	(1.700)
SharingCount	–2.242	–2.242	–2.242	–2.242	–2.242
	(1.270)	(1.189)	(1.263)	(1.232)	(1.233)
Intervention × *SharingCount*	0.400	2.384	0.302	0.301	1.269
	(1.740)	(1.647)	(1.810)	(1.730)	(1.687)
TimeOnline	–0.468	–0.468	–0.468	–0.468	–0.468
	(0.787)	(0.736)	(0.782)	(0.763)	(0.764)
Intervention × *TimeOnline*	–0.161	–0.051	0.410	–0.190	–0.976
	(0.850)	(0.845)	(0.857)	(0.872)	(1.180)
R^2	0.052	0.097	0.027	0.035	0.026
Adj. R^2	0.034	0.079	0.008	0.016	0.008
N	375	370	371	372	372

Table S10: **Regression results for discernment of deepfakes while controlling for participants' social media use and the interactions with the interventions.** Parentheses report standard errors. Variance inflation factor (VIF) values for the predictors in the models did not exceed 10, indicating that multicollinearity is not a major concern [80] and that intervention effects remain reliable. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E.9 Differences in sociodemographics and discernment for attrition group

The response rate for our follow-up was 72.93% of our participants. Here, we check whether there are significant differences between the participants that returned for the follow-up (respondents) and those who did not (dropouts) using Mann-Whitney U tests. We find that both groups show similar discernment abilities for real images and deepfakes. Also, the sociodemographics are largely similar, expect for age, where participants who dropped out were slightly younger on average than participants who returned for the follow-up (see Supplementary Table S11).

Variable	**Mean value (respondents)**	**Mean value (dropouts)**	U**-statistic**	p**-value**
Gender	0.54	0.49	123252.00	0.133
Age	43.11	38.54	105464.00	< 0.000
Ethnicity	0.57	0.56	135005.00	0.603
Education	1.08	1.13	136166.00	0.214
Religion	0.43	0.37	115084.00	0.081
PoliticalOrientation	3.95	3.90	130480.00	0.618
Income	1.31	1.30	128194.00	0.954
Real image discernment	83.26%	82.54%	152803.00	0.514
Deepfake discernment	66.66%	65.33%	152803.00	0.288

Table S11: **Differences of sociodemographic and discernment abilities for participants who completed the follow-up (respondents) and those who did not (dropouts) using Mann-Whitney U tests.**

E.10 Regression with image set covariates

To test for possible effects of the images that were used for the image discernment task, we conducted a robustness check which includes the image set as a control (as defined in Equation 3). Supplementary Table S12 shows the results for deepfake discernment. We find that, while participants were better at discerning image set B, there were no interaction effects with our interventions. This shows that the effectiveness of our interventions is not influenced by the image set. This suggests that our interventions are effective, regardless of which images participants see. This implies that our findings can be generalized to a large variety of deepfakes.

	Coef.	**s.e.**	**Lower CI**	**Upper CI**	**p-value**
Intercept	57.65	(2.260)	53.22	62.09	< 0.001
Textual	4.74	(3.188)	–1.51	10.99	0.137
Visual	13.94	(3.214)	7.63	20.25	< 0.001
Gamified	0.93	(3.196)	–5.33	7.20	0.770
Feedback	–5.09	(3.196)	–11.36	1.18	0.112
Knowledge	3.25	(3.188)	–2.99	9.51	0.307
Set	7.30	(3.188)	1.04	13.55	0.022
Textual × Set	5.50	(4.502)	–3.33	14.33	0.222
Visual × Set	–1.92	(4.533)	–10.81	6.97	0.672
Gamified × Set	7.06	(4.527)	–1.81	15.94	0.119
Feedback × Set	7.64	(4.521)	–1.22	16.51	0.091
Knowledge × Set	–0.05	(4.521)	–8.92	8.81	0.990
Observations: 1112					

Table S12: **OLS regression results for intervention effect on discernment of deepfakes while controlling for the image set.** Reported are 95% confidence intervals (CIs). Standard errors (s.e.) are reported in parentheses.

Similarly, for discernment of real images, we find that participants were better at discerning image set B, but there was no interaction effect with our interventions (see Supplementary Table S13). Hence, the effectiveness of our intervention was independent of the image set that participants were exposed to in the image discernment task.

	Coef.	**s.e.**	**Lower CI**	**Upper CI**	p**-value**
Intercept	77.74	(2.127)	73.56	81.91	< 0.001
Textual	2.22	(2.999)	–3.66	8.10	0.459
Visual	–1.77	(3.024)	–7.70	4.16	0.558
Gamified	–2.84	(3.007)	–8.75	3.05	0.344
Feedback	1.00	(3.007)	–4.89	6.90	0.739
Knowledge	1.03	(2.999)	–4.85	6.92	0.730
Set	12.38	(2.999)	6.49	18.26	< 0.001
Textual × Set	–6.33	(4.236)	–14.64	1.97	0.135
Visual × Set	–4.38	(4.265)	–12.75	3.98	0.304
Gamified × Set	0.69	(4.259)	–7.66	9.04	0.871
Feedback × Set	–1.29	(4.253)	–9.63	7.05	0.762
Knowledge × Set	–1.76	(4.253)	–10.10	6.58	0.679
Observations: 1112					

Table S13: **OLS regression results for intervention effect on discernment of real images while controlling for the image set.** Reported are 95% confidence intervals (CIs). Standard errors (s.e.) are reported in parentheses.

