A Study of Training Strategies on Enhancing Human Detection of AI-Synthesized Faces

Ester Chen*[1], Haeseung Seo*[1], Margie Ruffin[2], Dongwon Lee[1], Gang Wang[2], Aiping Xiong[1]

[1]The Pennsylvania State University, USA
[2]University of Illinois Urbana-Champaign, USA

esterchen@mail.rit.edu, hxs378@gmail.com, mruffin2@illinois.edu, dul13@psu.edu, gangw@illinois.edu, axx29@psu.edu

Abstract

Artificial intelligence (AI) synthesized faces—so called deepfake images—have been increasingly used for malicious intent and have resulted in prominently adverse impact. Because online users must contend with discerning fake from real, great emphasis has been placed on enhancing human detection of deepfake images. We conducted an online human-subject study (N=237), investigating the effect of three training strategies (explicit training with visible *artifacts* in synthetic faces, implicit training with experiencing the *generation* of synthetic faces using real human faces, and a combination of both *artifact* and *generation*) on participants' detection of synthetic faces generated by the state-of-the-art StyleGAN techniques. Comparing participants' deepfake detection across three phases (baseline in phase 1 without any training, phase 2 after one training session, and phase 3 after the other training session), we found that all training strategies effectively enhanced participants' detection of AI-synthesized faces and their decision confidence. We also explored factors that impact participants' learning and decision-making of deepfake detection. Responses to the open-ended question revealed that participants developed generalized strategies and utilized artifacts beyond the training. Our quantitative and qualitative results provide nuanced insights into the promises and limitations of the training strategies. In addition to advancing theoretical understanding of human training in the context of deepfake image detection, our study findings hold practical implications for interface design.

Introduction

Deepfake is one kind of artificial intelligence (AI)-generated synthesized content. It intentionally manipulates individuals' articulations and actions, and human facial information, resulting in photorealistic faces of non-existing humans or vivid depictions of a person saying/doing something they did not really do (Karras et al. 2020; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Van Den Oord et al. 2016). While the use of photorealistic AI-generated images started in the realm of art and entertainment, such as (re)creating characters or scenarios (Shah 2018), its adverse influence has become increasingly prominent (Atleson 2023).

The prevalence of AI-synthesized faces (also called *deepfake images*), which this work focuses on, has been increasingly evident among various types of fraud globally (Zholudev and Kalaydin 2023). The malicious use of synthetic faces has been reported or demonstrated, from non-consensual sexual imagery being weaponized against (primarily) women (Saner 2024) and other marginalized populations (Maung et al. 2024) to the deceptive persona on social media platforms for small- and large-scale social engineering attacks and disinformation campaigns (Brooks et al. 2018; Diakopoulos and Johnson 2021). For instance, using AI-synthesized faces, deepfake profiles were created on social media platforms, posing as journalists or military personnel during the Ukraine-Russian war to spread misinformation (Chadwick 2022).

To mitigate the negative influence of deepfake images, researchers have developed automatic detection techniques to detect AI-synthesized faces (Cozzolino et al. 2018; Li et al. 2020; Marra et al. 2018). However, due to the evolving techniques of facial synthesis and generative AI, these state-of-the-art detection techniques will become less effective over time (Farid 2022). As a result, great emphasis has also been placed on user-centered solutions, because online users must contend with distinguishing between real content and fake content (Hancock and Bailenson 2021).

Compared to false textual claims, fake content in richer modalities such as images has caused participants to perceive fake content as more credible (Garry and Wade 2005). Existing work has also shown that deepfake-enabled social media profiles increased participants' perceived accuracy of fake claims (Ruffin et al. 2024). While AI-synthesized faces are highly photorealistic, nearly indistinguishable from real human faces (Lago et al. 2021), and perceived as more trustworthy than real human faces (Nightingale and Farid 2022), those images often contain human-detectable artifacts.

Human detection of AI-synthesized faces can be affected by various factors, such as training, feedback, and AI-related experience (Liu, Qi, and Torr 2020; Nightingale and Farid 2022; Hulzebosch, Ibrahimi, and Worring 2020). Those initial efforts have provided a preliminary understanding of how humans develop resilience to deepfake images. However, given the prominently negative impact of misusing AI-synthesized faces, it is critical to further explore effective training strategies for enhancing the human detection.

*These authors contributed equally to this work.

Beyond *explicit* training such as a short tutorial describing specific artifacts to identify AI-synthesized faces by Nightingale and Farid (2022), we examine the effectiveness of *implicit* training informed by the literature on skilled acquisition and training (Johnson and Proctor 2016). Specifically, the implicit training provides a short experience about how AI-synthesized faces are generated from real human faces, mimicking in everyday life that online users who view both deepfake and real images can learn the detection through the opportunities of fake and real comparisons. Moreover, prior studies have primarily relied on quantitative measures (e.g., 2-alternative forced choice of image credibility and scale rating of trustworthiness), but have not explored factors impacting humans' learning and decision-making with regard to deepfake detection.

It is now more critical than ever to enable online users to stay vigilant and fortify their defense against AI-synthesized faces and downstream attacks such as in disinformation campaigns. We design an online experiment to investigate different training strategies (i.e., implicit, explicit, and implicit and explicit) on enhancing human's ability to discern AI-synthesized faces from real human faces. We leverage two StyleGAN techniques (Karras et al. 2020, 2021) to generate stimuli of AI-synthesized faces used in our study. We aim to answer the following research questions (**RQs**):

- **RQ1.** Compared to implicit training (generation), is explicit training (artifact) more effective in helping participants distinguish state-of-the-art GAN-synthesized faces from real human faces?
- **RQ2.** Do participants who have both implicit and explicit training (artifact and generation) further increase their distinguishability between AI-synthesized faces and real human faces, compared to those who only have the implicit or explicit training?
- **RQ3.** Do participants continue increasing their distinguishability between AI-synthesized faces and real human faces if extra training is provided?
- **RQ4.** What are the primary factors impacting participants' learning and decision-making regarding the deepfake detection?

To address the **RQs**, we conducted an online human-subject study (N=237). In the experiment, we manipulated the training strategy as a between-subjects factor at three conditions [explicit training with *artifact*, implicit training with *generation*, explicit and implicit training (i.e., *artifact* and *generation*)]. In each training condition, we measured participants' performance of detecting real human faces and the state-of-the-art StyleGAN synthesized faces (GAN2 and GAN3) at three phases (i.e., baseline at *phase 1* without any training, *phase 2* after one training session, *phase 3* after the other training session). We also asked participants to describe what they have learned throughout the training sessions. Our main findings are the following.

First, we found that all training strategies were effective in enhancing participants' detection of AI-synthesized faces, as well as their detection confidence. *Second*, we obtained the benefits of combining the two training strategies in participants' continuously increased detection confidence and its promise to enhance participant's better detection of AI-synthesized faces devoid of obvious artifacts. *Third*, compared to the other conditions, the extra training in the artifact condition tended to deteriorate the participants' detection accuracy on real human faces. Such findings suggest that obtaining ground truths for real human faces as in the other conditions can help participants mitigate decision bias (i.e., a tendency to judge all images as AI-synthesized). *Moreover*, we found that participants developed generalized strategies and used artifacts beyond the training for AI-synthesized face detection. Altogether, the results of current work advance our understanding of human deepfake detection training. We discuss the theoretical and practical implications of our findings for detecting AI-synthesized faces.

Background and Related Work

AI-Synthesized Faces. Generative adversarial networks (GANs) are one of the most successful artificial intelligence (AI) techniques for synthesizing content, such as images of fictional persons (Goodfellow et al. 2014). To generate a synthetic face, two neural networks in each GAN system—a generator and a discriminator—compete one and the other in an adversarial game. The generator starts by randomly seeding an array of pixels and then iteratively updates its guess through the discriminator's feedback. In each round, if the discriminator equipped with a large database of real faces can distinguish the synthetic face from real human faces, the discriminator penalizes the generator. This process continues until the generator produces a synthetic face that the discriminator cannot distinguish from real human faces. The StyleGAN family (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021) is considered the state-of-the-art technique in synthetic faces for GANs.

Human Detection of AI-Synthesized Faces. Researchers have examined whether humans can distinguish AI-synthesized faces from real human faces. While synthetically generated faces were nearly indistinguishable from real human faces for online participants (Hulzebosch, Ibrahimi, and Worring 2020), human's detection of AI-synthesized faces can be enhanced when receiving immediate feedback (Hulzebosch, Ibrahimi, and Worring 2020), with AI guidance (Boyd et al. 2023), or after viewing many synthetic examples (Liu, Qi, and Torr 2020). Individuals' AI-related experience also has impact on their detection performance of synthetic faces (Hulzebosch, Ibrahimi, and Worring 2020).

Nightingale and Farid (2022) conducted a series of experiments examining participants' susceptibility to GAN-synthesized faces. In their Study 2, they asked untrained and trained participants to classify a face image as either real or synthetic. Each participant was tested with 128 faces (half real), which were randomly sampled from a dataset of 400 real and 400 StyleGAN2 (Karras et al. 2020) synthetic faces. After each response, performance feedback was also provided. Critically, before the test, examples of artifacts that can be used to identify AI-synthesized faces were provided to the trained participants but not the untrained participants. Their results showed that the untrained participants had difficulty in detecting AI-synthesized faces, with about half AI-synthesized faces labeled incorrectly. However, the perfor-

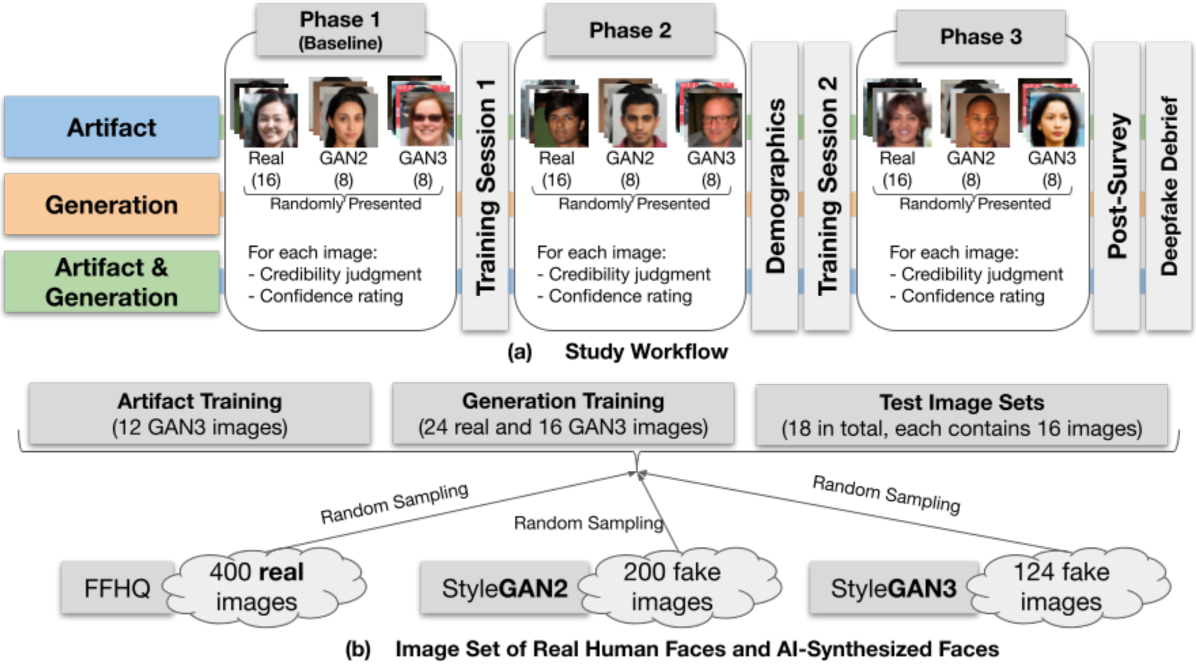


Figure 1: **An Overview of the Experiment Design.** Figure 1(a) illustrates the study workflow. Participants in each condition evaluated a set of 32 images (16 real, 8 StyleGAN3, 8 StyleGAN2) in each phase. Phase 1 served as a baseline. Depending on the condition, in each training session, participants received a short tutorial describing examples of artifacts to identify synthetic faces (i.e., the *artifact* training), experienced how synthetic faces are generated using real human faces (i.e., the *generation* training), or received both training. Figure 1(b) depicts the images used in the study.

mance of trained participants was about 10% higher, indicating the effectiveness of a short training with artifacts. No performance improvement was obtained over time, suggesting the limited impact of the trial-by-trial feedback.

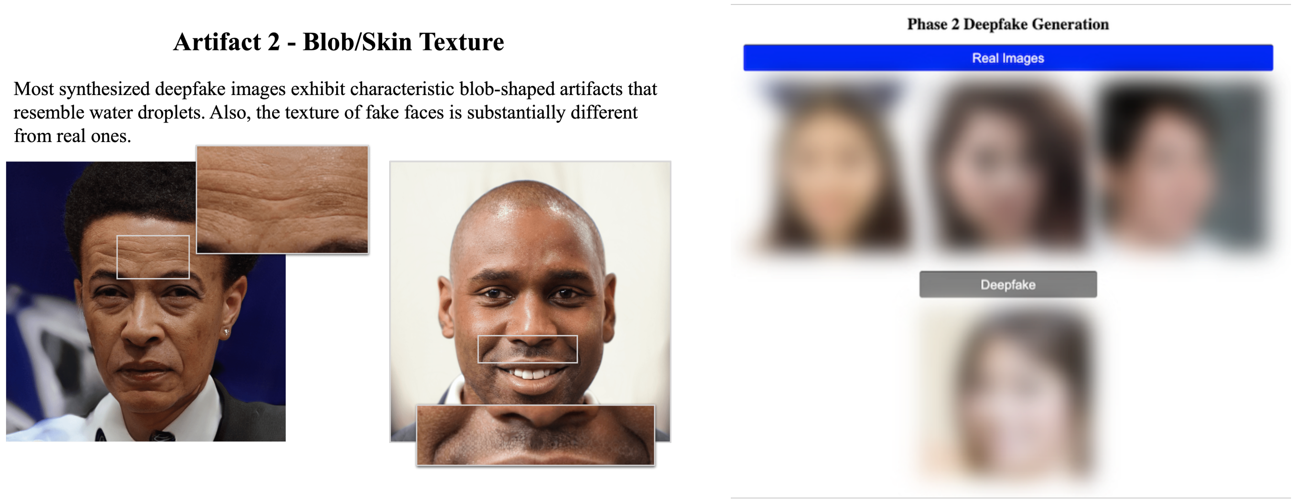
Building on the prior work, our study examines effect of training on detecting synthesized faces generated by both StyleGAN2 and StyleGAN3 techniques. Compared to StyleGAN2, StyleGAN3 is able to generate images that rotate and translate smoothly and without texture sticking.

Factors Affecting Learning and Skill Acquisition. Given the possibly substantial impact of deepfake images on the individual level (e.g., non-consensual sexual images) and the societal level (e.g., amplifying disinformation campaigns), it is essential to understanding processes that can lead to enhanced human perception and detection of AI-synthesized faces. Perceptual skill has been defined as “the enhanced ability to discriminate between and to classify stimuli based on perceivable properties”(Johnson and Proctor 2016, p.29). In contrast to learning through *explicit* training (e.g., artifacts evaluated by Nightingale and Farid, 2022), *implicit* training refers to learning without a specific objective or without paying attention to the relevant information. Thus, an investigation on the implicit training could reveal the role of *experience* in perceptual learning. Such an investigation is critical for deepfake image detection because online users typically have few opportunities to learn specific artifacts to detect AI-synthesized faces in the wild. We propose to examine whether a short experience about how AI-synthesized faces are generated based on real human faces affects human detection of synthetic faces.