E.11 Regression for differences in discernment for viral and non-viral deepfakes

Deepfakes that went viral may be structurally different from those that did not. To analyze the heterogeneity of our effects across both types of deepfakes, we estimated a linear mixed effects regression as defined in Equation 4. Supplementary Table S14 shows that participants were 12.55% better at detecting viral deepfakes than non-viral deepfakes. Moreover, our *Textual* and *Gamified* interventions were less effective for viral deepfakes compared to non-viral deepfakes.

	Coef.	**s.e.**	**Lower CI**	**Upper CI**	p**-value**
Intercept	55.00	(1.957)	51.17	58.84	< 0.001
Textual	10.87	(2.764)	5.46	16.29	< 0.001
Visual	15.47	(2.783)	10.02	20.92	< 0.001
Gamified	8.02	(2.779)	2.57	13.47	0.004
Feedback	–1.91	(2.775)	–7.35	3.52	0.489
Knowledge	2.30	(2.775)	–3.13	7.74	0.406
Source	12.55	(2.126)	8.39	16.72	< 0.001
Textual × Source	–6.19	(3.003)	–12.07	–0.30	0.039
Visual × Source	–4.88	(3.023)	–10.81	1.03	0.106
Gamified × Source	–7.25	(3.019)	–13.16	–1.33	0.016
Feedback × Source	1.56	(3.015)	–4.34	7.47	0.603
Knowledge × Source	2.10	(3.015)	–3.81	8.00	0.486
Observations: 2224					

Table S14: **Linear mixed effects regression results for intervention effect on discernment of viral vs. non-viral deepfakes.** Reported are 95% confidence intervals (CIs). Standard errors (s.e.) are reported in parentheses. Standard errors are reported in parentheses.

E.12 Regression analysis for long-term effects

To test the long-term effectiveness of our interventions, we estimated a linear mixed effects regression where we model deepfake discernment of each participant at time points T1 and T2. The direct effects capture the effectiveness of our intervention at time point T1 (immediately after), while the interaction effects how each intervention's effectiveness changed over time, i.e., whether the effect differs in the follow-up. Supplementary Table S15 shows that our main intervention effects remain robust immediately after the intervention, with the *Textual* and *Visual* intervention boosting participants significantly in their discernment. Over time, the effect of both the *Textual* and *Knowledge* interventions decreased, although not significantly, while the effects of the *Visual* and *Gamified* intervention decreased significantly over time. Only the effect of the *Feedback* intervention increased over time, however, this cancels out with the negative direct effect. This confirms our results regarding Hypotheses H1 and H3 from Section 3.1.1 in that both the *Textual* and *Visual* interventions were able to boost discernment immediately after the intervention but not long-term.

	Coef.	**s.e.**	**Lower CI**	**Upper CI**	p**-value**
Intercept	61.40	(1.95)	57.58	65.21	< 0.001
Textual	7.53	(2.71)	2.21	12.84	0.006
Visual	12.17	(2.78)	6.72	17.62	< 0.001
Gamified	3.50	(2.78)	–1.94	8.94	0.207
Feedback	–1.62	(2.72)	–6.95	3.72	0.553
Knowledge	2.38	(2.77)	–3.05	7.81	0.391
Time Point	2.98	(2.19)	–1.31	7.27	0.173
Textual × Time Point	–3.45	(3.05)	–9.42	2.53	0.258
Visual × Time Point	–6.89	(3.13)	–13.02	–0.76	0.028
Gamified × Time Point	–6.35	(3.12)	–12.46	–0.23	0.042
Feedback × Time Point	1.14	(3.06)	–4.86	7.13	0.710
Knowledge × Time Point	–2.49	(3.11)	–8.59	3.61	0.424
Observations: 2224					

Table S15: **Linear mixed effects regression results for intervention effect on discernment of deepfakes over time.** Reported are 95% confidence intervals (CIs). Standard errors are reported in parentheses.

F Learning effect of image discernment task

To check for possible learning effects from the task itself, we recruited another round of $N = 200$ participants (in addition to $N = 1200$ participants from the main study), who first received an unrelated emotion discernment task at time point T1 and then the image discernment task only in the follow-up at time point T2. Here, half of the participants saw image set A for the discernment task, while the other half saw image set B. None of the participants received an intervention. We then compared the results from the discernment task of the new robustness condition with the results from the follow-up of the control conditions. For this, we use Mann-Whitney U tests for difference in means for the images sets. This means that we compare the follow-up results of the robustness condition on image set A with the follow-up results of the control condition on image set A (the same procedure is applied to image set B). By doing so, we can see whether exposure to the task itself (=the deepfake discernment) has had a significant effect on the performance of participants.