Repeated execution of a task is one key factor impacting human skill acquisition (Proctor and Van Zandt 2008). Because deepfake detection is not easy to learn all at once, training may become more effective when it is repeated. To the best of our knowledge, previous work has not examined such a factor. Our study aims to fill the gap. To understand possible interaction of explicit and implicit training, we also consider a condition with both explicit and implicit training.

Method

We conducted an online study investigating different training strategies on enhancing human detection on AI-synthesized faces. As shown in Figure 1, we evaluated explicit training (e.g., a short tutorial describing examples of different artifacts to identify synthetic faces), implicit training (e.g., an interactive experience of how AI-synthesized faces are generated from real human faces), and a combination of both, using a between-subjects design (**RQ1** and **RQ2**). We also examined whether participants continued enhancing their distinguishability between the synthetic faces and real human faces when extra training is provided across phases (**RQ3**). Moreover, we explored factors that impact participants' learning and decision-making regarding the detection of AI-synthesized faces (**RQ4**).


Artifact 2 - Blob/Skin Texture

Most synthesized deepfake images exhibit characteristic blob-shaped artifacts that resemble water droplets. Also, the texture of fake faces is substantially different from real ones.

Figure 2: Illustrations of an artifact (i.e., Blob/Skin Texture) explained in the artifact (explicit) training condition (left panel) and how a synthetic face is generated by using real human faces in the generation (implicit) training condition (right panel).

Materials

Image Stimuli. Following previous work (Lago et al. 2021; Nightingale and Farid 2022), we established an image pool comprising 400 real images, 200 images synthesized by StyleGAN2 (GAN2), and 124 images synthesized by StyleGAN3 (GAN3) for the study [see Figure 1(b)]. There is an equal distribution of race and gender for each image type. The real images are the same as Nightingale and Farid (2022). We randomly sampled GAN2 images from their synthetic dataset. We created the GAN3 image set as follows.

GAN3 Image Generation. Using an NVIDIA lab GAN3 pre-trained model (NVIDIA 2022), we generated $5,000$ synthetic faces initially. When generating images from the pre-trained models, one can control the quality of image production in different ways. For example, the pre-trained models for GAN3 have either configuration T (translation equiv.) or configuration R (translation and rotation equiv.). We opted to use the model with configuration R because it helped to control for the texture-sticking artifacts noticeable in GAN2. We also chose to generate aligned images because they most closely resembled the images in our GAN2 pool. The pre-trained model we used was also trained on the Flicker-Face-HQ (FFHQ) dataset, which was originally created as a benchmark for GANS (Goodfellow et al. 2014). It is the same dataset used to train GAN2, which produced images used by Nightingale and Farid (2022).

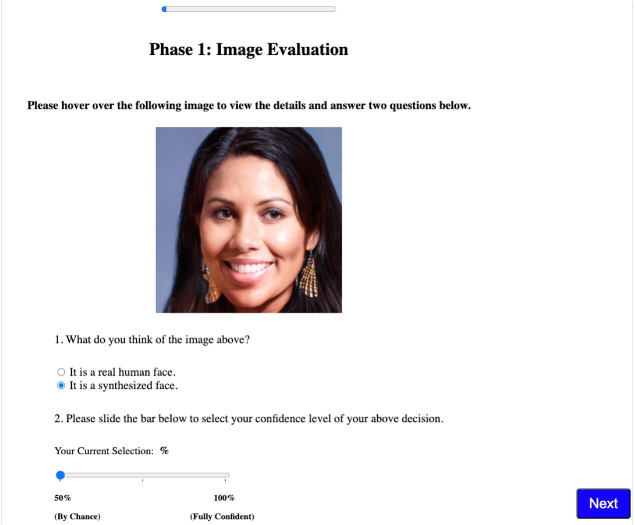
GAN3 Image Evaluation. After creating those images, we underwent a meticulous process to select experimentally suitable GAN3 images based on both internal and external evaluations.

Internally, three authors assessed and labeled the images for outwardly visible demographic information and for quality. For demographic characteristics, we verified race, gender, and age. Race and gender categories adhered to Nightingale and Farid (2022)'s classification: race (White/Caucasian, Black/African American, East Asian, South Asian) and gender (Male, Female). Age labeling was performed to exclude images of children in order to prevent potential infringements on child portrait rights. For the quality evaluation, we categorized image quality into high and low, and subsequently excluded images that were of low quality. "Low quality" is characterized by significant artifacts or multiple minor distortions that are easily noticeable upon overall inspection. The artifacts include pronounced facial or background distortions, images that do not convincingly depict a person, excessively distorted proportions, non-human like skin patterns, distorted individuals in the surroundings, irregular background contours, and instances of mismatched subjects and backgrounds. Our evaluation ignored intricate details such as pupils, hair textures, and teeth alignment, which typically require scrupulous scrutiny.

Among the $5,000$ images, at least two out of three authors annotated a total of 510 images as "high" for quality. Considering the two-gender and four-race categories, we counted these images and found that the "black female" category had the fewest instances, totaling 22 images. To align with this count, a random selection of 22 images was performed for each remaining category.

Externally, the 176 images that were internally selected by the authors underwent a labeling process for race and gender by diverse Prolific workers, following an Institutional Review Board (IRB) approval from the authors' institution. A total of 11 sets of images were created such that a single worker would label two images for each combination of demographic characteristics. We recruited 70 Prolific workers such that each image was evaluated by different workers at least five times. As a result, out of the 176 labeled images, we excluded two images labeled as "Other," which we had provided as a labeling choice alongside the existing categories to the Prolific workers, along with ten images labeled with various races in equal proportions, and three images labeled with different gender in equal proportions.

Considering both internal and external evaluations, we further excluded 14 images from the remaining 161 images



Phase 1: Image Evaluation

Please hover over the following image to view the details and answer two questions below.

1. What do you think of the image above?

○ It is a real human face.
● It is a synthesized face.

2. Please slide the bar below to select your confidence level of your above decision.

Your Current Selection: %

Figure 3: A screenshot of our online study interface.

due to disagreements between Prolific workers and three authors. Ultimately, after this step, we examined the distribution of race and gender among the remaining 147 images. Taking the relatively modest count of 12 "Female South Asian" images as a reference, we randomly selected 12 images from each of the remaining categories and curated a total of 96 images for the study. Because GAN2 images of Nightingale and Farid (2022) were without any obvious artifacts, we created our training materials using another 28 GAN3 images (12 for the tutorial and 16 for the generation).

Image Set for Evaluation. In the experiment, each participant underwent 32 rounds of image evaluation for every phase, with 16 real images and 16 synthetic images (half GAN2 and the remaining half GAN3). A total of 18 sets of images were created by sampling the image pool (9 for real images and 9 for synthetic images). Each set contained 16 images, with factors such as image credibility (real or GAN2 & GAN3) and image categories (two genders and four races) being considered during the assembly. To make the experimental design more efficient, instead of fully randomization, we applied a Latin square design to distribute the image sets among the participants. Two synthesized images from Nightingale, Wade, and Watson (2017) were used for two attention check questions. See details of the image set in our supplementary materials (Chen et al. 2025).

Training Design

Explicit Training (Artifact). Informed by Nightingale and Farid (2022), we designed a short tutorial of three specific artifacts of deepfake images. The three artifacts are accessories (e.g., eyeglasses and earrings), blob/skin texture, and background/boundary. In each tutorial page, a succinct description of the artifact is presented along with two illustrative examples. In each example image, we enlarge the part(s) showing the corresponding artifact(s) of deepfake (see Figure 2). Considering the two training sessions, we created two tutorial pages for each artifact, using different illustrative examples. We expected that participants would learn these *explicit* cues (i.e., artifacts) to identify AI-synthesized faces throughout the training.

Implicit Training (Generation). In this condition, we offered participants an interactive experience to understand the relationship between real images and AI-synthesized images (see Figure 2). To generate a synthetic face, participants first press a "Real Images" button to view the real human faces that are potential sources of synthesized "deepfake" images. Specifically, three real images are randomly assigned, which consist of an image with the same gender as the synthesized image, an image with the same race, and an image with both identical gender and race. By pressing a "Deepfake" button below the real images, participants are able to view an example of the generated synthetic face, which incorporates the race and gender features of the real images (see Figure 2).

Considering the combination of two genders and four races, we provided eight such opportunities in each training session. For each combination, we presented different synthetic faces across the two training sessions. Throughout the generation experience, we aimed to foster participants' comprehension of how deepfake images are synthesized in an intuitive manner. We also expected that participants would detect the differences between real and synthetic faces *implicitly*.

Explicit and Implicit Training (Artifact and Generation). A combination of the aforementioned training was also examined. Specifically, we first presented the artifact training and then the generation training. Considering the complementary roles of the artifacts (i.e., what to detect) and the generation (i.e., how to detect), we expected an additive effect: participants in the artifact and generation condition would further enhance their performance on detecting AI-synthesized faces compared to those in the explicit condition or the implicit condition.

Study Interface. To run our online study, we designed a web-based application utilizing the PyFlask web framework, which offered us great flexibility to design the different training sessions. Considering the impact of image resolution (Hulzebosch, Ibrahimi, and Worring 2020), we presented each image with a resolution of 1024 × 1024. We also prevented participants using mobile devices from accessing our study. We presented one image (real or synthetic) in each evaluation page (see Figure 3). Moreover, we implemented and provided a hover-over feature directly on each image. Such a feature scaled up the image by a factor of 2.7, which enabled participants to closely examine the image details and search artifacts described in the conditions with the artifact training. Below each image, the credibility judgment and confidence rating questions were shown. To help participants better understand the study progress, we also showed the study phase and a progress bar on the interface. After completing both questions, participants pressed the "Next" button to proceed.

Procedure

After informed consent, the participants were randomly assigned to one of the three training conditions. In phase 1, the participants first evaluated 16 synthetic faces (half GAN2

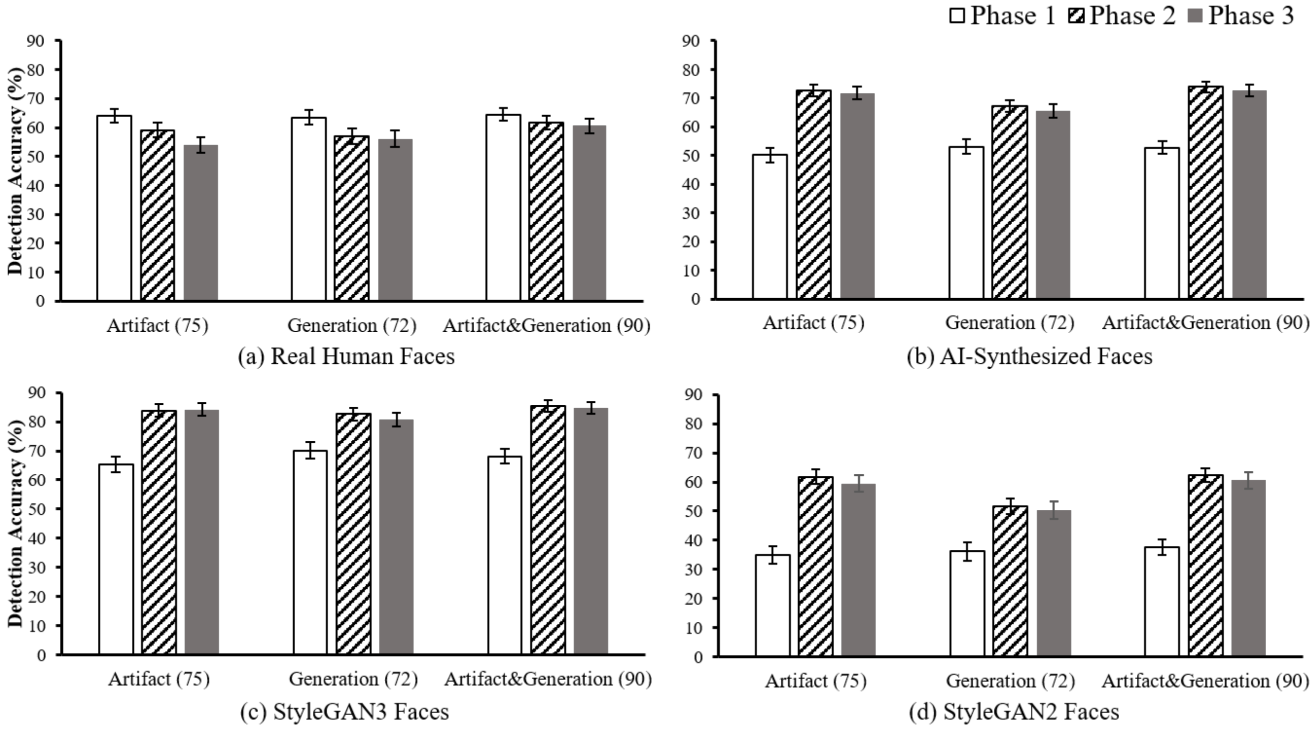


Figure 4: **Detection Accuracy**. Results of real human faces (a), AI-synthesized faces (b), StyleGAN3 faces (c), and StyleGAN2 faces (d) as a function of **phase** (1,2,3) and **condition** (artifact, generation, artifact & generation). Numbers in the parentheses indicate the number of participants at each condition. Error bars represent ± one standard error.

and the remaining half GAN3) and 16 real human faces presented in randomized order. We asked two questions for each image. Participants first answered the question of "What do you think of the image above?" with two options ("It is a real human face." and "It is a synthesized face."). Then, they were promoted to slide a bar to indicate their confidence level in answering the first question. The range of the slider was presented from 50%, indicating the minimum confidence level to make a decision (Singh et al. 2019), to 100%. Below each number, "By Chance" and "Fully Confident" were also presented (see Figure 3). Without any intervention at phase 1, participants' performance could reveal their baseline detection of AI-synthesized faces.

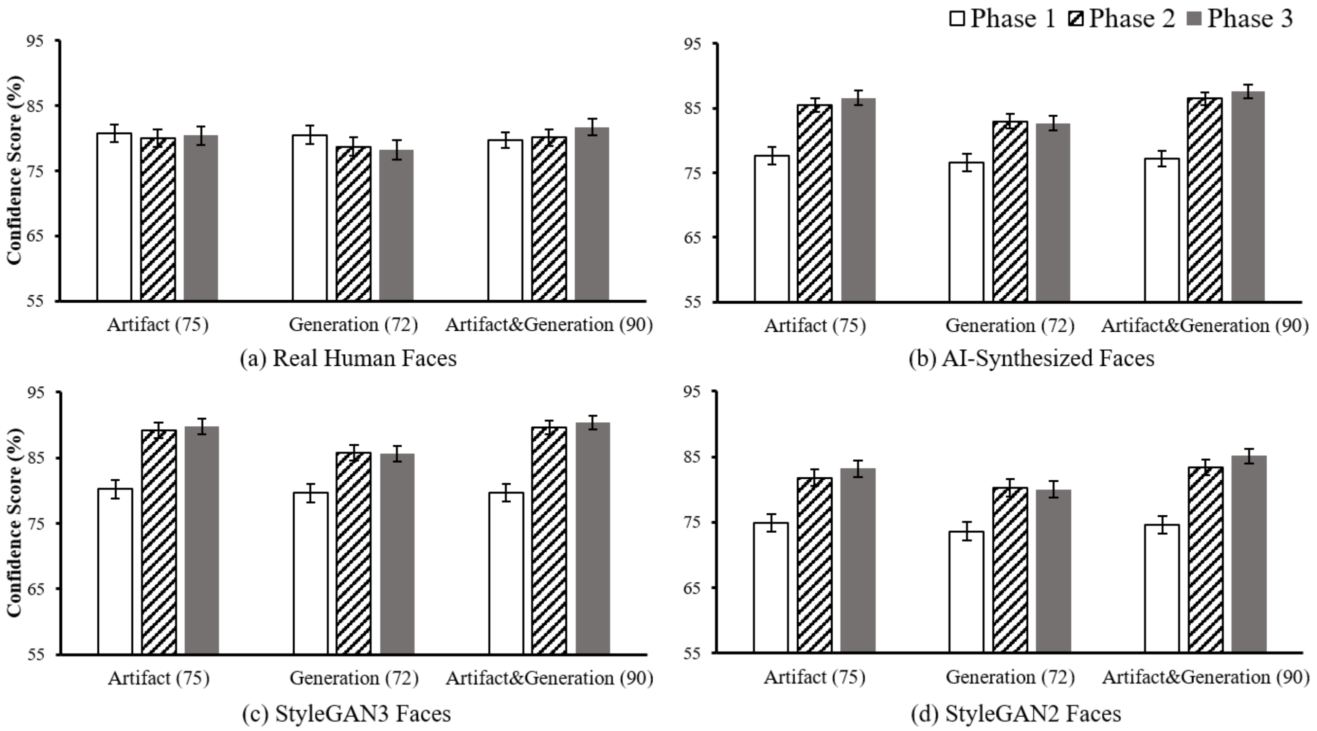
After the first training session in each condition, participants evaluated another set of 16 synthetic faces and 16 real human faces in phase 2. After the second training session, participants in each condition evaluated another set of 32 images in phase 3. The procedure of each phase was the same. We also presented one attention check question at phases 1 and 2, respectively. We clearly described the attention check before the study started and the question resembled the image-credibility evaluation in each phase. For participants who failed both attention check questions, we excluded their responses from the data analysis (Prolific 2024).