We find no statistically significant difference between the results of the control condition (average discernment of real images $\mu = 78.05\%$, average disicernment of deepfakes $\mu = 56.69\%$) and the robustness condition (average discernment of real images $\mu = 72.93\%$, average discernment of deeofakes $\mu = 53.17\%$) that saw image discernment set A (Mann-Whitney $U = 2304.5$, $p = 0.147$ for real images and Mann-Whitney $U = 2158.5$, $p = 0.491$ for deepfakes) and the control condition (average discernment of real images $\mu = 85.07\%$, average discernment of deepfakes $\mu = 69.71\%$) and robustness condition (average discernment of real images $\mu = 87.93\%$, average discernment of deepfakes $\mu = 67.59\%$) that saw image discernment set B (Mann-Whitney $U = 1853.5$, $p = 0.282$ for real images and Mann-Whitney $U = 2169.0$, $p = 0.601$ for deepfakes). Hence, mere exposure to the task did not lead to a learning effect.

G Images

Image	Prompt	Model	Date
	A photo of a couple holding hands at dinner in a restaurant	DALL-E 3	17.09.2024
	A photo of a woman on a street	DALL-E 3	17.09.2024
	A photo of two friends having coffee in Paris	DALL-E 3	17.09.2024
	A photo of a man brushing his teeth in his bathroom	DALL-E 3	17.09.2024

	https://piktochart.com/blog/viral-ai-images/	17.09.2024

Table S16: Example image sources used in the interventions *Visual*, *Gamified*, and *Feedback*.

Image	Source	Date	Image Set
	https://www.flickr.com/photos/t_riel/5675863222	16.09.2024	A
	https://www.flickr.com/photos/benedictflett/19672183508	17.09.2024	A
	https://www.flickr.com/photos/theopendoor/26187763682	16.09.2024	A
	https://www.flickr.com/photos/46924752@N03/4920054002	16.09.2024	A
	https://www.flickr.com/photos/148042613@N02/54000168419	16.09.2024	A
	https://www.flickr.com/photos/126108832@N06/48638756656	16.09.2024	B

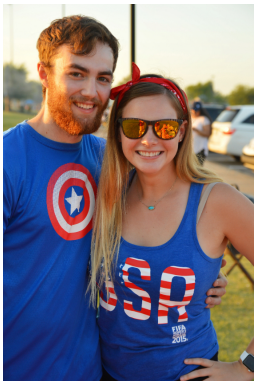
Image	Source	Date	Image Set
	https://www.flickr.com/photos/17101115@N00/2497641135	16.09.2024	B
	https://www.flickr.com/photos/radargeek/35769697725	16.09.2024	B
	https://www.flickr.com/photos/americanfarmschool/52470077293	16.09.2024	B
	https://www.flickr.com/photos/kgs/180105927	17.09.2024	B

Table S17: Real images used in the experiment.

Image	Prompt	Model	Date	Image Set
	A photo of an angry protester in a red baseball cap inside a polling place, voting booths are visible in the background	DreamStudio	22.02.2024	A
	A photo of election workers damaging the machinery of voting machines	DreamStudio	06.02.2024	A
	A photo of Mexican immigrants getting off of a bus outside a polling place	DreamStudio	06.02.2024	A
	A photo of a white man going door to door in a red baseball cap picking up mail in ballots, talking to an old lady, ballot harvesting	DreamStudio	06.02.2024	A
	A photo of Trump supporters in red baseball hats, holding guns marching towards the United states capitol	DreamStudio	06.02.2024	A

Image	Prompt	Model	Date	Image Set
	a couple holding hands	DALL-E 3	17.09.2024	B
	A photo of Mexican immigrants getting off of a bus outside a polling place	Image Creator (Bing)	06.02.2024	B
	A photo of an election worker marking ballots with pen in a mundane looking office	Image Creator (Bing)	06.02.2024	B
	a politician in front of an audience says "The sun will shine again in the Netherlands"	Midjourney	06.02.2024	B
	A photo of an election worker marking ballots with pen in a mundane looking office	DreamStudio	06.02.2024	B

Table S18: Deepfakes that are not viral we used in the experiment.

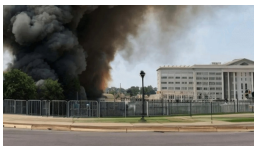
Image	Source	Date	Image Set
	https://www.washingtonpost.com/technology/2023/11/23/stock-photos-ai-images-controversy/	17.09.2024	A
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	A
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	A
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	A
	https://petapixel.com/2023/12/28/the-ai-images-that-shook-the-photography-world-in-2023/	17.09.2024	A
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	B

Image	Source	Date	Image Set
	https://www.washingtonpost.com/technology/2023/11/23/stock-photos-ai-images-controversy/	17.09.2024	B
	https://twitter.com/heyBarsee/status/1641746873210814467	17.09.2024	B
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	B
	https://piktochart.com/blog/viral-ai-images/	17.09.2024	B

Table S19: Viral deepfakes used in the experiment.