After phase 2, participants answered demographic questions such as age, gender, and education level. We also asked about their experience with deepfakes, Photoshop, and other image editing tools. In the post survey after phase 3, participants described what they had learned in the study. At the end, participants were also provided with information about the deepfake images used in the study.

Participants

We recruited participants aged 18 and above who are residents of the U.S. through Prolific. We collected complete responses from a total of 245 participants.* Among them, we filtered out four respondents who failed both attention check questions. Additionally, we excluded two participants who consistently selected 100% confidence for all decisions, as well as two participants with task completion time falling below 25% of the median duration (about 39 min for the tutorial condition). As a result, responses of 237 participants

*Power analysis using G*Power 3.1 (Faul et al. 2007) indicated $n = 204$ participants to detect a small effect size (Cohen's $f = 0.1$) of the interaction of training strategies and phases, with a power of .80 [analysis of variance (ANOVA) test], $\alpha = .05$. To account for potential submission removals while ensuring the statistical power, we recruited 346 participants initially. There were 101 incomplete responses during the data collection. We contacted those Prolific workers. Most of them indicated extended image loading, possibly due to network speed issue and use of VPN, which might have hindered their proper responses.



(a) Real Human Faces

(b) AI-Synthesized Faces

(c) StyleGAN3 Faces

(d) StyleGAN2 Faces

Figure 5: **Detection Confidence**. Results of real human faces (a), AI-synthesized faces (b), StyleGAN3 faces (c), and StyleGAN2 faces (d) as a function of **phase** (1,2,3) and **condition** (artifact, generation, artifact & generation). Numbers in the parentheses indicate the number of participants at each condition. Error bars represent ± one standard error.

were included for data analysis. Participants were paid based on an hourly rate of \$8, following the minimum wage under Prolific's payment principles.[*]

Among the 237 participants, there were 111 females (46.8%), 120 males (50.6%), and six others (2.5%). In terms of race, White Caucasians constituted the largest group with 149 individuals (62.9%), followed by 33 Black/African Americans (13.9%), 25 Hispanic/Latino (10.5%), and 17 Asians (7.2%). Regarding education level, most participants (82, 34.6%) had a bachelor's degree, followed by 59 participants (24.9%) with some college credit, 42 with a Master's degree (17.7%), and 35 with a high school education (14.8%). In terms of age groups, the $30-39$ age range was the most prevalent with 75 individuals, followed by $18-29$ years with 62 participants (26.2%) and $40-49$ years with 50 participants (21.1%).

Results

Shown in Figures 4 and 5 are descriptive statistics of participants' detection accuracy and confidence score. To quantify the effect of training, detection accuracy and confidence score were entered into 2 (*image credibility*: real, synthetic) × 3 (*phase*: 1, 2, 3) × 3 (*condition*: artifact, generation, artifact and generation) mixed analysis of variances (ANOVAs) with a significance level of .05, respectively. Post-hoc tests with Bonferroni correction were performed. The number of participants included for data analysis in each condition is as follows: 75 (artifact), 72 (generation), and 90 (artifact and generation).[*]

Detection Accuracy

Synthetic vs. Real across Phases. As shown in Figures 4(a) and 4(b), participants were generally better at detecting synthetic faces (64.4%) than real human faces (60.0%, $F_{(1,234)} = 7.65, p = .006, \eta_p^2 = .032$). Moreover, such a detection gap varied across phases ($F_{(2,468)} = 90.27, p < .001, \eta_p^2 = .278$). Critically, in phase 1, without any intervention, participants' baseline detection of synthetic images (52.0%) was worse than that of real images ($64\%, p < .001$). However, such a pattern was reversed at phase 2 (synthetic: 71.3% vs. real: 59.2%, $p < .001$) and phase 3 (synthetic: 70.0% vs. real: 56.8%, $p < .001$), indicating the effect of training.

Participants' average detection accuracy also varied across phases (phase 1: 58.0%, phase 2: 65.3%, phase 3: 63.4%, $F_{(2,468)} = 44.25, p < .001, \eta_p^2 = .159$). Post-hoc pairwise comparisons showed that the differences across

[*] Prolific's payment principles: https://researcher-help.prolific.com/en/article/2273bd

[*] Due to the exclusion was not evenly distributed across conditions, the resulting number of participants in each condition was similar but not the same.

phases were all significant ($ps \leq .035$). In particular, contrary to our prediction, the average detection accuracy slightly decreased from phase 2 to phase 3 ($p = .035$). Moreover, the main effect of *phase* was qualified by the two-way interaction of *phase* × *condition* ($F_{(4,468)} = 3.09, p = .047, \eta_{\text{p}}^2 = .026$). Post-hoc pairwise comparisons showed that the decreased accuracy from phase 2 and phase 3 was mainly contributed by the artifact condition ($p = .055$) but not the other two conditions with generation ($p \geq .335$).

Such results were consistent with a pattern shown in Figure 4(a) for the real image evaluation: from phase 2 to phase 3, there was a trend of more decrease for detection accuracy in the artifact condition (59.1% to 54%) than the other conditions (generation: 56.9% to 56%, artifact and generation: 61.7% to 60.5%). To gain a better understanding of the artifact training, we conducted further investigation in Exploratory Analysis. No other effect involving condition was significant ($F_S \leq 2.18$).

GAN2 vs. GAN3. To further understand the training effect, we analyzed whether there was any detection difference between GAN2 and GAN3 images. Detection accuracy results were entered into 2 (*style*: GAN2, GAN3) × 3 (*phase*: 1, 2, 3) × 3 (*condition*: artifact, generation, artifact and generation) ANOVA.

As shown in Figures 4(c) and 4(d), there was a noticeable difference between the correct detection of GAN3 (78.3%) and GAN2 (50.5%, $F_{(1,234)} = 631.76, p < .001, \eta_{\text{p}}^2 = .730$). Such a result was not surprising, because 1) all the synthetic images in the training interventions were GAN3 and 2) GAN2 images of Nightingale and Farid (2022) devoid of obvious rendering artifacts were hard to detect.

The interaction of *style* × *phase* was also significant ($F_{(2,468)} = 5.47, p = .020, \eta_{\text{p}}^2 = .023$). Specifically, the initial detection accuracy gap between GAN3 and GAN2 in phase 1 (31.7%) was reduced in phase 2 (25.3%, $p < .001$) and phase 3 (26.4%, $p < .001$), indicating the effect of training for both styles. Nevertheless, the detection differences of phases 2 and 3 showed no statistical significance ($p = .865$), suggesting limited effect of training repetition.

Moreover, the two-way interaction of *style* × *condition* ($F_{(2,468)} = 3.11, p = .047, \eta_{\text{p}}^2 = .026$) was significant. While the detection of GAN3 showed no difference across conditions ($p \geq .999$), the detection of GAN2 in the generation condition (46.1%) was similar to that of the artifact condition (52.1%, $p = .193$), but worse than that of the artifact and generation condition (53.5%, $p = .050$). Because baseline (phase 1) performance is similar between the two conditions (generation: 36.1%, artifact and generation conditions: 37.6%), the discernible difference suggests the additive effect of the explicit and implicit training on detecting synthetic faces. No other effect was significant ($F_s \leq 1.62$).

Detection Confidence

Participants were confident about their deepfake detection in general (see Figure 5). ANOVA results were in line with what we observed in the detection accuracy analysis.

Synthetic vs. Real across Phases. As shown in Figures 5(a) and 5(b)[*], participants showed higher confidence in their decisions of AI-synthesized faces (82.6%) than the decisions about real human faces (80.0%, $F_{(1,234)} = 42.83, p < .001, \eta_{\text{p}}^2 = .155$). The main effect of *phase* ($F_{(2,468)} = 55.82, p < .001, \eta_{\text{p}}^2 = .193$) was qualified by its interaction with *image credibility* ($F_{(2,468)} = 138.10, p < .001, \eta_{\text{p}}^2 = .371$), revealing the effect of training. Specifically, participants' confidence in evaluating synthetic faces at phase 1 (77.1%) was increased in phase 2 (85.0%, $p < .001$) and maintained at phase 3 (85.7%, $p < .001$). However, there was no significant difference for the decision confidence of real human faces (phase 1: 80.4%, phase 2: 79.6%, phase 3: 80.1%, $p_s \geq 352$).

The effect of *phase* also varied across conditions ($F_{(4,468)} = 4.41, p = .013, \eta_{\text{p}}^2 = .036$). While participants in the artifact condition increased their confidence from phase 1 (79.2%) to phase 2 (82.8%, $p < .001$) and maintained the confidence level at phase 3 (83.4%, $p < .001$), participants in the generation condition showed similar confidence throughout the study (phase 1: 78.6%, phase 2: 80.9%, phase 3: 80.5%, $p_s \geq .138$). In the case of the artifact and generation condition, significant differences were observed across all conditions (phase 1: 78.4%, phase 2: 83.3%, phase 3: 84.6%, $ps \leq .047$), implying the effectiveness of the explicit and implicit training.

GAN2 vs. GAN3. We conducted the same analysis as the accuracy measure. Only the main effect of *style* was significant ($F_{(1,234)} = 182.77, p < .001, \eta_{\text{p}}^2 = .439$). In agreement with the detection accuracy results, participants were more confident in detecting GAN3 images (85.5%) than GAN2 images [79.7%, see Figures 5(c) and 5(d)]. No other terms involving style were significant or approached the significant level ($Fs < 1.0$).

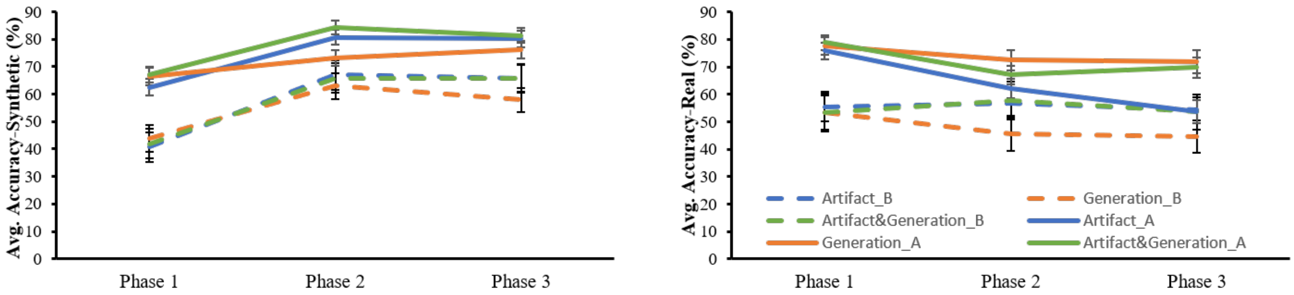
What Did Participants Learn from the Training?

To understand factors that impact the learning of AI-synthesized face detection, we asked participants in each condition what they learned from training. In total, we obtained 236 responses. We performed a thematic analysis (Braun and Clarke 2006) of all responses based on the training condition. Two coders independently coded the responses and three themes were identified.

Trained Artifacts. One hundred and sixty-eight (71.2%) participants described at least one trained artifact in their responses. The recall rate was 84.0% for the artifact condition and 73.3% for the artifact and generation condition. Interestingly, 39 (54.1%) participants in the generation condition also mentioned at least one of the trained artifacts. Thus, without explicit training, participants could detect those artifacts through seeing examples of real human faces and synthetic faces. Those results suggest that trained artifacts are easy to detect and recall.

Artifacts Beyond Training. Twenty-four (10.2%) participants detected AI-synthesized faces using artifacts beyond

[*] In phase 1, some participants failed to detect any GAN2 images. We replaced missing confidence of these participants with the average confidence of all participants.



(a) AI-Synthesized Faces

(b) Real Human Faces

Figure 6: **Exploratory Analysis using Baseline Performance**. Results of AI-synthesized faces (a) and real human faces (b) as a function of **phase** (1,2,3) and **condition** (artifact, generation, artifact & generation). **Dash lines** and **solid lines** show the results of participants whose baseline (phase 1) performance was **below** (B) or **above** (A) the overall median, respectively. Error bars represent ± one standard error.

the training (e.g., "... *Also, the teeth were a pretty big indicator.* (P118)" and "... *They might have an extra tooth in their mouth.* (P140)"). Instead of focusing on specific artifacts, 34 (14.4%) participants answered that they learned to pay attention to the details in general after the training. For example, P92 described, "*To really pay attention to details,*" and P42 responded "*Zoom in and see if there are things that are not normal ...*" The number of participants was similar across conditions.

Hard to Detect Deepfake. Twelve (5.1%) participants responded that deepfake detection was difficult. For example, P114 answered "*Some [deepfake images] are incredibly difficult to spot,*" corresponding to GAN2 images used in our study. Most of the participants (9) were from the generation condition (e.g., "*I had a very hard time telling the difference even after seeing the originals and the deepfakes.* (P131)"). Yet, no participants in the artifact and generation condition mentioned the detection difficulty. Such results highlight that an explicit training seems to be necessary, at least for some participants.

Overall, the qualitative analysis suggests that explicit cues (e.g., artifacts) are essential to help participants detect AI-synthesized faces. However, given the difficulty of detecting certain deepfakes (e.g., GAN2 images in our study), relying on a single intervention might not be sufficient.

Exploratory Analysis

We also performed exploratory analysis to understand whether the characteristics of AI-synthesized faces and the individual differences of the participants had influenced the detection accuracy of deepfake.

Effect of Image Features. We first explored the impact of the racial and gender features of the utilized images on deepfake detection. By adding those factors to the ANOVA, we found that the detection precision of white male synthetic faces (55.8%) was significantly lower compared to other image categories (average detection: 65.9%, $ps < .001$). Moreover, white female synthetic faces showed the second lowest detection accuracy (62.5%). Those results not only replicate what Nightingale and Farid (2022) obtained but also extend the findings to StyleGAN3 images.

Effect of Prior Experience in Photo Editing and Deepfakes. We asked participants whether they had encountered deepfake images or had experience in using photo editing tools before the study. Among the 146 participants (61.6% of 237) who had seen deepfakes before, most reported encountering deepfake images on social media (102) and news websites (23). Participants with at least some experience with photo editing tools constituted 53.6% (127). After adding each factor in the ANOVA, we observed that participants who have seen deepfakes demonstrated better performance (64.8%) than those without experience (58.1%, $F_{(1,232)} = 17.51, p < .001, \eta_{\mathrm{p}}^2 = .070$). Participants' previous experience with photo editing tools did not show any statistical significance ($Fs \leq 2.87$).

Effect of Baseline Performance. Given that the aforementioned impact of individual differences was based on self-reported data from the participants, which may have been influenced by the social desirability effect (Dodou and de Winter 2014), we performed an additional exploratory analysis using the baseline performance of the participants in phase 1. Specifically, we categorized the participants based on whether their detection performance in all images exceeded the median value in phase 1. There were 101 participants whose performance was above the median and 136 participants whose performance was at or below the median.

We added the factor in the ANOVAs of real and synthetic images, respectively. For synthetic faces, the main effect of *baseline performance* (below: 56.9%, above: 74.6%, $F_{(1,232)} = 102.45, p < .001, \eta_{\mathrm{p}}^2 = .307$) and its interaction with *phase* were both significant ($F_{(2,464)} = 7.62, p = .006, \eta_{\mathrm{p}}^2 = .032$). As shown in Figure 6(a), such results suggest that the initial gap between the two groups (below: 42.2%, above: 65.3%) was reduced in phase 2 (below: 65.3%, above: 79.4%) and phase 3 (below: 63.2%, above: 79.2%), indicating the effect of training, particularly for participants who are initially worse at deepfake detection.

For real images, the main effect of *baseline performance* (below: 52.7%, above: 69.9%, $F_{(1,232)} = 63.82, p < .001, \eta_{\text{p}}^2 = .216$) and its interaction with *phase* ($F_{(2,464)} = 7.53, p = .007, \eta_{\text{p}}^2 = .032$) were also significant. Moreover, the two-way interaction of *baseline performance* × *condition* ($F_{(2,232)} = 5.33, p = .005, \eta_{\text{p}}^2 = .044$) was qualified by the three-way interaction of *baseline performance* × *phase* × *condition* ($F_{(4,464)} = 3.94, p = .021, \eta_{\text{p}}^2 = .033$). Across conditions, worse performance was obtained in the artifact condition for the above-median group but in the generation condition for the below-median group. Figure 6(b) shows a continuously decreasing accuracy of real image detection for the above-median group in the artifact condition (phase 1: 75.8%, phase 2: 62.1%, phase 3: 53.7%) but not the other conditions [generation (phase 1: 77.7%, phase 2: 72.5%, phase 3: 71.9%); artifact and generation (phase 1: 78.8%, phase 2: 67.0%, phase 3: 69.7%)]. Such results suggest that repeated training only on the artifacts could reinforce the participants' bias to judge all images as fake.

General Discussion

It is clear from our findings that all training strategies were effective in enhancing participants' detection of AI-synthesized faces and their detection confidence (**RQ1** & **RQ2**). As reflected by the tendency of better detecting style-GAN2 images (**RQ2**) and continuously increased detection confidence (**RQ3**), the training effect could become more robust when both strategies (i.e., artifact and generation) were provided. While artifacts were essential for participants to detect AI-synthesized faces, repeated training only on artifacts could increase participants' bias to judge all images as synthetic (**RQ3**). Participants also learned extra artifacts beyond the training, and developed some generalized strategies for deepfake detection (**RQ4**). While participants may have learned both aspects through an online search, the results suggest that the training motivated them to further explore the learning space for effective deepfake detection.

Our findings have theoretical and practical implications. We discuss each aspect in detail as follows.

Theoretical Implications. First, we expanded the investigation of explicit training (i.e., artifacts of the state-of-the-art StyleGAN techniques) and compared it with implicit training (i.e., interactive experience of how AI-synthesized faces are generated using real human faces) and a combination of explicit and implicit training. Although all training strategies were effective in improving deepfake detection and detection confidence, only participants in the explicit training (artifact) condition were biased to judge real human faces as deepfake [see Figures 4(a) and 6(b)]. In our analysis, we found that missing or uncertain ground truth of real human faces could be the primary reason for the bias. As shown in Figure 1, participants in the artifact condition only viewed AI-synthesized faces in the training. In contrast, participants in the other two conditions had the opportunity to view both AI-synthesized faces and real human faces. We note that Nightingale and Farid (2022) did not report such bias. A comparison with their study design suggests that the trial-based feedback mechanism in their work could have played a similar role as the generation experience in our study, with which participants learned the ground truth of real human faces. Thus, our results highlight the *importance of including both positive (signal, e.g., synthetic faces) and negative (noise, e.g., real human faces) examples on human skill acquisition to detect AI-synthesized content.*

Second, we created our training materials based on images generated using StyleGAN3 technique, in which the artifacts of AI-synthesized faces are easy to detect. Thus, it was expected that the detection performance of GAN3 images was better than that of GAN2 images. On the one hand, such a result indicates the development of the desired capability after the training. On the other hand, it suggests that skills gained during GAN3 images training failed to find appropriate application in the evaluation of images generated by StyleGAN2 technique. Ideally, skill transfer across different deepfake techniques is desired. In line with the additive hypothesis of training, we also investigated a condition by concatenating two training strategies. The benefits of additivity were suggested in the relatively better detection of GAN2 images in the artifact and generation condition than in the generation condition. Because the baseline performance in phase 1 was similar between the two conditions, the detection difference were mainly due to the gap in phases 2 and 3 [see Figure 4(d)]. While these results seem promising, they are not conclusive. Thus, to better detect AI-synthesized faces without noticeable artifacts (e.g., GAN2 images), it would have to explore alternative training materials–such as opportunities to compare GAN2 images with real human faces–to facilitate the detection.

Furthermore, responses to the open-ended question revealed that participants noticed or searched artifacts beyond the training. They also integrated and organized what they learned (e.g., "pay attention to details (P92) and zoom in to see things that are abnormal (P42)"). Such engaging and self-construction activities lend support to the idea that active (Chase and Simon 1973) or constructive (Schwartz and Bransford 1998) learning could also be considered to enhance human deepfake detection. For example, to encourage and elicit constructive learning, direct prompting can be added after training, asking participants to identify new artifacts and integrate different artifacts for generalization.

Practical Implications. Our findings also carry important practical implications. Key questions include how to design interfaces and systems that support effective training and detection, and how to equip users to promptly and efficiently identify AI-synthesized faces. First, the findings of our study suggest the effect of explicit and implicit training in deepfake detection. Considering the current emphasis on explainable AI (Gunning and Aha 2019), we recommend including these training strategies in the explanations of AI systems. Second, online social media platforms could consider embedding hover-over or zoom-in functions to allowing detailed overview of user profile images. Such an interactive interface could increase the opportunities for online users to check profile images and potentially detect deepfake images. Also, we believe that the training intervention itself is not sufficient for deepfake detection, given the rapid evolvement of generative AI technology. We also suggest implementing supplementary technical solutions (e.g., automatic deepfake detection) to facilitate effective and prompt detection (Farid 2022).

Limitations. Several limitations of this study are worth noting. *First*, we chose to recruit Prolific workers in the U.S. for high data quality (Eyal et al. 2021). Although the participants' demographics are diverse, they tend to be relatively young and have higher education levels, leading to concerns about the generalizability of our findings. Future studies could benefit from recruiting more balanced and representative samples, as well as participants from other countries, regions, or cultural backgrounds. *Second*, our study focused on detecting AI-synthesized faces instead of investigating the detection of AI-synthesized faces in social engineering or disinformation campaigns. Future efforts need to explore deepfake image detection and training in more ecologically valid settings. *Third*, all our training materials are styleGAN3 images. Future studies may consider including styleGAN2 images in the training. *Fourth*, our training methods are driven by theoretical considerations. Future work should explore the feasibility of these training methods in the wild and investigate other types of training beyond the current work. *Finally*, we only investigated images generated using styleGAN2 and styleGAN3 techniques. It is important to acknowledge that there are other types of AI-synthetic techniques. During our study, diffusion models, such as stable diffusion (Rombach et al. 2022) and DaLL-E (Ramesh et al. 2021) that can generate various image variants from text prompts using a large language model, have been the focus of generative AI. While their use has been mainly to generative artwork, recent studies showed that those models can be used to generate unsafe images (Qu et al. 2023). Because the malicious use of generated unsafe images is evident (Walker 2023), future work should devote efforts to designing tools and techniques to help online users detect AI-synthesized faces generated by Text-to-Image models.

Conclusion

The dissemination of AI-synthesized faces (i.e., deepfake images) on social media platforms has become a global issue with critical implications individually and societally. Mitigating the negative impact of AI-synthesized faces on humans requires a better understanding of effective training methods that can enhance humans' ability to distinguish real human faces and synthetic faces. Our study extended previous research on human deepfake-detection training and showed the effectiveness of explicit training (artifacts in AI-synthesized faces), implicit training (experiencing the generation of synthetic faces using real human faces), and a combination of explicit and implicit training, on human detection accuracy and their decision confidence. The benefits of combining different training strategies were revealed in participants' continuously increased detection confidence and their tendency to better detect AI-synthesized faces without noticeable artifacts. Yet, we found that training only on artifacts increased participants' bias to judge all images as synthetic, highlighting the importance of knowing both signal (synthetic faces) and noise (real human faces) for the detection. We also observed that participants developed generalized strategies and leveraged artifacts beyond the training for the detection. We hope this work can provide a basis for the future development of training for human detection of AI-synthesized content.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and suggestions. We also thank Shih-Chieh Dai for his help in survey interface design and data analysis. This research was supported in part by the National Science Foundation under grants #1915801, #1934782, #2121097, #2229876, and the Graduate Research Fellowship Program under Grant No 21-46756.

References

Atleson, M. 2023. Chatbots, deepfakes, and voice clones: AI deception for sale. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale. Accessed: 2023-07-10.

Boyd, A.; Tinsley, P.; Bowyer, K.; and Czajka, A. 2023. The value of ai guidance in human examination of synthetically-generated faces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5930–5938.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.

Brooks, T.; Princess, G.; Heatley, J.; Jeremy, J.; Kim, S.; Samantha, M.; Parks, S.; Reardon, M.; Rohrbacher, H.; Sahin, B.; Shani, S.; James, S.; Oliver, T.; and Richard, V. 2018. Increasing thread of deepfake identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf. Accessed: 2024-01-10.

Chadwick, J. 2022. Russian trolls are busted trying to undermine President Zelensky by creating FAKE Facebook profiles for AI-generated Ukrainian citizens who 'want to escape their country's neo-Nazi dictatorship'. https://www.dailymail.co.uk/sciencetech/article-10570087/Russia-accused-creating-social-media-accounts-fake-Ukrainians.html. Accessed: 2023-07-10.

Chase, W. G.; and Simon, H. A. 1973. Perception in chess. *Cognitive Psychology*, 4(1): 55–81.

Chen, E.; Seo, H.; Ruffin, M.; Lee, D.; Wang, G.; and Xiong, A. 2025. Image Set Details. https://osf.io/7cgak/. Accessed: 2025-03-30.

Cozzolino, D.; Thies, J.; Rössler, A.; Riess, C.; Nießner, M.; and Verdoliva, L. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*.

Diakopoulos, N.; and Johnson, D. 2021. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7): 2072–2098.

Dodou, D.; and de Winter, J. C. F. 2014. Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36: 487–495.

Eyal, P.; David, R.; Andrew, G.; Zak, E.; and Ekaterina, D. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20.

Farid, H. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4).

Faul, F.; Erdfelder, E.; Lang, A.-G.; and Buchner, A. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2): 175–191.

Garry, M.; and Wade, K. A. 2005. Actually, a picture is worth less than 45 words: Narratives produce more false memories than photographs do. *Psychonomic Bulletin & Review*, 12: 359–366.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Gunning, D.; and Aha, D. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2): 44–58.

Hancock, J. T.; and Bailenson, J. N. 2021. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3): 149–152.

Hulzebosch, N.; Ibrahimi, S.; and Worring, M. 2020. Detecting CNN-generated facial images in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 642–643.

Johnson, A.; and Proctor, R. W. 2016. *Skill acquisition and training: Achieving expertise in simple and complex tasks*. New York, NY: Routledge/Taylor & Francis Group.

Karras, T.; Aittala, M.; Laine, S.; Harkonen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-Free generative adversarial networks. In *Advances in Neural Information Processing Systems (NEURIPS)*, 1049–5258.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4217–4228.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Lago, F.; Pasquini, C.; Böhme, R.; Dumont, H.; Goffaux, V.; and Boato, G. 2021. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1): 109–116.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207–3216.

Liu, Z.; Qi, X.; and Torr, P. H. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8060–8069.

Marra, F.; Gragnaniello, D.; Cozzolino, D.; and Verdoliva, L. 2018. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 384–389. IEEE.

Maung, B. M.; McBride, K.; Lucas, J. S.; Tabar, M.; and Lee, D. 2024. Generative AI Disproportionately Harms Long Tail Users. *IEEE Computer*, 57(11): 82–85.

Nightingale, S. J.; and Farid, H. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8): e2120481119.

Nightingale, S. J.; Wade, K. A.; and Watson, D. G. 2017. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications*, 2: 1–21.

NVIDIA. 2022. StyleGAN3 pretrained models. https://catalog.ngc.nvidia.com/orgs/nvidia/teams/research/models/stylegan3. Accessed: 2022-07-10.

Proctor, R. W.; and Van Zandt, T. 2008. *Human factors in simple and complex systems*. Boca Raton, FL: CRC Press.

Prolific. 2024. Prolific's attention and comprehension check policy. https://researcher-help.prolific.com/en/article/fb63bb. Accessed: 2024-03-10.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 3403–3417.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruffin, M.; Seo, H.; Xiong, A.; and Wang, G. 2024. Does It Matter Who Said It? Exploring the Impact of Deepfake-Enabled Profiles on User Perception towards Disinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1328–1341.

Saner, E. 2024. Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box. Accessed: 2024-03-10.

Schwartz, D. L.; and Bransford, J. D. 1998. A time for telling. *Cognition and Instruction*, 16(4): 475–5223.

Shah, A. 2018. Deepfakes : How much of what we see is real? — (Part 3). https://anvayshah.medium.com/the-2016-star-wars-movie-rogue-one-a-prequel-to-the-1977-film-a-new-hope-used-cgi-to-bring-back-3b6133520f5f. Accessed: 2023-01-10.

Singh, K.; Aggarwal, P.; Rajivan, P.; and Gonzalez, C. 2019. Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, 453–457. SAGE Publications Sage CA: Los Angeles, CA.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K.; et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Walker, A. 2023. Pope in white puffer jacket. https://knowyourmeme.com/memes/pope-in-white-puffer-jacket-pope-francis-drip. Accessed: 2024-09-20.

Zholudev, V.; and Kalaydin, P. G. 2023. Deepfakes are the new big threat to business. How can we stop them? https://sumsub.com/blog/liveness-and-deepfake-detection/. Accessed: 2024-03-10.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No, because the potential risk of misuse is minimal.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, it is included as a URL in the References section.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, it is included in the Procedure section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, our data does not contain any PII or offensive content.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? No, but we present critical instructions, stimuli, and questions in the paper and the appendix.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? No, we discussed how data is stored, shared, and deidentified in the IRB protocol, but we did not discuss it in our paper.

Broader Impact and Ethical Statement

Our research protocol was approved by the Institutional Review Board (IRB) of the authors' institution. We asked for informed consent form each participants. We also took suitable steps in our data collection and analysis to ensure an ethical study and preserve user privacy. In addition, we did not name any Prolific account in this paper to protect participants' privacy. Moreover, we debriefed the AI-synthesized images to the participants. With the development of generative AI, the chance of everyday users encountering AI-synthetic content (e.g., text, images, and videos) has increased. Our study is to gain a better understanding of human skill acquisition and training in a nuanced setting. The findings reveal the promises and limitations of different training strategies. Thus, it is essential to explore diverse training methods to provide viable mitigation to the negative consequences of synthetic content.